Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

Optimized Clustering Algorithm with Swarm Intelligence for Large Dataset

Rajendra DevidasMandaokar

SATI, Vidisha, MP

rdmandaokar@rediffmail.com

Dr. Shailesh Jaloree

SATI, Vidisha, MP

Shailesh_jaloree@rediffmail.com

ABSTRACT

The compactness and content validation of the clustering algorithm is a significant challenge. The partition-based clustering algorithms achieve the paramount of compactness and validation. The swarm-based clustering algorithm has great protentional over the large dataset. This paper presents the modification of K-means algorithms with particle swarm optimization and ant colony optimization. Both swarm algorithms enhance the performance of the K-means algorithm. The derive objective function convert into the fitness function of the swarm algorithm. The modified algorithm called KACO and KPSO implement in MATLAB software. To evaluate the performance of the modified algorithm, applied one real dataset and two synthetic datasets. The results of the modified algorithm are better than the FCM and K-means algorithm. The validation of cluster centers distance measure by different distance formulae such as Euclidean distance, cosine distance, and others. The applied distance formula influences the results of the clustering algorithm. (The optimal solution set of centers formed an optimal cluster.) For empirical evaluation, measure three parameters like the number of iterations, standard deviation, error rate. The minimization factor of validation parameters indicates the compactness and validation of the clustering algorithm.

Keywords: - Clustering, Partition, K-means, FCM, ACO, PSO, MATLAB, Large dataset.

1. INTRODUCTION

Clustering algorithms are the backbone of data engineering and pattern analysis. The unknown attribute of data and large-size faces of bottleneck problem of grouping and estimating data similarity is resolved by the clustering algorithm[1, 2]. The diverse nature of clustering algorithm divided into several groups such as partition clustering, hierarchical clustering, density-based clustering and micro clustering[3]. The principle of clustering further explodes into three sections such as overlapping, partitional and hierarchical[4, 5]. The efficient and straightforward clustering process is called partition clustering. The partition clustering has several variants of algorithms such as K-means, K-mode and FCM. The FCM clustering algorithm is also called soft clustering algorithm[6, 7, 8]. The significant challenge for handling clustering algorithms is the number of iteration and selection of seed for cluster formation[9, 10, 11]. Despite these problems, clustering is the first option for pattern analysis of an extensive database. The reported survey withstands two clustering processes, partitioning and hierarchical clusteringthe hierarchical clustering is in two approaches, agglomerative and divisive[12]. The formation of these clustering algorithms is top to bottom in the bottom-up, and another side, the top-down clustering is bottom to top. The utilization of partition clustering and hierarchical clustering is very high in extensive dataset analysis[13]. The performance of these clustering algorithms is a significant issue. A large number of iteration and random selection of cluster selection degraded the performance of the clustering algorithm[14]. The various authors and research work in partition clustering with other algorithms and improve clustering algorithms' performance. The counter race between partition clustering and hierarchical clustering due to their performance. The quality indicator of hierarchical clustering is suitable in respect of partition clustering[15]. But the reallocation of the object is very poor. Due to this reason, the hierarchical clustering algorithm cannot handle large scale of data. Partition clustering's major advantage is to take large scale dataset is

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

very efficient and manage time complexity. This paper mainly focuses on partition clustering algorithms, such as K-means and FCM. The limitation of K-means and FCM overcome with swarm intelligence algorithms[16, 17]. Swarm intelligence-based algorithm move on next stage of data clustering and pattern analysis. Swarm intelligence is a bio-inspired heuristic function to help in the case of searching and merging intermediate cluster. The behaviors of swarm-based algorithms defined into two categories such as single algorithm and population-based algorithm[7, 12]. The single based algorithm has a specific limitation, but the population-based algorithm provides ample searching space for mapping data and provides potential results as per requirements. Many population-based algorithms such as genetic algorithm (GA), particle swarm optimization (PSO), ant colony optimization (ACO), artificial Bee colony algorithm (ABC), glowworm swarm optimization (GSO) and many more meta-heuristics algorithms implied to solve the clustering problem in large dataset. The complexity of partition-based clustering algorithms is almost linear due to every field of pattern analysis adopt this process of clustering algorithms. The K-means and FCM clustering process are widely applied in large dataset[9, 10, 15]. This paper applied two swarm intelligence algorithms such as particle swarm optimization and ant colony optimization, with the K-means algorithm and improved the algorithm's efficiency. Also, study of FCM (fuzzy C-means) algorithm on the same dataset. The analysis of the algorithm uses extensive weblog data, image dataset. The performance of algorithms validates with an evaluation of standard parameters.

2. RELATED WORK

Farmer, Jocelyn R. Et al. [15] CVID is increasingly recognized for its association with autoimmune and inflammatory complications. Despite recent advances in immunophenotypic and genetic discovery, clinical care of CVID remains limited by their inability to accurately model risk for non-infectious disease development. Herein, Authors demonstrate the utility of unbiased network clustering as a novel method to analyse inter-relationships between non-infectious disease outcomes in CVID using databases at the USIDNET, the centralized immuno-deficiency registry of the United States, and Partners, a tertiary care network in Boston, MA, USA, with a shared electronic medical record amenable to natural language processing. Immunophenotypes were comparable in terms of native antibody deficiencies, low titer response to pneumococcus, and B cell maturation arrest. Salem, Semeh Ben Et al. [16] Partitional clustering algorithms represent an interesting issue in pattern recognition due to their high scalability and efficiency. The k-means, discussed since 1965, had shown great efficiency for numeric clustering but is unfortunately inadequate for categorical clustering. In 1998, the k-modes were discussed as an extension of the k-means to cluster categorical datasets. A new categorical method based on partitions called MFk-M is detailed. It aims to convert the initial categorical data into numeric values using the relative frequency of each modality in the attributes. Patibandla, RSM Lakshmi Et al. [17] clustering algorithms have been emerged learning aid to generate and analyse the huge volumes of data. The foremost clustering objective is to classify same type of data has been grouped with in the same Cluster while they are similar according to precise metrics. For various applications, clustering is one of the techniques to classify and analyse the large amount of data. On the other hand, the main issues of applying clustering algorithms for big data that causes uncertainty among the practitioners require consent in the definition of their properties in addition to be deficient in proper classification.

Caron, Mathilde Et al. [18] Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large-scale datasets. Tripathi, Ashish Kumar Et al. [19] With advancement of the technology, data size is increasing rapidly. For making intelligent decisions based on data, efficacious analytic methods are required. Data clustering, a prominent analytic method of data mining, is being efficiently employed in data analytics. To analyse massive data sets, the improvement in the traditional methods is the urge of today's scenario. An efficient clustering method, MR-EGWO, is presented for clustering large-scale data sets. Fränti, Pasi Et al. [20] This paper has two contributions. First, Authors introduce a clustering basic benchmark. Second, Authors study the performance of k-means using this benchmark. Specifically, Authors measure how the performance depends on four factors: overlap of clusters, number of clusters, dimensionality, and unbalance of cluster sizes.

Steinegger, Martin Et al. [21] Metagenomic datasets contain billions of protein sequences that could greatly enhance large-scale functional annotation and structure prediction. Utilizing this enormous resource would require reducing its redundancy by similarity clustering. Tardioli, Giovanni Et al. [22] The formulation of energy policies for urban building stock frequently requires the evaluation of the energy use of large numbers of buildings. When urban energy modelling is utilized as part of this process, the identification of building groups and associated representative buildings can play a critical role. Rida, Mohamad Et al. [23] In WSNs, hundreds or thousands of nodes are deployed in order to provide high quality monitoring. Nowadays, they constitute one of the most important sources of big data. Such amount of collected data is a real challenge for sensor nodes

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

suffering from many limitations, especially, energy constraints. Therefore, to address big data issue, research efforts have been done today to design efficient data management techniques for WSNs. The main objective of these works is to reduce the amount of transmitted data over the network while preserving their properties.

Xu, Xiao Et al. [24] an energy efficient data management approach dedicated to big data applications in PSNs, named EK-means, has been discussed. their suggested approach is twofold: first, Authors discussed a data aggregation method at the sensors level in order to eliminate the similar data generated at each period; second, Authors discussed an enhanced version of K-means algorithm in order to cluster similar data at the aggregators level. It was shown, through simulations on real sensor data, that their approach can be effectively used to reduce the energy consumption in PSNs and accelerating the clustering of data using the EK-means algorithm.Wolf, F. Alexander Et al. [25] Single-cell RNA-seq quantifies biological heterogeneity across both discrete cell types and continuous cell transitions. PAGA provides an interpretable graph-like map of the arising data manifold, based on estimating connectivity of manifold partitions. Kiselev, Vladimir Yu Et al. [26] Single-cell RNA sequencing allows researchers to collect large catalogues detailing the transcriptomes of individual cells. Unsupervised clustering is of central importance for the analysis of these data, as it is used to identify putative cell types.

Duò, Angelo Et al. [27] Subpopulation identification, usually via some form of unsupervised clustering, is a fundamental step in the analysis of many single-cell RNA-seq data sets. This has motivated the development and application of a broad range of clustering methods, based on various underlying algorithms. Huang, Wenzhun Et al. [28] The rapid growth of Internet has vast amounts of information over online. The correct information can be provided by the source only if the information is processed, analysed and linked. The efficient store and manage model is required to access and to protect these large data. These data are structured and unstructured which is available in online in order to process such data an intense technology is required. The cloud computing satisfies the need of store and manage model. Azad, Ariful Et al. [29] Biological networks capture structural or functional properties of relevant entities such as molecules, proteins or genes. Characteristic examples are gene expression networks or protein–protein interaction networks, which hold information about functional affinities or structural similarities. Such networks have been expanding in size due to increasing scale and abundance of biological data. While various clustering algorithms have been discussed to find highly connected regions, MCL has been one of the most successful approaches to cluster sequence similarity or expression networks.

Du, Mingjing Et al. [30] The DP clustering approach, a novel density-based clustering algorithm, detects clusters with arbitrary shape. Rodriguez, Mayra Z. Et al. [31] Authors per-formed a systematic comparison of 9 well-known clustering methods available in the R language assuming normally distributed data. In order to account for the many possible variations of data, Authors considered artificial datasets with several tunable properties. Shakeel, P. Mohamed Et al. [32] Cloud computing is an interesting computing model suitable for accommodating huge volume of dynamic data. To overcome the data handling problems this, work focused on Hadoop framework along with clustering technique. This work also predicts the occurrence of diabetes under various circumstances which is more useful for the human. Sinha, Debajyoti Et al. [33] Droplet based single cell transcriptomics has recently enabled parallel screening of tens of thou-sands of single cells. Clustering methods that scale for such high dimensional data without compromising accuracy are scarce. Authors exploit Locality Sensitive Hashing, an approximate nearest neighbour search technique to develop a de novo clustering algorithm for large-scale single cell data. On a number of real datasets, dropClust outperformed the existing best practice methods in terms of execution time, clustering accuracy and detectability of minor cell sub-types.

Wolf, F. Alexander Et al. [34] SCANPY is a scalable toolkit for analysing single-cell gene expression data. It includes methods for pre-processing, visualization, clustering, pseudo-time and trajectory inference, differential expression testing, and simulation of gene regulatory networks. Its Python-based implementation efficiently deals with data sets of more than one million cells.Yang, Yan Et al. [35] the data are generated from different sources or observed from different views. These data are referred to as multi-view data. Unleashing the power of knowledge in multi-view data is very important in big data mining and analysis. This calls for advanced techniques that consider the diversity of different views, while fusing these data. MvC has attracted increasing attention in recent years by aiming to exploit complementary and consensus information across multiple views. This paper summarizes a large number of multi-view clustering algorithms, provides a taxonomy according to the mechanisms and principals involved, and classifies these algorithms into five categories, namely, co-training style algorithms, multi-kernel learning, multi-view graph clustering, multi-view subspace clustering, and multi-task multi-view clustering.

Liu, Qidong Et al. [36] Path-based clustering algorithms usually generate clusters by optimizing a criterion function. Most of state-of-the-art optimization methods give a solution close to the global optimum. By analysing the minimax distance, Authors find that cluster centres have the minimum density in their own

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

clusters. Inspired by this, Authors discussed an IPC by mining the cluster centres of the dataset. IPC solves this problem by the process of elimination since it is difficult to mine these cluster centres directly. Patibandla, RSM Lakshmi Et al. [37] Large quantity of data has been accumulating tremendously due to digitalization. But the accumulated data are not converted into useful patterns. This gap is conquered by using exploratory data analysis techniques. Clustering is one of the vital technologies in exploratory data analysis. It is a methodology to arrange data objects as per their characteristics. Traditional clustering approaches, namely leader, K-means, ISODATA and evolutionary-based approaches like genetic algorithm, particle swarm optimization, social group optimization methods, are also implemented on benchmark data set. Riaz, Sumbal Et al. [38] Authors performed sentiment analysis on the customer review real-world data at phrase level to find out customer preference by analysing subjective expressions. Then Authors calculated the strength of sentiment word to find out the intensity of each expression and applied clustering for placing the words in various clusters based on their intensity. Zou, Quan Et al. [39] Sequence clustering is a basic bioinformatics task that is attracting renewed attention with the development of metagenomics and microbiomics. The latest sequencing techniques have decreased costs and as a result, massive amounts of DNA/RNA sequences are being produced. The challenge is to cluster the sequence data using stable, quick and accurate methods. For microbiome sequencing data, 16S ribosomal RNA operational taxonomic units are typically used. Deng, Tingquan Et al. [40] Subspace techniques have gained much attention for their remarkable efficiency in rep- resenting high-dimensional data, in which SSC and LRR are two commonly used prototypes in the fields of pattern recognition, computer vision and signal processing. Both of them aim at constructing a block sparse matrix via linearly representing data to make them be embedded into linear sub- spaces.

3. METHODOLOGY

This section describes the K-means clustering, FCM clustering, ACO and PSO. Also describe the methods proposed with particle swarm intelligence and ant colony optimization algorithms.

3.1. K-MEANS CLUSTERING

The K-means clustering is member of partition-based clustering algorithms. The nature of k-means clustering algorithm is very simple and categorised in section of unsupervised learning. The process of K-means clustering deals with generation of centre points with merging of intermediate cluster. The maximum iteration of clustering process declines the performance of K-means algorithm[5].

Consider we want to categorise the data into different K groups. The objective of algorithms is minimizing the value of fitness function as squared error describe as

$$Fs = \sum_{j=1}^{k} \sum_{j=1}^{n} \|xi - cj\|^2 \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Here Xi isith point of data samples Cj is the cluster centres
The process of clustering algorithm follows following steps

1. Define the value of K and represents the data points for the initial groups of centres select randomly.
2. Measure the distance with centre point and assign object in these groups
3. Redefine the centres of intermediate clusters
4. Merger intermediate cluster
5. Repeat step 2 and 3 until the condition is met. And validate the cluster

3.2. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization resolve the problem of partition K-means clustering algorithm. The problem of K-means algorithm is minimization of objective function of similarity index. The fitness function of particle swarm optimization converted into objective function of clustering algorithm. The particle swarm optimization algorithm is dynamic population based meta-heuristic function. The working of algorithm deals in two manners local and global ways. The local mapping deals with selection of data points and global maintain the objective function of clustering algorithm. The processing of PSO algorithm based on the concept of bird fork and movements of fish in wheel[1]. The velocity and direction of particle moving with problem space of data. the mapping of process of data space in velocity and direction shown in figure[3].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

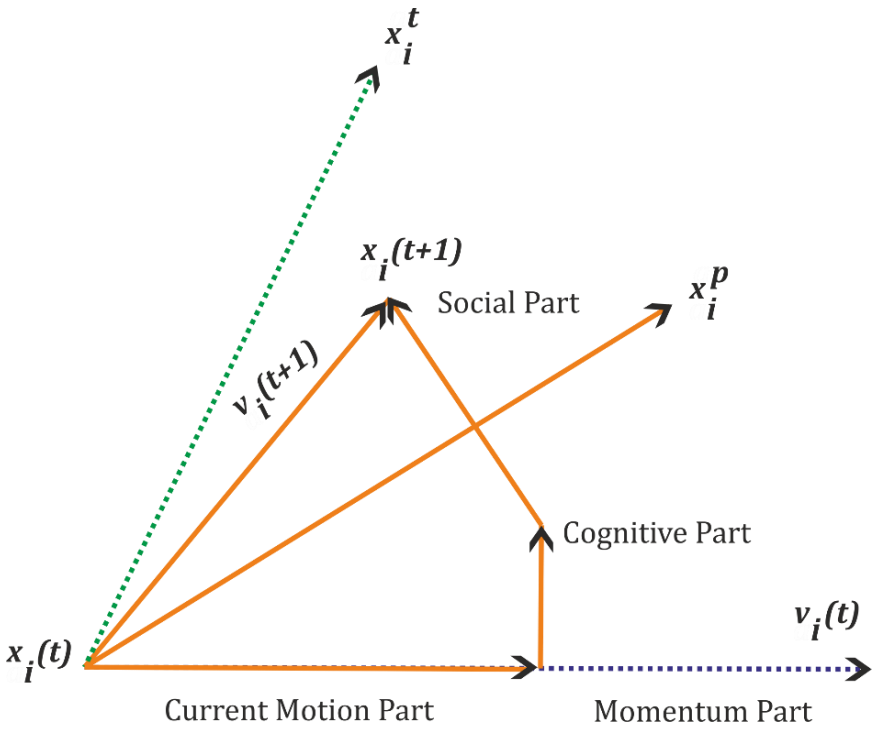

Figure 1: Represents the velocity and direction of particle in search space for the objective function.

the processing of particle swarm optimization algorithm is described here.

1. Define the population of particle as data points A
   (a) For i= 0 to M where M is maximum of particle
   (b) Initialize A[i]
2. Define the speed of particle
   (a) For i= 0 to M
   (b) Velocity[i]=0
3. Estimate particle in M
4. Reallocate the position that represents the data point of data samples
5. Generate search space D
6. Define memory o each particle
7. Compute the speed of particle
   Velcoty[i]= W*Velocity[i]+R1*(Pbset[i]-M[i])+R2*(data points[h]-A[i])

   Where the range value of R is[0,1]
8. Estimate new position of particle
   A[i]=A[i]+velocity[i]
9. Measure current position of particle
   Pbest[i]=A[i]
10. Increment of counter
11. End

3.3. ANT COLONY OPTIMIZATION (ACO)

The working nature of ant colony optimization is distributed and it influence the various real-time problem-solving capacity. The principle of distributed and similarity of path solve the problem of objective function minimization problem in K-means clustering algorithm. The various authors and research reported variants of clustering algorithm with ant colony optimization. The ant colony optimization is also based on dynamic population based meta-heuristic function. The processing of ACO algorithm process on phenomenon value and position of iteration. The process of iteration derives rule of transition[2, 7].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

Every iteration of the ACO, All ants move from a $stater$ to state s to get other intermediate solution. The k^{th} ant from $stater$ to $states$ is $selected\ among\ the\ unvisited\ states\ memorized$ in j_r^k considering to the numerical equation:

$$s = \underset{u \epsilon j_r^k}{^{arg}} max\left[\tau_i(r,u)^{\alpha}.\eta\ (r,u)^{\beta}\right]\ if\ q \leq q0(Exploitation) \qquad (2)$$

The trail level describes a posteriori notation of the move's expectation. Trails are regular modify when all ants have finished their route like $solution$, the stage of trails behalf to moves reduced or increased that were part of worst and $best\ solution\ results$. Normally, the k^{th} ant go through from state r to state $swith\ the\ probabilityp_k(r,s)$,

$$p_k(r,s) = \begin{cases} \frac{\tau_i(r,u)^{\alpha}.\eta\ (r,u)^{\beta}}{\Sigma_{u \epsilon j_r^k} \tau_i(r,u)^{\alpha}.\eta\ (r,u)^{\beta}} & if\ s \in j_r^k \\ 0 \qquad otherwise \end{cases} \qquad (3)$$

where,

$p_k(r,s)$ - $transition\ probability$

$\tau(r,s)$ - $pheromone\ concentration\ between\ the\ stater$ and the state u in the $i^{th} population$

$\eta(r,s)$- $length\ of\ the\ trail\ from\ the\ state\ r\ and\ the\ state\ u$

J_r^k - $set\ of\ unvisited\ states\ of\ thek^{th} ant\ in\ the\ i^{th}\ population,$

α and β - $control\ parameters$

q - $uniform\ probability$ $[0, 1]$

3.4. MODIFIED K-MEANS WITH PSO(KPSO)

The particle swarm optimization (PSO) algorithm improves the efficiency of k-means algorithm. The PSO algorithm control the objective function of k-means as fitness constraints function. The fitness function belongs to the distance function. The process of optimization minimizes the similarity index matrix and generate the optimal cluster[5, 7]. The objective function describes as

$$F(P) = \sum_{i=1}^{n} dist(Ci, Pi) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Here F(P) is fitness function of particle swarm optimization and Ci is center of K-means and Pi is particle assigned for data points.

Now update the position of centers and data points according to their direction and velocity

$$Po_j^i = Po_j^i + V_i^j$$

$$V_j^i = wV_i^j + c1r\left(p_i^j - po_i^j\right) + c2r\left(C_i^j - po_j^i\right) \ldots\ldots\ldots\ldots (5)$$

Here w is weight and c1 and c2 is constant r is acceleration function. The Po is new position of data points. The processing of algorithm is described here

1. Set the value of W, C1, C2 and population as p={p1,p2,…….,pn}
2. Estimation the fitness function by equation (1)
3. Define the centers of cluster C=[c1,c2,c3,……..,cn]
4. Measure fitness value F=[f1,f2,f3,……..fn]
5. Initialize velocity V=[v1,v2,v3,…..,vn]
6. Start iteration do
7. For each Ci do
8. Update Ci by equation (5)
9. Measure fitness by equation (4)
10. If Fi<PoFi then
11. PCi=Fi

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

12. End
13. Process terminated
14. Optimal cluster

3.5. MODIFIED K-MEANS WITH ACO(KACO)

The modification of k-means algorithm with ant colony optimization proceeds in manners of control of objective function and minimization of error. The ant colony optimization algorithms represent the all-data points of data in terms of ants and assigned the value of pheromone for the processing of ants[2, 7]. For the estimation of distance applied different distance estimation function is described in 3.6.

Define the value of phenomenon and start the process of clustering with this derivation

$$S_{(i,Xn)} = \frac{\tau^{\alpha}_{(i,xn)}\eta^{\beta}_{(i,Xn)}}{\sum_{i=0}^{k}\tau^{\alpha}_{(i,xn)}\eta^{\beta}_{(i,Xn)}} \dots\dots\dots\dots\dots\dots\dots\dots (6)$$

Here's is probability of ants, τ and η define the phenomenon and data state and K is number of clusters. The search space of heuristic function estimated by

$$\eta_{(i,Xn)} = \frac{k}{distance\ \ (Xn,Ci)X\ objloc\ \ (Xn,Ci)} \dots\dots\dots\dots\dots\dots\dots (7)$$

Here distance is function of difference estimation and objloc is distance with center point of k cluster.

Now the process of phenomenon update describe as

$$\tau_{(i,Xn)} <-- (1-s)\tau(i,Xn) + \sum_i \triangle\tau(i.Xn) \dots\dots\dots\dots\dots\dots\dots (8)$$

Step of algorithm

1. Define the number of ants M, define cluster K, define value of pheromone level 1
2. The data points mapped into M ants matrix and select randomly center Ci
3. Measure the probability of data points belong with objective function (6)
4. Formed new cluster. If the formation of new cluster similar to old one goto step 3
5. Estimation best solution of M ants
6. Increment and decrement the value of pheromone for continuous update
7. Update cluster centers with best solutions of ants
8. Check the condition of maximum iteration otherwise go to step 3
9. Optimal cluster

3.6. DISTANCE FORMULA

The following formula applied to measure distance between center points and data points

Standard Euclidean distance $(\sum_{l=1}^{d}|Xli - xlj|^2)$

Makowski distance $(\sum_{l=1}^{d}\left|\frac{xli-xlj}{sl}\right|^2)$

Cosine distance $1 - cos\alpha = \frac{x_i^T xj}{||xj||\,||xj||}$

Mahala Nobis distance $\sqrt{(xi-xj)TS^{-1}}(xi-xj)$

4. EXPERIMENTAL ANALYSIS

The experimental process is evaluated based on real dataset and synthetic dataset. The process of algorithm implemented in MATLAB software; the version of software is R2014(a). Thehardware configuration of the system is RAM 16GB, processor I7 and 1TB HDD. The operating system windows 10 professional. The performance of clustering algorithms measure in terms of standard deviation, computed error and number of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

iterations for the processing of clustering. For the validation of modified algorithm of data clustering validated with real dataset as well as real time dataset. The real time dataset is web log data. the web log data is very complex structure and it contains string, number and constant[2, 4, 29]. The processing of web log data is very difficult, now it needs to data transformation. The dataset contains IP address, port number, flags, time, payload and many other attributes. The other two dataset is wine dataset and image dataset. The both datasets obtained from UCI machine learning repository.

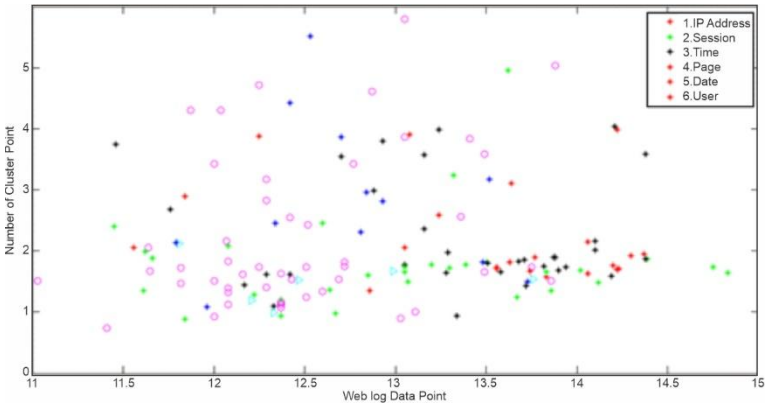

Figure 2: window show that the output web log data point and here, input field of initial value of seed is 5 and when hitting on K-Means technique button here loaded the dataset logdataset1 and imported in our simulation GUI window.

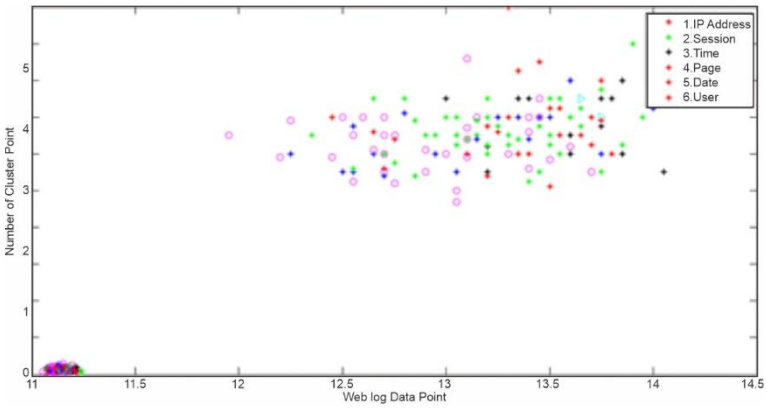

Figure 3: window show that the output web log data point and here, input field of initial value of seed is 5 and when hitting on K-ACO technique button here loaded the dataset logdataset2 and imported in our simulation GUI window.

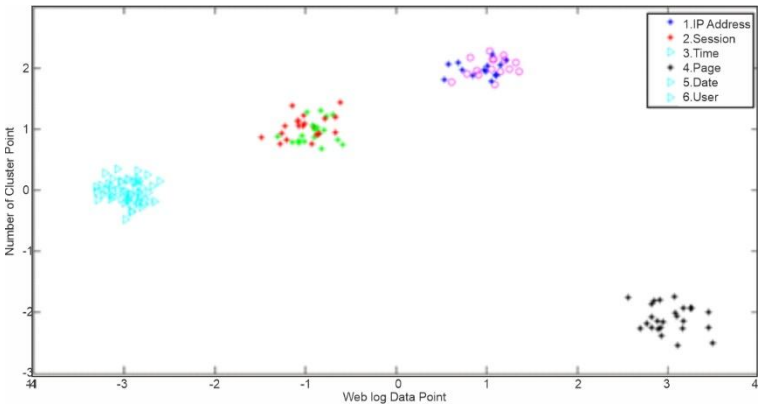

Figure 4: window show that output web log data point and here, input field of initial value of seed is 5 and when hitting on K-PSO technique button here loaded the dataset logdataset3 and imported in our simulation GUI window.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

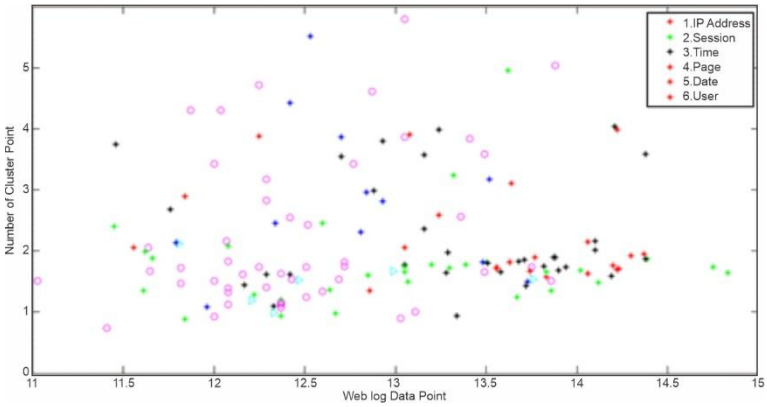

Figure 5: window show that output web log data point and here, input field of initial value of seed is 5 and when hitting on FCM technique button here loaded the dataset logdataset4 and imported in our simulation GUI window.

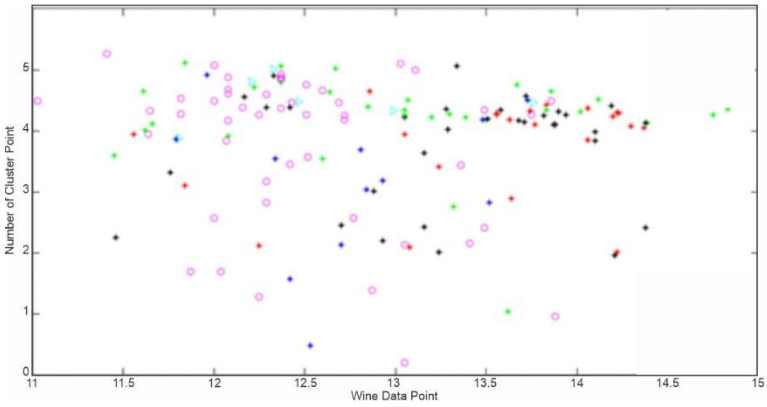

Figure 6: window show that the output wine data point and here, input field of initial value of seed is 8 and when hitting on K-Means technique button here loaded the dataset winedataset1 and imported in our simulation GUI window.

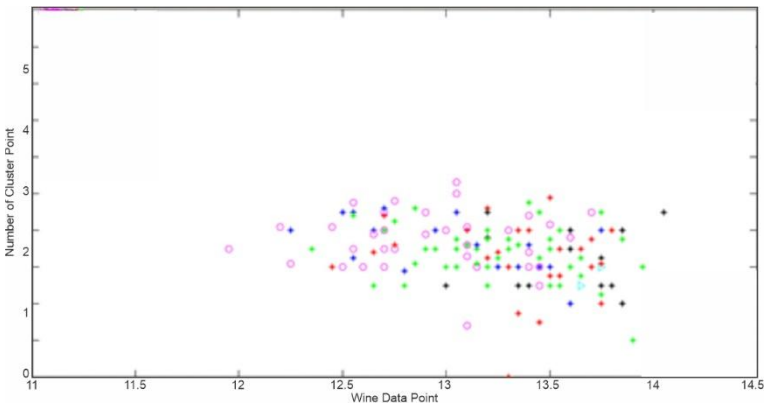

Figure 7: window show that the output wine data point and here, input field of initial value of seed is 8 and when hitting on K-ACO technique button here loaded the dataset winedataset2 and imported in our simulation GUI window.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

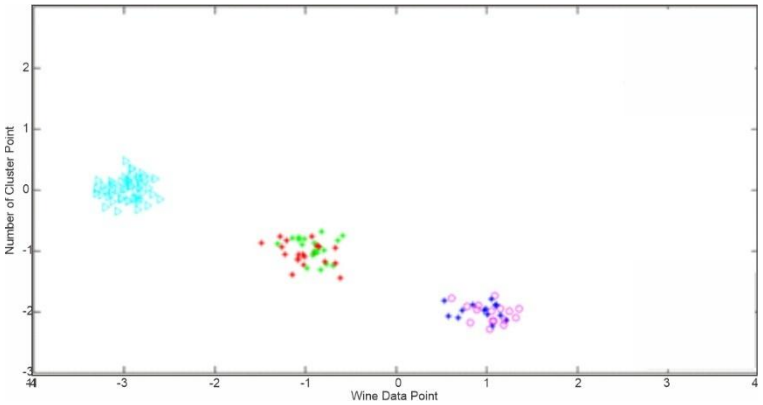

Figure 8: window show that output wine data point and here, input field of initial value of seed is 8 and when hitting on K-PSO technique button here loaded the dataset winedataset3 and imported in our simulation GUI window.

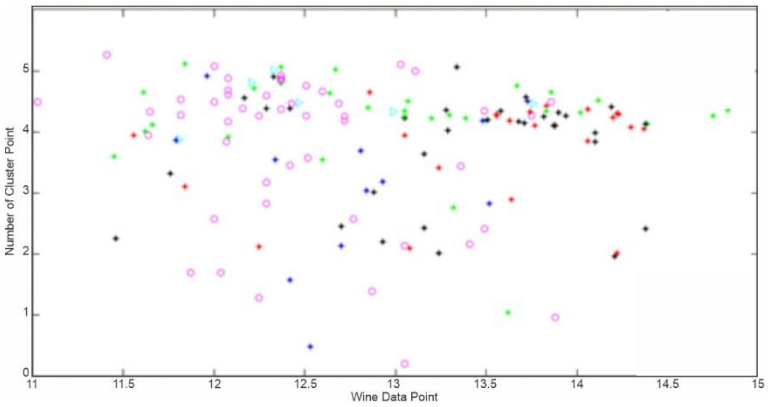

Figure 9: window show that output wine data point and here, input field of initial value of seed is 8 and when hitting on FCM technique button here loaded the dataset winedataset4 and imported in our simulation GUI window.

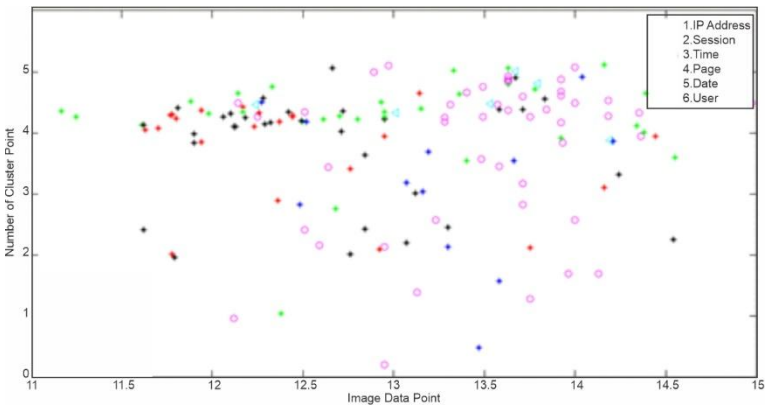

Figure 10: window show that the output image data point and here, input field of initial value of seed is 10 and when hitting on K-Mean's technique button here loaded the dataset imagedataset1 and imported in our simulation GUI window.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

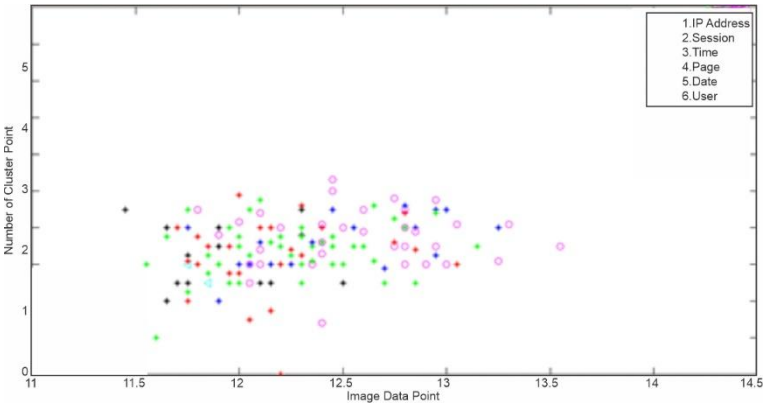

Figure 11: window show that the output image data point and here, input field of initial value of seed is 10 and when hitting on K-ACO technique button here loaded the dataset imagedataset2 and imported in our simulation GUI window.

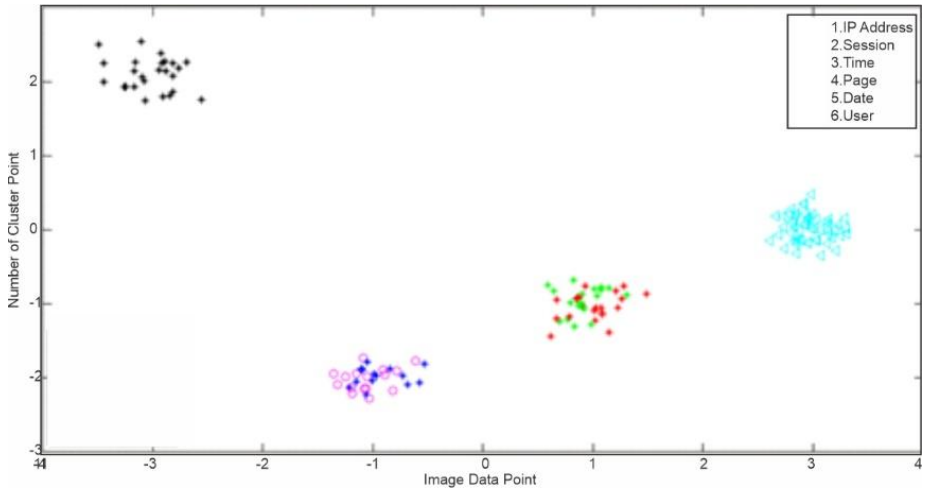

Figure 12: window show that output image data point and here, input field of initial value of seed is 10 and when hitting on K-PSO technique button here loaded the dataset imagedataset3 and imported in our simulation GUI window.

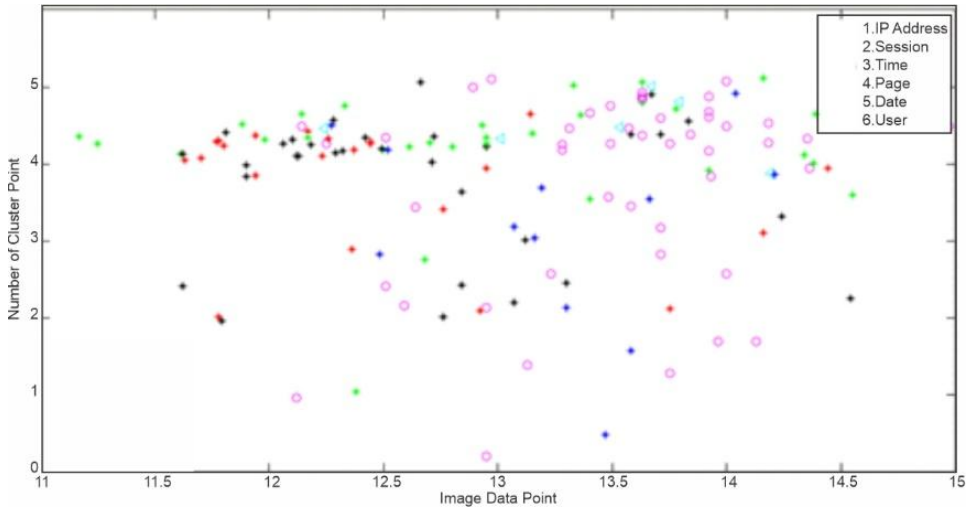

Figure 13: window show that output image data point and here, input field of initial value of seed is 10 and when hitting on FCM technique button here loaded the dataset imagedataset4 and imported in our simulation GUI window.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

Table 1: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 5 and Logdata1 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	5	7	6	4	Logdata1 dataset
Computed Error	5.72	8.22	6.54	4.79	
Standard Deviation	1.68	2.72	2.02	1.4	

Table 2: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 5 and Logdata2 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	10	9	9	7	Logdata2 dataset
Computed Error	9.57	7.25	8.64	6.59	
Standard Deviation	2.96	2.78	1.98	1.51	

Table 3: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 5 and Logdata3 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	14	16	12	11	Logdata3 dataset
Computed Error	16.85	17.64	15.24	13.45	
Standard Deviation	1.96	1.73	1.80	1.69	

Table 4: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 5 and Logdata4 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	18	17	16	15	Logdata4 dataset
Computed Error	18.67	18.07	17.49	16.55	
Standard Deviation	3.01	2.84	2.81	2.66	

Table 5: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 8 and Winedata1 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	6	7	8	5	Winedata1 dataset
Computed Error	8.94	7.44	8.25	6.47	
Standard Deviation	1.89	1.37	1.84	1.03	

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

Table 6: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 8 and Winedata2 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	18	16	16	12	Winedata2 dataset
Computed Error	5.88	6.18	8.95	5.09	
Standard Deviation	2.51	2.11	2.84	2.21	

Table 7: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 8 and Winedata3 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	10	11	10	8	Winedata3 dataset
Computed Error	23.55	20.78	22.34	18.27	
Standard Deviation	1.05	1.11	1.27	0.98	

Table 8: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 8 and Winedata4 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	20	20	22	18	Winedata4 dataset
Computed Error	10.84	11.84	12.51	10.25	
Standard Deviation	2.25	2.69	2.54	2.01	

Table 9: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 10 and Imagedata1 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	8	9	6	5	Imagedata1 dataset
Computed Error	12.84	11.35	11.67	10.37	
Standard Deviation	1.78	1.89	1.97	1.18	

Table 10: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 10 and Imagedata2 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

Number of Iteration	15	18	16	14	Imagedata2 dataset
Computed Error	9.85	9.78	8.47	8.04	
Standard Deviation	1.84	2.59	1.74	1.47	

Table 11: Comparative Analysis of K-Means, KACO, FCM and KPSOtechniques using value of seed is 10 and Imagedata3 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	9	8	7	6	Imagedata3 dataset
Computed Error	12.58	12.59	14.57	11.66	
Standard Deviation	3.05	3.22	2.48	2.11	

Table 12: Comparative Analysis of K-means, K-ACO, K-PSO and Proposed technique using value of seed is 10 and Imagedata4 dataset with given parameters Number of Iteration, Computed Error, Standard Deviation.

	K-Means[5]	KACO[2]	FCM[6]	KPSO	Dataset
Number of Iteration	12	11	10	9	Imagedata4 dataset
Computed Error	15.02	16.35	15.66	14.08	
Standard Deviation	1.98	1.87	1.99	1.56	

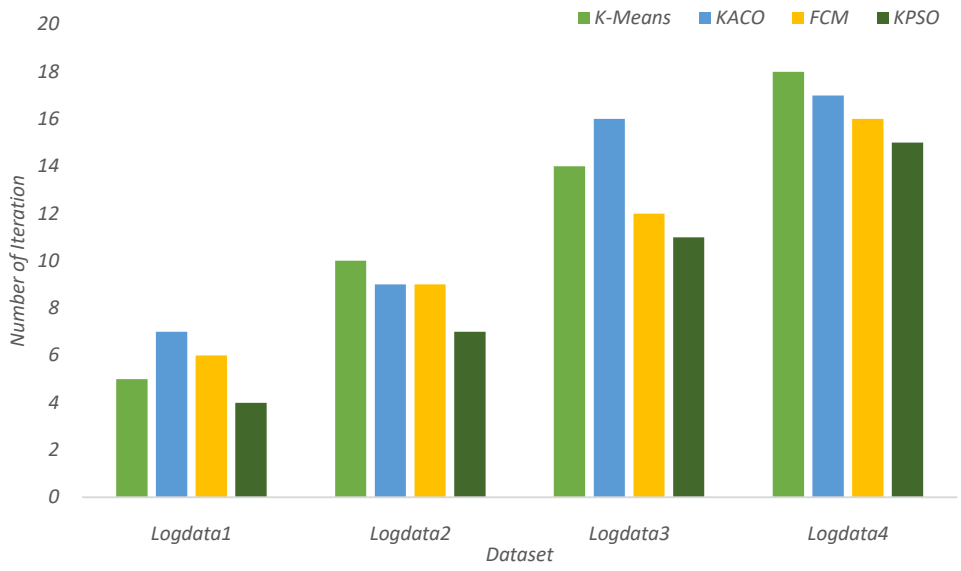

Figure 14: Comparative performance analysis of number of iterations using K-Means, KACO, FCM and KPSOtechniques with logdata1, logdata2, logdata3, logdata4 datasets.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

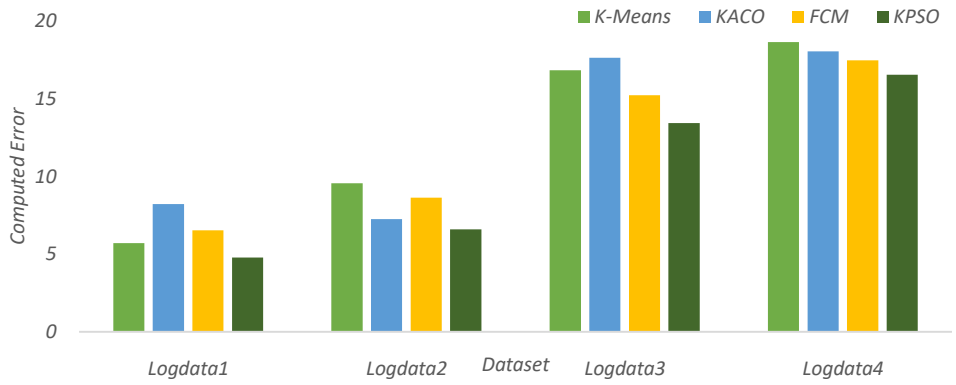

Figure 15: Comparative performance analysis of computed error using K-Means, KACO, FCM and KPSOtechniqueswith logdata1, logdata2, logdata3, logdata4 datasets.

Figure 16: Comparative performance analysis of standard deviation using K-Means, KACO, FCM and KPSOtechniqueswith logdata1, logdata2, logdata3, logdata4 datasets.

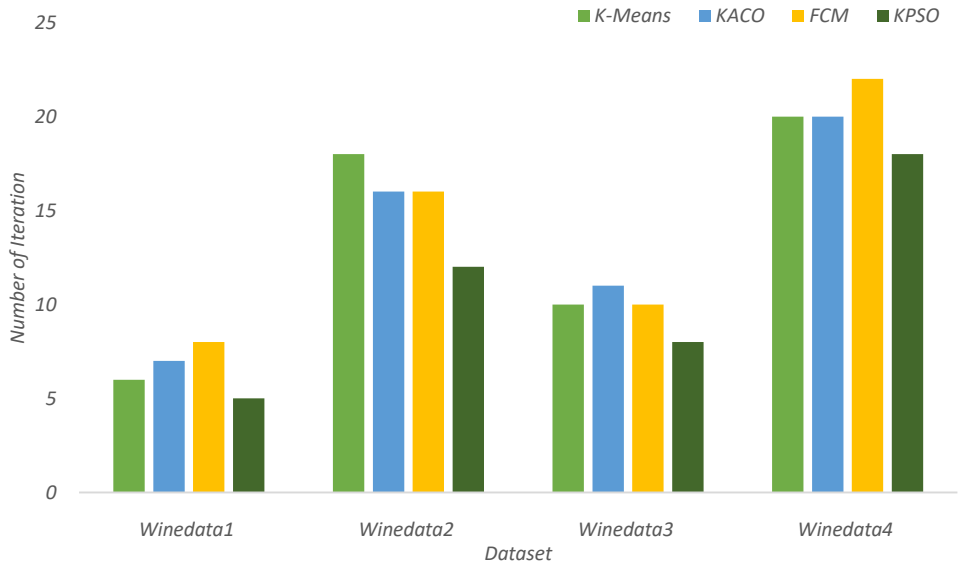

Figure 17: Comparative performance analysis of number of iterations using K-Means, KACO, FCM and KPSOtechniqueswith winedata1, winedata2, winedata3, winedata4 datasets.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

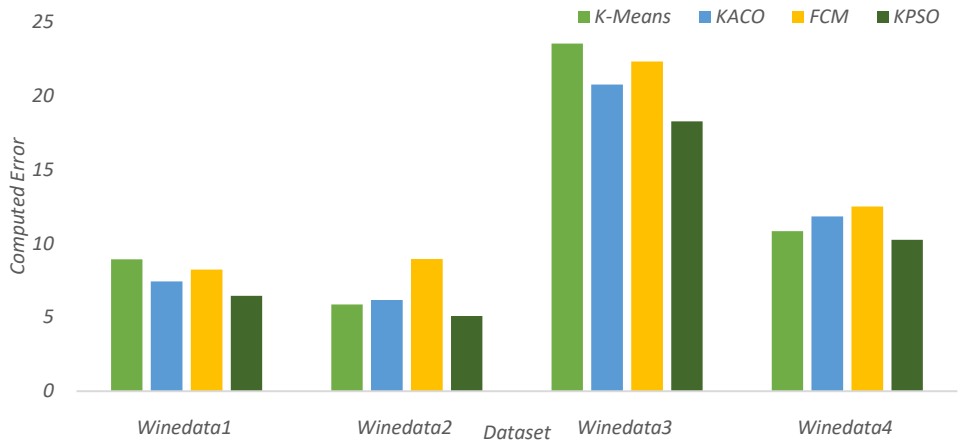

Figure 18: Comparative performance analysis of computed error using K-Means, KACO, FCM and KPSOtechniques with winedata1, winedata2, winedata3, winedata4 datasets.

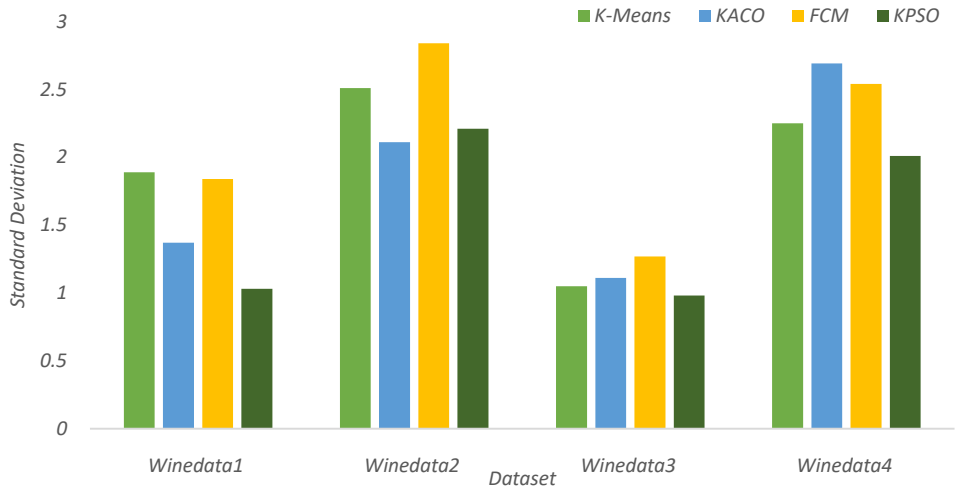

Figure 19: Comparative performance analysis of standard deviation using K-Means, KACO, FCMand KPSOtechniques with winedata1, winedata2, winedata3, winedata4 datasets.

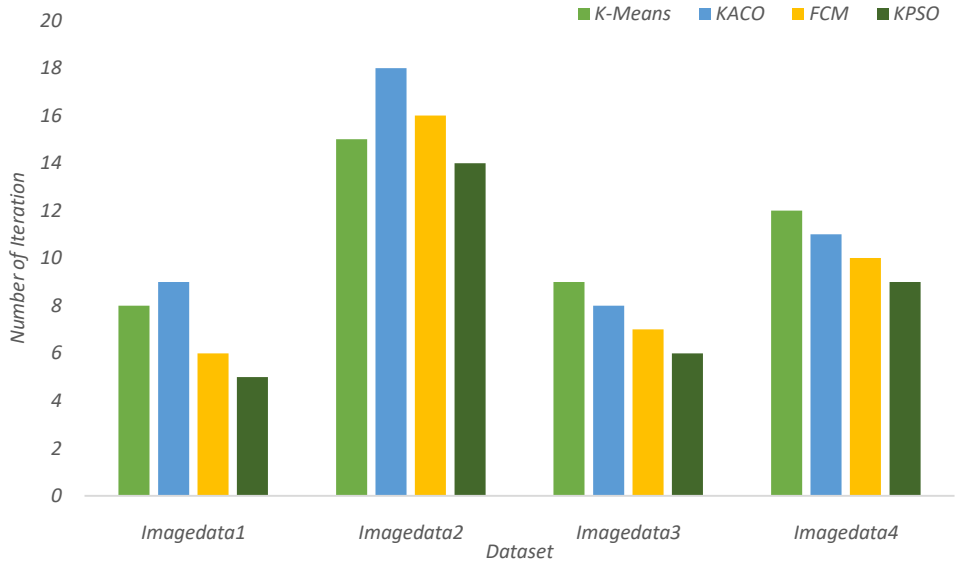

Figure 20: Comparative performance analysis of number of iterations using K-Means, KACO, FCM and KPSOtechniqueswith imagedata1, imagedata2, imagedata3, imagedata4 datasets.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

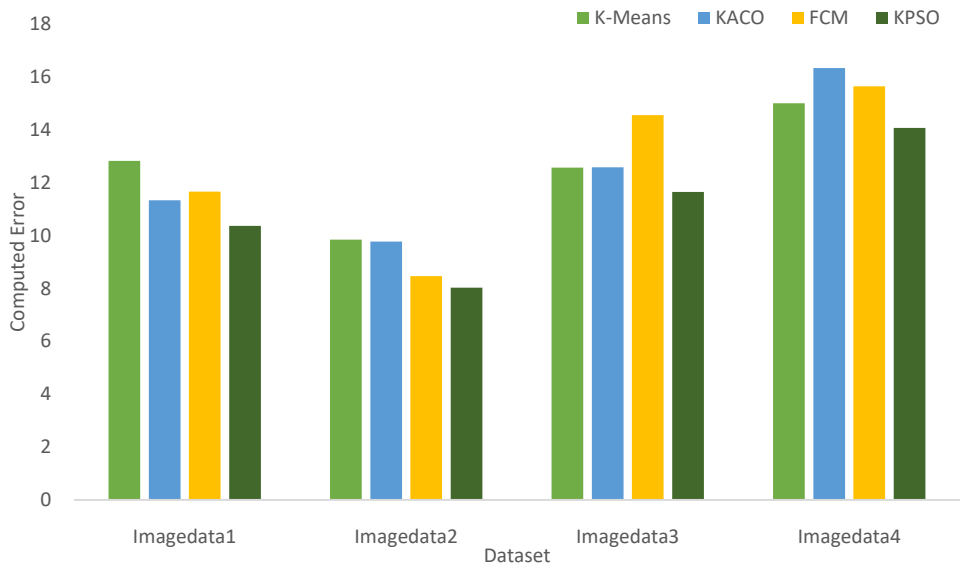

Figure 21: Comparative performance analysis of computed error using K-Means, KACO, FCM and KPSOtechniqueswith imagedata1, imagedata2, imagedata3, imagedata4 datasets.

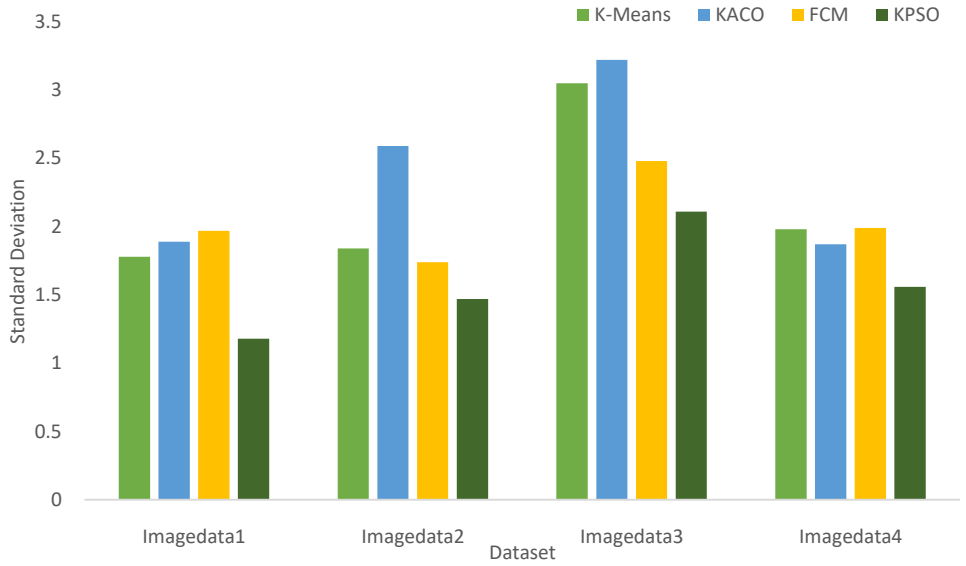

Figure 22: Comparative performance analysis of standard deviation using K-Means, KACO, FCM and KPSOtechniqueswith imagedata1, imagedata2, imagedata3, imagedata4 datasets.

5. CONCLUSION & FUTURE WORK

The modification of the K-means algorithm with PSO and ACO improves the extensive dataset clustering process's efficiency. The problem of data mapping and minimization of objective function resolve by two swarm intelligence algorithms. The modified algorithms also maintain the complexity of time and reduce the computational error due to the minimization of sample data's standard deviation. The modified algorithm tested with one real-time dataset and two synthetic datasets. The performance of the algorithm indicates that the modification of the algorithm achieves the objective of clustering. The ant colony optimization algorithm mapped with ant space, and the selection of centers controlled with pheromone value validated the cluster's index and spied with the intermediate cluster. Despite ant colony optimization, particle swarm optimization also improves the clustering algorithm's efficiency—the velocity and direction of particle move according to their data points. The selection of fitness function with mapped data decreases the iteration of cluster formation. The modified clustering algorithm also validated the cluster contains data. The modified algorithm also compares with the existing algorithm FCM. The validation of results with three parameters number of iterations, standard

deviation and computer errors. The process of clustering algorithms applied different distance formula for similarity measure. The size of data is directly proportional to time complexity. The high dimension of data increases the time complexity and break the limit of the linear order of time in the K-means algorithm. The dimension complexity of data handling with hybrid swarm intelligence algorithm in future.

REFERENCES

[1]. Ghorpade-Aher, Jayshree, and Vishakha A. Metre. "Clustering multidimensional data with PSO based algorithm." arXiv preprint arXiv:1402.6428 (2014).
[2]. Srinivasan, Thenmozhi, and BalasubramaniePalanisamy. "Scalable clustering of high-dimensional data technique using SPCM with ant colony optimization intelligence." The Scientific World Journal 2015 (2015).
[3]. Li, Yaping. "Glowworm Swarm Optimization Algorithm-and K-Prototypes Algorithm-Based Metadata Tree Clustering." Mathematical Problems in Engineering 2021 (2021).
[4]. Gong, Xueyuan, Liansheng Liu, Simon Fong, Qiwen Xu, Tingxi Wen, and Zhihua Liu. "Comparative research of swarm intelligence clustering algorithms for analyzing medical data." IEEE Access 7 (2019): 137560-137569.
[5]. El-Khatib, Samer, Sergey Rodzin, and Yuri Skobtsov. "Investigation of optimal heuristical parameters for mixed ACO-k-means segmentation algorithm for MRI images." In Information Technologies in Science, Management, Social Sphere and Medicine. Atlantis Press, 2016.
[6]. Dutta, Ashit Kumar. "Fuzzy Clustering with Particle Swarm Intelligence for Large Dataset Classification." TEM Journal 7, no. 4 (2018): 738-743.
[7]. Xiao, Xingxing, and Haining Huang. "A Clustering Routing Algorithm Based on Improved Ant Colony Optimization Algorithms for Underwater Wireless Sensor Networks." Algorithms 13, no. 10 (2020): 250.
[8]. Abualigah, Laith, Amir H. Gandomi, Mohamed Abd Elaziz, Abdelazim G. Hussien, Ahmad M. Khasawneh, Mohammad Alshinwan, and Essam H. Houssein. "Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis." Algorithms 13, no. 12 (2020): 345.
[9]. Cui, Xiaohui, Thomas E. Potok, and Paul Palathingal. "Document clustering using particle swarm optimization." In Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005., pp. 185-191. IEEE, 2005.
[10]. Abualigah, Laith, Amir H. Gandomi, Mohamed Abd Elaziz, Husam Al Hamad, Mahmoud Omari, Mohammad Alshinwan, and Ahmad M. Khasawneh. "Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering." Electronics 10, no. 2 (2021): 101.
[11]. Al-Baity, Heyam, SouhamMeshoul, Ata Kaban, and Lilac Al Safadi. "Quantum behaved particle swarm optimization for data clustering with multiple objectives." In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 215-220. IEEE, 2014.
[12]. Aishwarya, M. S., H. Bhargavi, Kavya R. Jyothi Narayan, and Shailesh Shetty Harisha. "Handling Big Data Analytics Using Swarm Intelligence."
[13]. Gong, Congcong, Haisong Chen, Weixiong He, and Zhanliang Zhang. "Improved multi-objective clustering algorithm using particle swarm optimization." PloS one 12, no. 12 (2017): e0188815.
[14]. Abraham, Ajith, Swagatam Das, and Sandip Roy. "Swarm intelligence algorithms for data clustering." In Soft computing for knowledge discovery and data mining, pp. 279-313. Springer, Boston, MA, 2008.
[15]. Farmer, Jocelyn R., Mei-Sing Ong, Sara Barmettler, Lael M. Yonker, Ramsay Fuleihan, Kathleen E. Sullivan, Charlotte Cunningham-Rundles, Jolan E. Walter, and USIDNET Consortium. "Common variable immunodeficiency non-infectious disease endotypes redefined using unbiased network clustering in large electronic datasets." Frontiers in immunology 8 (2018): 1740.

[16]. Salem, Semeh Ben, Sami Naouali, and ZiedChtourou. "A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach." Computers & Electrical Engineering 68 (2018): 463-483.
[17]. Patibandla, RSM Lakshmi, and N. Veeranjaneyulu. "Survey on clustering algorithms for unstructured data." In Intelligent Engineering Informatics, pp. 421-429. Springer, Singapore, 2018.
[18]. Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 132-149. 2018.
[19]. Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "A novel clustering method using enhanced grey wolf optimizer and mapreduce." Big data research 14 (2018): 93-100.
[20]. Fränti, Pasi, and Sami Sieranoja. "K-means properties on six clustering benchmark datasets." Applied Intelligence 48, no. 12 (2018): 4743-4759.
[21]. Steinegger, Martin, and Johannes Söding. "Clustering huge protein sequence sets in linear time." Nature communications 9, no. 1 (2018): 1-8.
[22]. Tardioli, Giovanni, Ruth Kerrigan, Mike Oates, James O'Donnell, and Donal P. Finn. "Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach." Building and Environment 140 (2018): 90-106.
[23]. Rida, Mohamad, Abdallah Makhoul, Hassan Harb, David Laiymani, and Mahmoud Barhamgi. "EK-means: A new clustering approach for datasets classification in sensor networks." Ad Hoc Networks 84 (2019): 158-169.
[24]. Xu, Xiao, Shifei Ding, and Zhongzhi Shi. "An improved density peaks clustering algorithm with fast finding cluster centers." Knowledge-Based Systems 158 (2018): 65-74.
[25]. Wolf, F. Alexander, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells." Genome biology 20, no. 1 (2019): 1-9.
[26]. Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data." Nature Reviews Genetics 20, no. 5 (2019): 273-282.
[27]. Duò, Angelo, Mark D. Robinson, and Charlotte Soneson. "A systematic performance evaluation of clustering methods for single-cell RNA-seq data." F1000Research 7 (2018).
[28]. Huang, Wenzhun, Haoxiang Wang, Yucheng Zhang, and Shanwen Zhang. "A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop." Cluster Computing 22, no. 6 (2019): 13077-13084.
[29]. Azad, Ariful, Georgios A. Pavlopoulos, Christos A. Ouzounis, Nikos C. Kyrpides, and Aydin Buluç. "HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks." Nucleic acids research 46, no. 6 (2018): e33-e33.
[30]. Du, Mingjing, Shifei Ding, and Yu Xue. "A robust density peaks clustering algorithm using fuzzy neighborhood." International Journal of Machine Learning and Cybernetics 9, no. 7 (2018): 1131-1140.
[31]. Rodriguez, Mayra Z., Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. "Clustering algorithms: A comparative approach." PloS one 14, no. 1 (2019): e0210236.
[32]. Shakeel, P. Mohamed, S. Baskar, VR Sarma Dhulipala, and Mustafa Musa Jaber. "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering." Health information science and systems 6, no. 1 (2018): 16.
[33]. Sinha, Debajyoti, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. "dropClust: efficient clustering of ultra-large scRNA-seq data." Nucleic acids research 46, no. 6 (2018): e36-e36.
[34]. Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis." Genome biology 19, no. 1 (2018): 15.
[35]. Yang, Yan, and Hao Wang. "Multi-view clustering: A survey." Big Data Mining and Analytics 1, no. 2 (2018): 83-107.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6348 - 6367
Received 25 April 2021; Accepted 08 May 2021.

[36]. Liu, Qidong, Ruisheng Zhang, Rongjing Hu, Guangjing Wang, Zhenghai Wang, and Zhili Zhao. "An improved path-based clustering algorithm." Knowledge-Based Systems 163 (2019): 69-81.
[37]. Patibandla, RSM Lakshmi, and N. Veeranjaneyulu. "Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria." Arabian Journal for Science and Engineering 43, no. 8 (2018): 4379-4390.
[38]. Riaz, Sumbal, Mehvish Fatima, Muhammad Kamran, and M. Wasif Nisar. "Opinion mining on large scale data using sentiment analysis and k-means clustering." Cluster Computing 22, no. 3 (2019): 7149-7164.
[39]. Zou, Quan, Gang Lin, Xingpeng Jiang, Xiangrong Liu, and Xiangxiang Zeng. "Sequence clustering in bioinformatics: an empirical study." Briefings in bioinformatics 21, no. 1 (2020): 1-10.
[40]. Deng, Tingquan, Dongsheng Ye, Rong Ma, Hamido Fujita, and LvnanXiong. "Low-rank local tangent space embedding for subspace clustering." Information Sciences 508 (2020): 1-21.