A survey of nature-inspired algorithm for partitional data clustering

S Suresh Babu and K Jayasudha

Department of Computer Science, Thiruvalluvar Government Arts College, Rasipuram, Tamil Nadu, India

E-Mail: 1sradhababu@gmail.com, 2jayasudhakaliannan@gmail.com

Abstract. The aim of the clustering is representing the huge amount of data objects by a smaller number of clusters or groups based on similarity. It is a task of good data analysis tool that required a rapid and precise partitioning the vast amount of data sets. The clustering problem is bring simplicity in modelling data and plays major role in the process of data mining and knowledge discovery. In the early stage, there are many conventional algorithm are used to solve the problem of data clustering. But, those conventional algorithms do not meet the requirement of clustering problem. Hence, the nature-inspired based approaches have been applied to fulfil the requirements data clustering problem and it can manage the shortcoming of conventional data clustering algorithm. This present paper is conducting a comprehensive review about the data clustering problem, discussed some of the machine learning datasets and performance metrics. This survey paper can helps to researcher in to the next steps in future.

Keywords: Bacterial Colony Optimization (BCO), Bacterial Foraging Optimization (BFO), Particle Swarm Optimization (PSO), K-Means clustering algorithm, Data Clustering.

1. Introduction

Data mining is a well-known data analysis tool which aims at the mining of unknown information from huge databases. Data mining is used to predict upcoming behaviors that offer businesses in order to build knowledge-driven decisions [1]. Clustering is an unsupervised learning technique that deals with the discovery a structure in a collection of unlabeled data [2]. Clustering is used to segregate the huge amount of data objects into smaller groups such that the similar data objects are in the same groups and dissimilar data sample are in different groups [3]. Developing an efficient clustering technique is a difficult task because of the clustering is NP-complete when the quantity of clusters is more than three clusters [4]. The clustering problem is applied to solve many applications including image segmentation [5], document clustering [6], Disease prediction [7], wireless sensor networks [8], social network analysis [9], network traffic identification [10], information retrieval [11], marketing [12].

The partitional clustering method is mainly applied method in order to solve many real world applications. It is the process of divides the data objects into small groups based on a definite measure called distance function. The number of groups is identified prior to the execution. First, it makes a set of partitions based on a number of groups based on a definite measure. Then data samples are distributed from one group to another group iteratively. It takes the center of data samples as the center of corresponding groups [13]. In the partitional clustering, the k-means algorithm is most fashionable and extensive partitioning clustering algorithms due to its better-quality practicability and
effectiveness in dealing with a huge amount of data objects. The foremost drawback of the k-means algorithm is sensitive to the selection of the initial cluster centers, converge to the local optima and fail to find global minima. In order to solve above mentioned shortcomings, many nature-inspired based algorithms have been carried out to solve the clustering problem [14]. Data clustering problem is solved by nature-inspired algorithm that have supposedly outperformed when compared with many classical methods. In the literature, there are many evolutionary and swarm intelligence based optimization techniques have been applied to solve clustering problem which are followed by nature-inspired behavior such as Genetic Algorithm (GA) [15], Particle Swarm Optimization (PSO) [16], Ant Colony Optimization (ACO) [3], Bacterial Chemotaxis (BC) [17], Bacterial Colony Optimization (BCO) [18], Social Spider Optimization (SSO) [19], Firefly Algorithm (FA) [14], Bacterial Foraging Optimization (BFO) [20], Efficient Stud Krill Herd (ESKH) [21]. The above mentioned nature-inspired algorithms are easily applied to solve clustering problem with achieve a global best possible solution for the given search space.

The major contribution of the paper is conducting a brief survey that related to the data clustering problem solved by especially using k-means, PSO, BFO and BCO. Also, presented mostly used UCI machine learning datasets and performance metrics. The remaining part of the paper is organized as follows: Section 2 is demonstrating the problem definition of data clustering. Section 3 presented the details of clustering algorithm with their related works. Section 4 is demonstrated the details of datasets which are mostly used in the literature for solving the clustering problem. Section 5 is demonstrated some performance metrics in order to estimate the performance of the clustering algorithms. The section 6 discussed the conclusions of this paper with future enhancements.

2. Data clustering

Given a data samples \( X = \{x_1, x_2,\ldots, x_n\} \) find out a partition of the \( N \) data samples into number of \( K \) cluster \( C = C_1, C_2,\ldots, C_K \). The objective of the clustering technique is to reduce the sum of squared error (SSE) [3] which can be defined as follows

\[
SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2
\]  

(1)

Where, \( \|x_i - c_k\|^2 \) is the distance values between data samples \( x_i \) and center value \( c_k \). The Euclidean distance function is used to calculate the SSE values. The low SSE value is considered as best optimal cluster group.

3. Clustering algorithms

The present section discuss the details about the k-means algorithm and related works, the details about PSO clustering algorithm with related works, the details of BFO with their related works, the details of the BCO clustering algorithm with its related works, and the details are given in the Table 2. Finally, the hybrid clustering algorithms are also presented with their related works and details are given in Table 1.

3.1. K-Means clustering technique

The k-means is a well-known clustering model and simple clustering algorithm which is used to solve a clustering problem by creating the group iteratively. The k-means clustering algorithm is beginning with cluster centroid values which are selected randomly. The each data objects are allocate to the closest cluster center. The values of cluster centroid are updated by using mean value of the associated with data points belong to the appropriate class. In the traditional k-means algorithm have many shortcomings such as sensitive to initial value of the cluster center and not a success to finding the value of global optima [22].
Table 1. Survey on hybrid clustering algorithms

| Ref. No. | Year | Method | Merits | Demerits |
|----------|------|--------|--------|----------|
| [3]      | 2010 | FAPSO-ACO-K hybrid clustering algorithm (BFCA) with BFA and K-Means | Higher clustering accuracy | High computational complexity |
| [47]     | 2013 | hybrid clustering algorithm (BFCA) with BFA and K-Means | Maintaining the local searching ability and global searching ability | High computational complexity |
| [51]     | 2014 | K-Means + MCI | High clustering accuracy | High computation complexity |
| [52]     | 2019 | PSO + Hard clustering algorithm | Classification accuracy and cluster compactness | Premature convergence |
| [53]     | 2015 | ECPSO | Producing the optimal population initialization | Premature convergence |
| [54]     | 2016 | K-Means + PSO | To enhance both ability of local and global search. Higher convergence rate, local search ability and remove the problem of the more number of functional estimates in conventional CS. | Time complexity |
| [55]     | 2018 | HCSPSO | | Computation complexity |

In the earlier stage, the clustering problem is solved by using k-means clustering model [23]. The conventional k-means algorithm is failure to find a global optimum value due to its initialization of centroid values at the beginning stage. Hence, many research works have been developed in order to solve the difficulty of k-means algorithm. Khan et al. [24] (2004) was developed a latest approach to initialize the clustering center at the beginning stage for every separate attribute. Zhao et al. (2009) developed parallel version of k-means is called a parallel k-means clustering algorithm which is speeding up the computation time when applied large scale datasets [25].

Zhao et al. (2014) [26] proposes an improved k-means by using PSO. The robustness of k-means clustering algorithm is disturbed by the original cluster centers. The proposed system is using the PSO for initializing cluster centers of the k-means. The proposed methods produced low execution time compared with traditional k-means clustering algorithm. Capo et al. (2017) [27] proposed an efficient estimate to the k-means clustering technique for enormous data. The proposed efficient k-means clustering technique is recursively partitioned the whole dataset into a tiny number of clusters or groups.

Murugesan et al. [28] (2020) proposed a new initialization method for k-means clustering model is developed in order to address the problem of initialization of rough k-means clustering algorithm. Zeta values in Peters algorithm are used to initialize the initialization values of k-means and a new performance metric S/O (S is the within-variance / O is the total-variance) index has been recognized for the rough clustering algorithm. The root mean square (RMS), standard deviation (SD), S/O index, and computational time complexity are used to authorize the performance of the developed method. However, the conventional and enhanced k-means clustering algorithm could not solve local optima problem and failure to accomplish a global optimum solution.

3.2. Particle swarm optimization
The PSO algorithm is a well-known population based optimization method which was developed by R Eberhart et al. in 1995 [29]. It is used to solve problem of clustering in order to achieve well-organized clustering accurateness [16]. The PSO is followed by the behavior of the fish schooling and
bird flying. The PSO clustering model is examining for finding best possible cluster center of their search space and moving toward to an optimum result [30],

\[ V(t+1) = \omega \times V(t) + (c_1 r_1) \times (pbest(t) - X(t)) + (c_2 r_2) \times (gbest(t) - X(t)) \]  

(2)

\[ X(t+1) = X(t) + V(t) \]  

(3)

Where, \( pbest \) is the value of personal best achieved from present particles and \( gbest \) is a global best value achieved by all particle from \( c \times n \). However, the searching ability of PSO is an inefficient due to the search process not carried out completely. Hence, the PSO algorithm is obtained more iteration to complete the given solution. In this situation, the efficiency of PSO algorithm can disturbed by a problem of premature convergence [3].

Van der Merwe et al. [31] (2004) proposed two new data clustering approaches by using PSO to grouping the given data samples. The proposed method uses PSO is used to locate the center values of a user specified quantity of clusters. Then, the PSO algorithm is an extended to employ the k-means clustering technique to fix the initial swarm value of the PSO. Li-Yeh Chuang et al. (2011) [32] developed chaotic PSO for solving data clustering problem. In which, the k-means clustering technique is applied in order to achieve optimum centroid values. Kushwaha et al. (2019) [33] developed teacher–learning based optimization (TLBO) with PSO (TLBO+PSO) in order to solve the problem of data clustering. The proposed TLBO-PSO algorithm investigates through arbitrary datasets for finding the appropriate cluster centroid in order to find the global optimum value powerfully.

Rashed et al (2020) [34] proposed the clustering problem has addressed by the multi-objective PSO (MOPSO) scheme. The proposed scheme is using a crowding distance (CD) method in order to stability the optimality of the objectives in Pareto optimal solution investigation. The CD method is performed based on the dominance technique and to guarantee endurance of the greatest solution. Kharche et al. (2015) [35] paper present a new hybrid algorithm combined ACO with PSO algorithm for solving clustering problem. ACO is applied for the find the centroid values with the help of stimulation of ant colony system and PSO is used to find optimal cluster by using different fitness function.

Armano et al. (2016) [36] proposed a new multi-objective optimization model is used for solving the problem of partitional clustering. Cohesion and Connectivity are the two objective functions in order to obtain connected, well-separated, and compact clusters. Niu et al. [37] proposed a new clustering method that incorporates with different PSOs and k-means clustering technique. The PSO-based data clustering technique is used various social exchanges among neighbors to create a number of particles escape from local optima in order to improve exploration.

Guo et al. (2012) [38] proposed a new clustering method based on niching PSO (nichePSO). The author concentrates two contributions in order to enhance the clustering accuracy. First is to improve the searching ability of the PSO with the help of main swarm training models. Second, the adaptive method is used to determine the radius of sub-swarms according to the problem of actual clustering. Lai et al. (2019) [39] proposed the semi-supervised PSO (ssPSO) algorithm in order solve the data clustering. The proposed method capture the strong point of semi-supervised fuzzy c-means (ssFCM) and PSO is used to permit for a more learned search using labelled data sample across a little number of iterations while managing diversity in the search process. ssFCM algorithms can locate significant groups using existing labelled data to direct the learning process.

Wang et al. (2018) [40] proposed chaotic starling PSO (CSPSO) is developed to improve the ability of global search mechanism of conventional PSO. In CSPSO, the inertia weight of PSO is updated using a nonlinear decreasing method and the acceleration coefficients of PSO are updated using a chaotic logistic mapping mechanism to keep away from premature of the search process. An enhancing the convergence of PSO by using the dynamic disturbance term (DDT) for updating the velocity. A local searching method is inspired by the behavior of starling bird’s exploit the information of the nearby neighbors is applied to make a decision a new cooperative position and a new collective velocity for choosing particles. Additionally, fitness function and Euclidean distance function are accepted to make sure the overall convergence of the search process.
Table 2. Survey on nature-inspired clustering algorithms

| Ref. No. | Year | Method | Merits | Demerits |
|----------|------|--------|--------|----------|
| [27]     | 2016 | Efficient K-Means Peters algorithm + K-Means | High accuracy | Local optima |
| [28]     | 2020 | Peters algorithm + K-Means | High accuracy | Local optima |
| [33]     | 2019 | TLB+PSO | High accuracy | Premature convergence |
| [46]     | 2017 | BFO | fast convergence rate and high accuracy | High computational cost |
| [34]     | 2020 | MOPSO with CD | High accuracy | Premature convergence |
| [35]     | 2015 | ACO+PSO | Finding optimal centroid values | Accuracy is low |
| [36]     | 2016 | Multi-objective optimization- PSO | High clustering accuracy | Premature convergence |
| [37]     | 2017 | Lloyd’s k-means + PSO | High clustering accuracy | High computation complexity |
| [38]     | 2012 | nichePSO | Enhancing the searching ability | Low clustering accuracy |
| [32]     | 2011 | K-Means+ Chaotic PSO | Enhancing the clustering accuracy | Premature convergence |
| [40]     | 2018 | CSPSO | enhance the global search performance | Computation complexity |
| [45]     | 2015 | Object-oriented implementation of the BFS | Enhancing clustering quality | Computation complexity |
| [50]     | 2019 | BCO | Enhancing the clustering quality | Computational complexity |
| [41]     | 2018 | Accelerated two-stage PSO (ATPSO) | Reduce the intra cluster distance | Convergence rate is low |
| [42]     | 2018 | PSO | Enhance the balance between exploitation and exploration | Premature convergence |

Xu et al. (2018) [41] developed an accelerated two-stage PSO (ATPSO) for solving the problem of data clustering. The k-means clustering technique is applied to speed up particles’ convergence during the initialization of population size of the PSO. A latest index reflecting the similarity between data objects within a cluster is applied called intra cluster cohesion (ICC). Alswa’it et al. (2018) [42] developed a new data clustering technique based on PSO. The developed clustering method is used to improve the balance between exploration and exploitation with the help of two aspects. First, the kernel density estimation method is related with latest bandwidth estimation techniques to solve the premature convergence. Second, multidimensional gravitational learning coefficients are calculated.

Cai et al. (2018) [43] proposed a new data clustering algorithm based on density peaks clustering (DPC) and PSO (PDPC). First, DPC technique is used to compute the cluster centroid value of PSO. Second, a new fitness function is used that find K global optimal solutions iteratively during the PSO algorithm. Third, each data sample is allocated to K initial center values according to the principle of minimum distance. At last, the author used for updates the cluster centers and reallocates the remaining data samples to the clusters closest to the cluster centers.

3.3. Bacterial foraging optimization

Passino et al. (2002) was introduced a new nature inspired optimization algorithm followed by the bacterial foraging behavior of Escherichia coli bacteria [44]. A group of bacteria are meeting an optimal cost by following four major stages such as chemotaxis step, swarming step, reproduction step, and elimination and dispersal step. In the chemotaxis step, the bacterium is traveling towards to
the identical direction is called run or different direction is called tumble. For each bacterium, a step fitness function value is computed during either run or tumble in each chemotaxis steps. In the swarming step, the bacterium is delivers an attractant indications to other bacteria to swarm towards it during each bacterium moves.

In the reproduction steps, the power of each bacterium is assessed as the sum of the step fitness values during its complete life. All the bacteria are prepared based on their strength in reverse order. Then, first half of the high healthier bacteria are continued then the remaining low healthy bacteria are exchanged in the same place using first half healthier bacteria. In the eliminations and dispersal, the elimination and dispersal steps are implemented after completing a few steps of the reproduction process. The particular bacteria are selected based on their probability value for killing and moving to another location within the search space.

M. Wan et al. (2012) [20] developed a bacterial foraging optimization for solving data clustering and the proposed method produced higher classification accuracy compared with some conventional clustering algorithms. Mahapatra et al. (2015) [45] developed an object-oriented based Bacteria Foraging System (BFS) in order to solve the data clustering problem. The data samples are divided into smaller meaningful groups called clusters without any prior knowledge. The developed data clustering scheme have compared with K-means, GA, BCO and PSO. Zhang et al. (2017) [46] proposed a new data clustering approaches using an adaptive chaotic BFO (ACBFO). The improved BFO algorithm contains two new features, chaotic perturbation operation in each chemotactic event and adaptive chemotaxis step setting. The adaptive chemotaxis step setting make possible quick convergence speed and high-quality convergence correctness of the BFO algorithm, and the bacteria chaotic perturbation operation further offers the search to run away from local optima and accomplish superior convergence accuracy. Niu et al. (2013) [47] proposed a hybrid clustering algorithm (BFCA) combined K-Means with BFO. The proposed hybrid clustering algorithm attempts to obtain full merits of outstanding local search abilities of k-means and global search capacities of BFA.

3.4. Bacterial colony optimization

The performance of BCO is enhanced by newly created chemotaxis process combined with a communication process compared with BC and BFO algorithms [18]. The communication mechanism is established based on an interactive communication scheme with each bacterium in order achieves global optimum values with the least iterative process. The BCO is recently proposed optimization techniques followed by the behavior of E.Coli bacterial colony and it was developed by Niu and Wang in 2012 [48]. The BCO is a combination of two bacterial foraging based optimization algorithms including BFO [49] and BC [17].

The BCO has five steps such as chemotaxis and communication step, elimination and reproduction step, migration process step [48]. In the chemotaxis and communication step, the communication process is a major action of the bacterial colonies that are come with two processes including runs and tumbles. Three different types of information exchanges are performed including random direction, group information and personal information during the communication process. The relationship between the individual and the group of the bacterial colonies are considered by the position updating process. In the Elimination and Reproduction method, the searching ability of each bacterial colony is noticeable with a degree of energy level in the searching space. The probability is used to determine the energy level of each bacterial colony. The energy level of bacterial colony is deciding the after sorted and analyzed.

If the bacterial colony moves left from preserving the area, then two mechanisms will be performed either produce new individuals to change the outer one or modify the forward way to retains them effectively. At this stage, the boundary control is very significant. Because, the outer bacterial colony is considered as an unhealthy and it is ready to eliminate from the populations. Then, reproduce by the healthy bacterial colonies. The migration process is to construct more nutrition which performance is not followed by given probability. The bacterial colony would move to a new random location in the search space according to the given conditions satisfied. The aim of BCO is performing the migration
The process is used to avoid the problem of local minima. Therefore, the updating the position of the each bacterial colony as follows,

\[
\text{Position}_i(T) = \text{Position}_i(T-1) + C(i) \times [f_i \times (G_{\text{best}} - \text{Position}_i(T-1)) + (1 - f_i) \\
\times (P_{\text{best}_i} - \text{Position}_i(T-1)) + \text{turb}_i]
\]

(4)

\[
\text{Position}_i(T) = \text{Position}_i(T-1) + C(i) \times [f_i \times (G_{\text{best}} - \text{Position}_i(T-1)) + (1 - f_i) \\
\times (P_{\text{best}_i} - \text{Position}_i(T-1))]
\]

(5)

\[
C(i) = C_{\text{min}} + \left(\frac{\text{Iter}_{\text{max}} - \text{Iter}_j}{\text{Iter}_{\text{max}}}\right) (C_{\text{max}} - C_{\text{min}})
\]

(6)

Where, \( \text{turb}_i \) is a turbulent direction variance, \( C(i) \) - denote the size of chemotaxis step, \( f_i \in (0, 1) \), \( G_{\text{best}} \) - denote the global best value and \( P_{\text{best}_i} \) - denote the personal best value. \( \text{Iter}_j \) - denote the current iteration. \( \text{Iter}_{\text{max}} \) - denote maximum iteration. In the BCO clustering technique, the \( n \) number of bacterial colonies of the position is defined as \( \text{Position} \) and that is produced \( D \times K \) cluster centroid. Hence, \( n \times K \) position will be moved for achieving minimum distance that moves to the global optimum values. J. Revathi et al. (2019) developed an efficient data clustering model using BCO and produced higher clustering accuracy compared with k-means and PSO [50].

### 3.5. Hybrid clustering algorithm

The k-means clustering technique is a well famous clustering method in data clustering. It is easy to develop, simple, easy to understand and low computation complexity. However, k-means has many shortcomings including converge to local optima due to the random initialization of cluster centers. In order to solve shortcomings of k-means algorithm, many optimization algorithms have been developed to overcome the problem of local optimum. However, these optimization algorithms are suffering many shortcomings.

Niknam et al. (2010) [3] was developed a new hybrid data clustering algorithm in order to solve problem of nonlinear partitional clustering. The developed hybrid data clustering algorithms are an integrated with k-means, fuzzy adaptive PSO (FAPSO) and ACO called FAPSO-ACO-K. Krishnasamy et al. (2014) [51] present an efficient hybrid evolutionary data clustering algorithm is called as k-means with modified cohort intelligence (MCI). In the proposed data clustering method, k-means technique is combined with the MCI algorithm for solving the problem of data clustering.

de Gusmao et al. (2019) [52] introduced two hybrid clustering technique for multi-view relational datasets. The proposed hybrid clustering method combines PSO with hard clustering algorithm based on multiple dissimilarity matrices. The proposed data clustering method is utilized the merits of the ability of global convergence of PSO and the ability of local searching of conventional clustering algorithms in order to enhance the balance between exploration and exploitation. Also, the developed provides adapted versions of 11 different fitness functions are analyzed in order to deal with multi-view relational data.

Shkari et al. (2015) [53] developed an extended CPSO (ECPSO) algorithm in order to find optimum number of groups. The develop clustering algorithm is an extended version from standard PSO. The proposed method uses a chaos trail for producing the optimum population. Atabay et al. (2016) [54] a new hybrid form of clustering algorithm which is combined PSO with K-means in order to enhance the performance of clustering problem with the help of the strength of both algorithms.

Tarkhanzhe et al. (2018) [55] developed a new hybrid data clustering technique use CS, PSO and k-means (HCSPSO). The proposed hybrid data clustering algorithm is to make use of k-means and PSO clustering algorithm in order to construct a new nest in conventional CS to achieve better outcomes. The mantegna levy distribution is used to find a higher convergence rate as well as local search ability.
A part of nests have assigned to every section of the algorithm to remove the problem of the highest quantity of functional evaluations in conventional CS.

4. Datasets
Some of the mainly used UCI machines learning datasets are used to analyze the robustness of the clustering algorithm. The details of datasets are briefly discussed in this section and the numerical values are listed in Table 3.

- Artificial data set 1 has two featured with four single classes. Six hundred data samples are extracted from 4 independent bivariate normal distributions.
- Artificial data set 2 has three featured and 5 unique classes which are collected from 250 data samples.
- Glass data set contains 214 data samples with 10 features extracted from 6 different classes.
- Contraceptive method choice (CMC) consists of 1473 data samples with 10 features from classes.
- The iris dataset contains 150 data samples of flowers from the 3 classes with 4 features.
- Seeds datasets contain 210 data samples with 7 attributes from 3 classes.
- Vowel dataset has 871 data samples. There are 6 different vowel classes and 3 input features.
- Wisconsin breast cancer (WBC) consists of 683 objects characterized by nine features with 2 classes.
- Wine data contains 178 data samples with 13 features from 3 classes.
- Zoo dataset has 101 data samples with 17 features from 7 different classes.

| Datasets | No. of Classes | No. of Features | No. of data Samples | References |
|----------|----------------|-----------------|---------------------|------------|
| Artset 1 | 4              | 2               | 600                 | [3, 38, 45, 47] |
| Artset 2 | 5              | 3               | 250                 | [3, 38, 45, 47] |
| CMC      | 3              | 10              | 1473                | [32, 33, 55, 56] |
| Glass    | 6              | 10              | 214                 | [3, 20, 34, 39, 41, 42, 53, 55] |
| Iris     | 3              | 4               | 150                 | [3, 20, 32, 33, 37-39, 41, 42, 50,53, 55] |
| Seeds    | 3              | 7               | 210                 | [39, 42, 53] |
| Vowel    | 6              | 3               | 817                 | [32, 33, 42, 55] |
| Wine     | 3              | 13              | 178                 | [3, 20, 32, 33, 37-39, 41, 42, 47,50, 55] |
| WBC      | 2              | 9               | 683                 | [3, 20, 33, 37, 38, 41, 50,55] |
| Zoo      | 7              | 17              | 101                 | [20, 56] |

5. Performance metrics
The performance metrics of the clustering problem are used to investigate the strength of clustering algorithms. Based on the clustering performance metrics, it can identify how much data samples are properly classified into corresponding groups. The various performance metrics are used in the literature such as accuracy, classification error percentage (CEP), convergence rate, F-meaures, intra / inter distance, mean and standard deviations, objective functions, purity, time complexity, and validity. Table- 4 shows the some performance measures used in the existing papers.
Table 4. Description of performance measures used in data clustering problem

| Performance Metrics | Reference Number | Performance Metrics | Reference Number |
|---------------------|------------------|---------------------|------------------|
| Accuracy (%)        | [38, 39, 41, 43] | Mean and Standard deviations | [3, 37, 41] |
| CER                 | [32, 47, 54, 55] | Objective functions | [3, 50] |
| Convergence rate    | [53]             | Purity              | [53]             |
| F-Measure           | [3]              | Time Complexity     | [3, 20, 28, 34, 38, 41, 47, 53] |
| Intra / Inter Distance | [20, 32, 33, 47, 53, 55] | Validity           | [53] |

6. Conclusions
In this paper, the brief survey conducted an overall partitioned clustering focused on only k-means, PSO, BFO, BCO and its hybrid clustering algorithms. At the beginning stage, the conventional clustering algorithm has applied to solve clustering problem with higher accuracy and low computational cost. However, the conventional clustering algorithm gets stuck in problem of local optima. Hence, the nature-inspired algorithm is an alternative technique in order to solve clustering problem with the help of its searching ability. This present paper discussed various different research papers to solve the clustering problem such as the merits and demerits of the clustering techniques, datasets, and performance evaluation parameters.

The future enhancements of this paper discussed as follows: There are many issues are found in existing research work which can consider as future enhancements. Although, nature inspired clustering algorithms has produced higher accuracy. However, the computational complexity of nature inspired algorithms is remarkable significant issues due to the dimensionality of datasets. The map reducing technique can help to nature inspired algorithm in order to reduce the computation complexity. Also, the BCO algorithm is applied least research work for solving clustering problem. So, the future work will consider some enhancement in BCO in order to achieve higher accuracy.

References
[1] P. P. Mohanty, S. K. Nayak, U. M. Mohapatra, and D. Mishra 2019 A survey on partitional clustering using single-objective metaheuristic approach. Int. J. Innovative Comput. Appl. vol.10 pp.207-226
[2] A. Abraham, S. Das, and S. Roy 2008 Swarm intelligence algorithms for data clustering in Soft computing for knowledge discovery and data mining Springer pp.279-313.
[3] T. Niknam and B. Amiri, 2010 An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis Appl. Soft Comput. Vol.10 pp.183-197
[4] G. Sahoo 2017 A two-step artificial bee colony algorithm for clustering Neural Comput. and Appl. vol.28 pp.537-551
[5] A. Nithya, A. Appathurai, N. Venkatadri, D. Ramji, and C. A. Palagan 2020 Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images Measurement vol.149
[6] A. Christy and G. M. Gandhi, 2020 Feature Selection and Clustering of Documents Using Random Feature Set Generation Technique in Advances in Data Science and Management: Springer pp.67-79.
[7] S. Ramasamy and K. Nirmala, 2020 Disease prediction in data mining using association rule mining and keyword based clustering algorithm Int. J. Comput. Appl. vol.42 pp.1-8.
[8] D. K. Kotary and S. J. Nanda, 2020 Distributed robust data clustering in wireless sensor networks using diffusion moth flame optimization Engineering Applications of Artificial Intelligence vol.87
[9] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, 2020 An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit Inf. Process. Lett. & Manage. vol.57
[10] J. Yang, Y. Han, Y. Wang, B. Jiang, Z. Lv, and H. Song, 2020 Optimization of real-time traffic network assignment based on IoT data using DBN and clustering model in smart city Future Gener. Comput. Syst. vol.108 pp.976-986.
[11] P. Bedi and S. Chawla, 2010 Agent based information retrieval system using information scent Int. J. Artif. Intell. vol.3 pp.20-238.
[12] B. Yue, 2020 Topological Data Analysis of Two Cases: Text Classification and Business Customer Relationship Management in J. Phys. Conf. Ser. vol.1550
[13] A. José-García and W. Gómez-Flores, 2016 Automatic clustering using nature-inspired metaheuristics : A survey Appl. Soft Compt. vol.41 pp.192-213.
[14] J. Senthilnath, S. Omkar, and V. Mani, 2011 Clustering using firefly algorithm: performance study Swarm Evol. Comput. pp.164-171.
[15] U. Maulik and S. Bandyopadhyay, 2000 Genetic algorithm-based clustering technique Pattern Recognit. Vol.33 pp.1455-1465.
[16] T. Cura, 2012 A particle swarm optimization approach to clustering Expert Syst. Appl. vol.39 pp.1582-1588.
[17] S. D. Muller, J. Marchetto, S. Airaghi, and P. Kournoutsakos, 2002 Optimization based on bacterial chemotaxis IEEE Trans. Evol. Comput.
[18] B. Niu and H. Wang, 2012 Bacterial colony optimization Discrete Dyn. Nat. Soc.
[19] P. Padmavathi, V. Eswaramurthy, and J. Revathi, 2018 Fuzzy Social Spider Optimization Algorithm for Fuzzy Clustering Analysis in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) IEEE pp.1-6.
[20] M. Wan, L. Li, J. Xiao, C. Wang, and Y. Yang 2012 Data clustering using bacterial foraging optimization Int. J. Intell. Inf. Technol. Syst. Vol.38 pp.321-341
[21] K. Baalamurugan and S. V. Bhanu, 2018 An efficient clustering scheme for cloud computing problems using metaheuristic algorithms Cluster Comput. pp.1-11.
[22] T. J. E. S. w. A. Cura, 2012 A particle swarm optimization approach to clustering vol.39
[23] A. Likas, N. Vlassis, and J. J. Verbeek 2003 The global k-means clustering algorithm Pattern recognti. Vol.36 pp.451-461
[24] S. S. Khan and A. Ahmad, 2004 Cluster center initialization algorithm for K-means clustering Pattern Recognit. Lett. vol.25 pp.2293-1302
[25] W. Zhao, H. Ma, and Q. He, 2009 Parallel K-Means Clustering Based on MapReduce Springer Berlin Heidelberg in Cloud Comput. pp.674-679.
[26] M. Zhao, H. Tang, J. Guo, and Y. Sun, 2014 Data Clustering Using Particle Swarm Optimization Springer Berlin Heidelberg in Future Information Technology pp.607-612.
[27] M. Capó, A. Pérez, and J. A. Lozano, 2017 An efficient approximation to the K-means clustering for massive data Knowledge Based Syst. vol.117 pp.56-69
[28] V. P. Murugesan and P. Murugesan, 2020 A new initialization and performance measure for the rough k-means clustering Soft Comput. pp.1-15
[29] R. Eberhart and J. Kennedy, 1995 Particle swarm optimization in Proceedings of the IEEE international conference on neural networks vol.4 pp.1942-1948.
[30] K. Velusamy and R. Manavalan, 2012 Performance Analysis of Unsupervised Classification based on Optimization Int. J. Comput. Appl. vol.975 no.8887
[31] D. Van der Merwe and A. P. Engelbrecht, 2003 Data clustering using particle swarm optimization in The 2003 Congress on Evolutionary Computation vol.1 pp.215-220.
[32] L.-Y. Chuang, C.-J. Hsiao, and C.-H. Yang, 2011 Chaotic particle swarm optimization for data clustering Exp. Sys. with Appl. vol.38 pp.14555-14563
[33] N. Kushwaha and M. Pant, 2019 A Teaching–Learning-Based Particle Swarm Optimization for Data Clustering in Machine Intelligence and Signal Analysis pp.223-233.
[34] A. B. Rashed, H. Hamdan, N. M. Sharef, M. N. Sulaiman, R. Yaakob, and M. Abubakar, 2020 Multi-objective clustering algorithm using particle swarm optimization with crowding distance (MCPSo-CD) Int. J. Adv. in Intell. Inf. vol.6 pp.2-81

[35] D. Kharche and A. Thakare, 2015 ACPSO: Hybridization of ant colony and particle swarm algorithm for optimization in data clustering using multiple objective functions in 2015 Global Conference on Communication Technologies (GCCT) IEEE pp.854-859.

[36] G. Armano and M. R. Farmani, 2016 Multiobjective clustering analysis using particle swarm optimization Exp. Syst. with Appl. vol.55 pp.184-193.

[37] B. Niu, Q. Duan, J. Liu, L. Tan, and Y. Liu, 2017 A population-based clustering technique using particle swarm optimization and k-means Natural Comput. vol.16 pp.45-59

[38] C. Guo and Y. Zang, 2012 Clustering algorithm based on density function and nichePSO J. Syst. Eng. Electron vol.23 pp.445-452

[39] D. T. Lai, M. Miyakawa, and Y. Sato, 2020 Semi-supervised data clustering using particle swarm optimisation Soft Comput. vol.24 pp.3499-3510.

[40] L. Wang, X. Liu, M. Sun, J. Qu, and Y. Wei, 2018 A New Chaotic Starling Particle Swarm Optimization Algorithm for Clustering Problems Math. Prob. Eng. vol.825 no.0480.

[41] X. Xu, J. Li, M. Zhou, J. Xu, and J. Cao, 2018 Accelerated two-stage particle swarm optimization for clustering not-well-separated data IEEE Trans. Syst. Man Cybern. Syst.

[42] M. Alswaitti, M. Albughdadi, and N. A. M. Isa, 2018 Density-based particle swarm optimization algorithm for data clustering Exp. Syst. with Appl. pp.170-186

[43] J. Cai, H. Wei, H. Yang, and X. Zhao, 2020 A Novel Clustering Algorithm Based on DPC and PSO IEEE Access vol.8 pp.88200-88214

[44] K. M. Passino, 2010 Bacterial foraging optimization Int. J. of Swarm Intell. Res. (IJSIR) 1 1-16

[45] G. Mahapatra and S. Banerjee, 2015 An object-oriented implementation of bacteria foraging system for data clustering application in 2015 International Conference and Workshop on Computing and Communication (IEMCON) vol.15-17 pp.1-7

[46] Y. Zhang and C. Huang, 2017 An improved bacterial foraging optimization based approach to data clustering in Proc. Int. Conf. Artif. Intell. Sci. Technol World Scientific pp.91-99.

[47] B. Niu, Q. Duan, and J. Liang, 2013 Hybrid Bacterial Foraging Algorithm for Data Clustering 2013: Springer Berlin Heidelberg, in Intelligent Data Engineering and Automated Learning – IDEAL pp.577-584.

[48] B. Niu, H. J. D. D. i. N. Wang, and Society, 2012 Bacterial colony optimization

[49] K. M. Passino, 2002 Biomimicry of bacterial foraging for distributed optimization and control IEEE Control Syst. Mag. Vol.22 pp.52-67

[50] J. Revathi, V. Eswararamurthy, and P. Padmavathi, 2019 Bacterial Colony Optimization for Data Clustering in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) IEEE pp.1-4.

[51] G. Krishnasamy, A. J. Kulkarni, and R. Paramesran, 2014 A hybrid approach for data clustering based on modified cohort intelligence and K-means Exp. Syst. with Appl. vol.41 pp.6009-6016.

[52] R. P. de Gusmão and F. d. A. de Carvalho, 2019 Clustering of multi-view relational data based on particle swarm optimization Exp. Syst. With Appl. vol.123 pp.34-53.

[53] M. Lashkari and M. H. Moattar, 2015 The improved K-means clustering algorithm using the proposed extended PSO algorithm in 2015 International Congress on Technology, Communication and Knowledge (ICTCK) IEEE pp.429-434.

[54] H. A. Atabay, M. J. Sheikhzadeh, and M. Torshizi, 2016 A clustering algorithm based on integration of K-Means and PSO in 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC) IEEE pp.59-63.

[55] O. Tarkhaneh, A. Isazadeh, and H. J. Khamnei 2018 A new hybrid strategy for data clustering using cuckoo search based on Mantegna levy distribution, PSO and k-means Int. J. of Comp. Appl. in Technol. pp.137-149.