

A Robust Fuzzy Approach For Gene Expression Data Clustering

Meskat Jahan (✉ meskat.jahan@gmail.com)

Comilla University <https://orcid.org/0000-0002-1037-5282>

Mahmudul Hasan

Saitama University

Research Article

Keywords: FCM, K-Means, Fuzzy Clustering, Meskat-Mahmudul (MM) clustering, MM Extended K-Means (MMK) Clustering, Clustering algorithm, Data mining

Posted Date: June 18th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-547452/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A Robust Fuzzy Approach for Gene Expression Data Clustering

Authors:

1. Meskat Jahan*,
Department of Computer Science and Engineering, Comilla University,
Cumilla-3506, Bangladesh.
Email: meskat.jahan@gmail.com, Phone: +8801912194688
2. Mahmudul Hasan
Department of Computer Science and Engineering
Comilla University, Cumilla -3506, Bangladesh
Phone: Cell: +81-70-4442-6086(JP)
Email: [Personal](#) [Official\(BD\)](#), [Official\(JP\)](#)
Email: mhasanraju@gmail.com

A Robust Fuzzy Approach for Gene Expression

Data Clustering

Abstract In the big data era, clustering is one of the most popular data mining method. The majority of clustering algorithms have complications like automatic cluster number determination, poor clustering precision, inconsistent clustering of various datasets and parameter-dependent etc. A new fuzzy autonomous solution for clustering named Meskat-Mahmudul (MM) clustering algorithm proposed to overcome the complexity of parameter-free automatic cluster number determination and clustering accuracy. MM clustering algorithm finds out the exact number of clusters based on Average Silhouette method in multivariate mixed attribute dataset, including real-time gene expression dataset and dealt missing values, noise and outliers. MM Extended K-Means (MMK) clustering algorithm is an enhancement of the K-Means algorithm, which serves the purpose for automatic cluster discovery and runtime cluster placement. Several validation methods used to evaluate cluster and certify optimum cluster partitioning and perfection. Some datasets used to assess the performance of the proposed algorithms to other algorithms in terms of time complexity and clustering efficiency. Finally, MM clustering and MMK clustering algorithms found superior over conventional algorithms.

Keywords FCM; K-Means; Fuzzy Clustering ; Meskat-Mahmudul (MM) clustering ; MM Extended K-Means (MMK) Clustering; Clustering algorithm ; Data mining

1 Introduction

Clustering is a process of recognizing similar objects and separated non-identical objects. Cluster analysis divided into two basic groups like hard and fuzzy clustering methods. In addition, other types of clustering methods are distribution-based clustering, Connectivity-based clustering and Constraint-based clustering. Many algorithms developed based on these categorizations. DIANA (Divisive Analysis) or AGNES (Agglomerative Nesting) methods used in Hierarchical clustering. However, Hierarchical clustering does not well suited for large dataset. Density-based clustering or Model-based methods fall into Distribution-based categories where methods consider density instead of distances. Using density reachability and density connectivity to create cluster by separate and differentiate region of varying density degree Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[9] algorithm is formed. Yet it is a slow process and required a complex algorithm to implement. Famous Centroids-based hard clustering is K-Means[11], K-Medians, K-Medoids[10] and some extended version of the K-means are Isodata algorithm, forgy algorithms etc. Fuzzy C-Means, Fuzzy K-NN are popular soft clustering schemas. Still, we have to provide the number of cluster centroids before conducting cluster for partitioning methods. K-means[11], K-medoids[10] and PAM[18] are sample algorithms of partitioning clustering. The mean value of items in a cluster used to measure similarity in the K-means algorithm. This algorithm aims to divide the dataset into K clusters, where K is a predefined number. Each cluster in each round made up of the points closest to the corresponding reference point and the centroid of each cluster used as the next round's reference point. Because of these iterations, the reference point chosen closer to the real cluster centroid and the clustering

effect improves. Since the PAM [17] algorithm evaluates all objects, the central point of each class is considered. For many combinations, the accuracy of the clustering results is calculated. In a small dataset, the PAM approach works fine but the result is not optimal with giant dataset. One of the biggest weaknesses of density and demarcation approaches is the difficulty in defining the clusters prior to certain algorithms including, in particular, the form of its clusters like most algorithms.

Therefore, we can summarize the drawbacks of partitioning algorithms as follow: (1) Number of clusters K determination; (2) The PAM approach works well for a small dataset, but it is not suitable for a large dataset; (3) If the data contains noise and outliers, it is prone to errors. To solve the above-stated problems, we proposed two approaches named Meskat-Mahmudul clustering (MM clustering) and MM Extended K-Means (MMK) clustering algorithm for non-parametric automatic cluster number determination and clustering accordingly. In MM clustering, we will use Average Silhouette based cluster number determination techniques and for clustering, we adopt the FCM algorithm. In addition, in MMK clustering, we used the K-Means clustering schema for cluster partitioning and the Average Silhouette method for cluster number determination. The dataset used in this paper includes Wisconsin Breast Cancer, Iris, Leukemia, Motor Cycle etc. These datasets all come from the UCI database[7]. Seven validation criteria used for evaluating outcomes. Results show that the proposed methods more enhanced compared with the other state-of-art algorithms.

In brief, The contribution of this work is as follows: (1) We present an autonomous solution for clustering which includes automated cluster number determination using the Average Silhouette method and clustering accordingly; (2) We incorporate a preprocessing mechanism into the proposed algorithms to manage unnecessary noise and outliers in the dataset; (3) Proposed methods are parameter-free and do not require any manual intervention in the process;

2 Literature Review

Tao Lei et al. [17] proposed a significantly fast and robust Fuzzy C-Means clustering algorithm. They introduced local spatial information to an objective function of the FCM algorithm for image segmentation based on morphological reconstruction for image detail-preservation and noise-immunity and membership filtering (FRFCM). Singa and Yang [24] highlight a novel U K-means clustering algorithm that automatically finds the optimal number of clusters without providing any parameter selection and initialization. Dhanachandra et al. [5] provide a methodology for image segmentation by using K-means and subtractive clustering algorithm. The initial centers produced using a subtractive cluster and these centers used in the k-means algorithm to segment the image. Finally, a medial filter added to the segmented image to eliminate any unwanted regions. Cal et al. [2] discovered that the cluster centers cannot be calculated automatically and that the selected cluster centers which fall into a local optimum and the random selection of the parameter cut-off distance d_c value when they examined the quick search and find of density peaks clustering (DPC) algorithm[22]. PDPC [2] clustering algorithm is proposed to address these issues. Particle swarm optimization (PSO)[8] is implemented because of its simple idea and strong global search ability, which allows it to find the best solution in a short amount of time.

Jinyin et al. [16] proposed cluster center fast determination clustering algorithm which can adapt the parameter of d_c and automatically get the optimal density radius and select the

cluster center in the clustering process. Nayak et al. [15] gather some drawbacks of FCM algorithm including (1) point-based membership [3], (2) Information loss when calculating similarity degree, (3) slow convergence rate, (4) sensitivity to initialization etc. Hou et al. [12] propose clustering as image segmentation techniques using a new parameter-free algorithm based on DSets-DBSCAN. Reddy et al. [19] using a hybrid Fuzzy ANN-based classifier for the diagnosis of diabetes. Also, Reddy et al. [20] forecasting heart disease by applying the Hybrid OFBAT combined with a Rule-Based Fuzzy Logic Model. Another work on heart disease classification presented by Reddy et al. [21]. Table 1 provides a rundown of the preceding discussion. Chen [4] developed the DC-MDACC algorithm which divides the dataset into graded dominant, numerical dominant and mixed attribute types. MH algorithm [14] also provides the solution for automatic cluster detection and identification. Jahan et al. [13] shows a comparative analysis of some traditional algorithms.

Table 1. Summary of Literature Review

Authors (Year)	Focus	Findings	Gaps/ Remarks
Tao Lei et al.(2018) [17]	Fast and Robust FCM algorithm	FRFCM is faster and more robust than FCM	the number of clusters is set experimentally in FRFCM
Singa and Yang(2020)[24]	Unsupervised K-Means Clustering Algorithm	Automatically finding optimal cluster number	After cluster number selection they set $\beta = 0$ where $\beta \in [0, 1]$ is used to control competitions.
Dhanachandra et al. (2015)[6]	Image Segmentation	Generate initial centers for k-means & medial filter to remove unwanted region	By using morphological operation output quality may improve
Cai et al. (2020)[2]	DPC and PSO	New fitness function and a method to calculate d_c parameter	Time complexity $O(n^2)$ where K-Means is $O(n)$
Jinyin et al. (2017)[16]	Cluster center fast determination	d_c parameter adaptation, optimal density radius calculation and automatically cluster center selection	Can't analyze and deal with noise points and different density level datasets
Hou et al. (2016)[12]	Image segmentation	Fusing dominant sets & DBSCAN algorithm.	Doesn't exist a unique image segmentation
Reddy et al. (2017)[19]	Hybrid Fuzzy ANN classifier	FFBAT-ANN reduced feature and classify disease using LPP	Optimized algorithms may improve classification accuracy & performance
Reddy et al. (2016)[20]	Predicting heart disease	RBFL algorithm based on FFBAT & fuzzy classifier	The accuracy rate should be raised
Reddy et al. (2019)[21]	Fusing Fuzzy classifier & hybrid genetic algorithm	Rough set for feature reductions and AGAFL predict the heart disease.	Adaptive bee colony, whale optimization & antlion algorithm etc. may greatly enhance this work.

3 Proposed Algorithms

Detaching heterogeneous data and assemble them into a homogeneous set based on similarity is the goal of clustering. Earlier, partition-based clustering is dependent on pre-specified cluster number, which is one of the main drawbacks of them. Here, we propose a hybrid method for the automatic clustering of data. Optimum cluster number selection based on the Average Silhouette method is integrated into the proposed algorithms. Following are the proposed approaches:

Approach 1: Meskat-Mahmudul (MM) Clustering Algorithm

Approach 2: MM Extended K-Means (MMK) Clustering Algorithm

In the following sections, we described approaches.

Meskat-Mahmudul (MM) Clustering Algorithm

Let, $U_{i=\{1, \dots, n\}}$ be the dataset, a_i is average distance, b_i is minimum average distance, s_i is silhouette value, U' is the normalized data point, $mean$ is the arithmetic mean, N is the maximum cluster number determined by domain expert, l is iteration step, μ_{ij} is fuzzy membership function, D_{ij} is Euclidean distance, $\epsilon = 1 \times 10^{-6}$ and $m=1$ to ∞ is the fuzziness index.

1. for $i=1$ to n ,

$$\text{Calculate } U' = \frac{U_i - \bar{U}}{\sigma} \quad /* \text{normalize dataset using z-score method} */$$

2. for $i=1$ to N ,

$$\text{get, } s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} [23] \text{ and determine, } \bar{s}_u = \text{mean} (s_i)$$

3. obtain average silhouette value,

$$\text{avg_silhouette} = \text{mean} (\bar{s}_u)$$

4. find the maximum average silhouette values,

$$x_max = \max (\bar{s}_u)$$

5. assign cluster number, $k = x_max$

6. for $j=1$ to k , compute fuzzy membership [25],

$$\mu_{ij}^{(l)} = 1 / \sum_{p=1}^k (D_{ij} - D_{ip})^{(2/m-1)};$$

7. for $j=1$ to k , calculate the fuzzy centers [2],

$$V_{ij} = \frac{(\sum_{i=1}^n (\mu_{ij})^m x_i)}{(\sum_{i=1}^n (\mu_{ij})^m)}$$

8. repeat step 6 to 7, until $\| U^{(l)} - U^{(l-1)} \| < \epsilon$

MM Extended K-Means Clustering Algorithm

Let, $U_{i=\{1, \dots, n\}}$ be the dataset, a_i is average distance, b_i is minimum average distance, s_i is silhouette value, U' is the normalized data point, $mean$ is the arithmetic mean, N is the maximum cluster number determined by a domain expert, v is cluster centers, k_i is the number of the data element in the i^{th} cluster.

1. for $i=1$ to n ,

$$\text{Calculate } U' = \frac{U_i - \bar{U}}{\sigma} \quad /* \text{normalize dataset using z-score method} */$$

2. for $i=1$ to N ,

$$\text{get, } s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} [23] \text{ and determine, } \bar{s}_u = \text{mean} (s_i)$$

3. obtain average silhouette value,

$$\text{avg_silhouette} = \text{mean} (\bar{s}_u)$$

4. find the maximum average silhouette values,

$$x_max = \max (\bar{s}_u)$$

5. assign cluster number, $k = x_max$

6. random selection of 'k' cluster centers

7. for $i=1$ to n ,

Calculate and search the minimal distance between a data element and cluster centers and assign the data point to the nearest cluster center

8. for $i=1$ to n ,

reappraise the cluster centers using

$$v_i = \frac{1}{k_i} \sum_{j=1}^{k_i} U_j$$

9. for $i=1$ to n ,

reappraise the data elements distance and update the cluster centers.

10. Repeat step 7 to 9, until no data elements will not reassigned.

3.1 Determination of Number of Clusters

We used optimization techniques for determining the number of clusters. Average silhouette values used for cluster number selection. First, we observe the silhouette values for over a range of cluster number. Then we obtain the average silhouette values from them. From the literature, we know that ranges of silhouette value vary from 1 to -1 and a high silhouette value is used to measure well-matched data element to its self-cluster and imperfectly matched to other clusters. The precise of clustering depends on high silhouette values for

most of the data element. The standard of clustering can be measured using average silhouette methods. It demonstrates how rigorously each data element lies within its cluster. For good clustering, a high average silhouette value recommended. The average silhouette values obtained from the mean silhouette values of various cluster number. From the average silhouette values, find the maximum average silhouette value and its corresponding cluster number. This cluster number is the desired optimized number of clusters. Below is the outcome of optimizing techniques for cluster number determination for all dataset. These shown in Fig 1. to Fig 4.

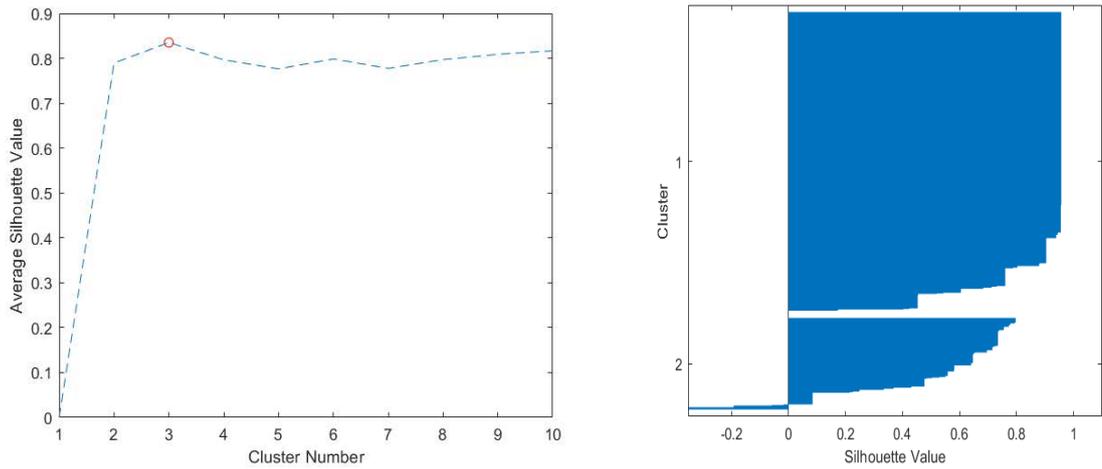


Fig 1. Determination of number of clusters for WBC Dataset

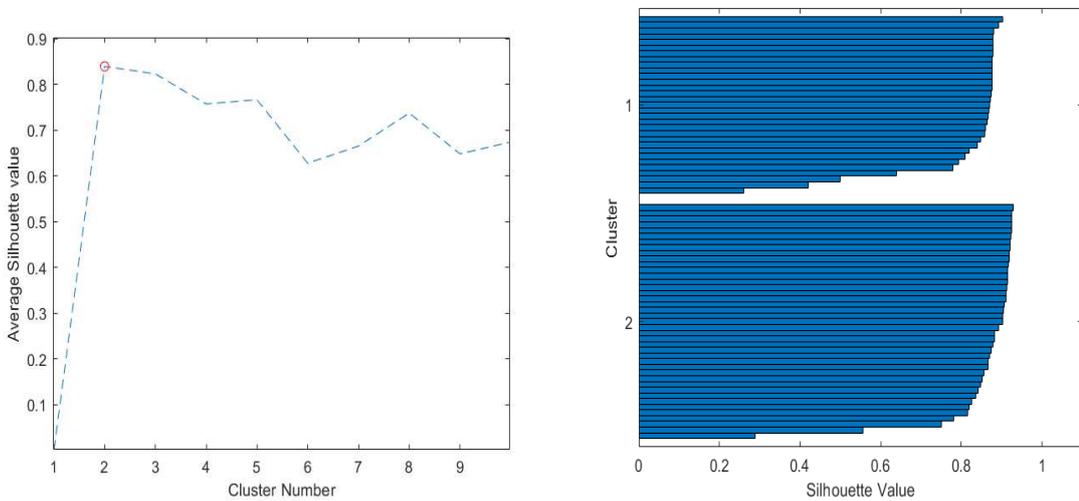


Fig 2. Determination of number of clusters for Leukemia Dataset

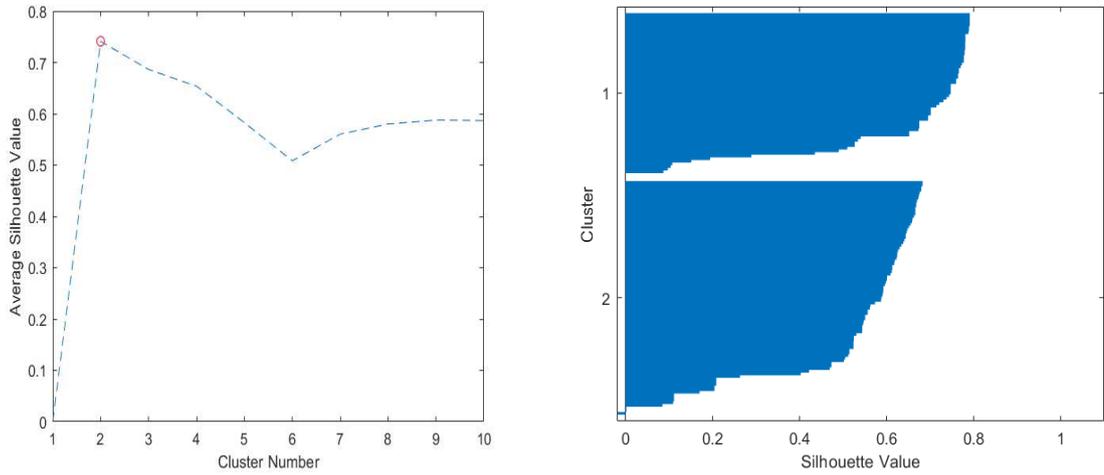


Fig 3. Determination of number of clusters for Iris dataset

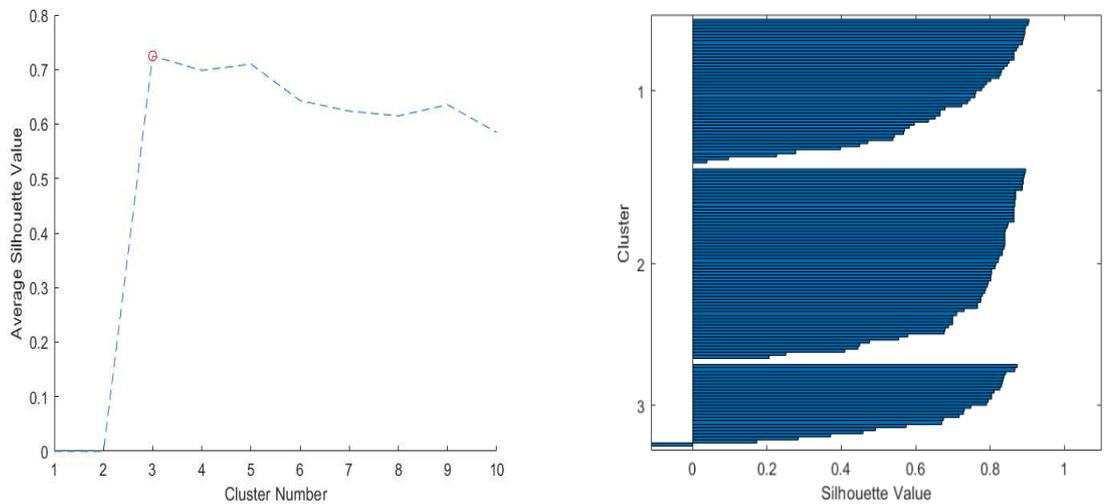


Fig 4. Determination of number of clusters for Motor Cycle Dataset

3.2 MM and MM Extended K-Means Clustering Algorithm

MM and MM Extended K-Means clustering algorithm both are sharing a common approach for determining the number of clusters where the MM algorithm is using a fuzzy approach and the MMK algorithm is a hard algorithm.

4 Performance Analysis among Clustering Algorithms

4.1 Algorithms Performance Evaluation based on Execution Time

A comparative analysis of algorithms performance based on execution time provided below. Algorithms are ordered from fuzzy algorithm to crisp algorithm. Here, algorithms are compared for every selected dataset altogether. Results are summarised in Fig 5. In Fig 5, we consider four dataset and seven algorithms performance is tested on them. In the WBC dataset, MM algorithms take acceptable execution time and the MMK algorithm also takes less time than the rest of the hardcore algorithms. In the Motorcycle dataset, all fuzzy

algorithms take the almost same time to execute and MMK algorithms need more time to execute because the MMK algorithm used an optimization mechanism for cluster number selection. In the leukaemia dataset both the MM algorithm and MMK algorithm execute in the shortest possible time than others. In the case of the Irises dataset, the MM algorithm takes less time and the MMK algorithms take acceptable time.

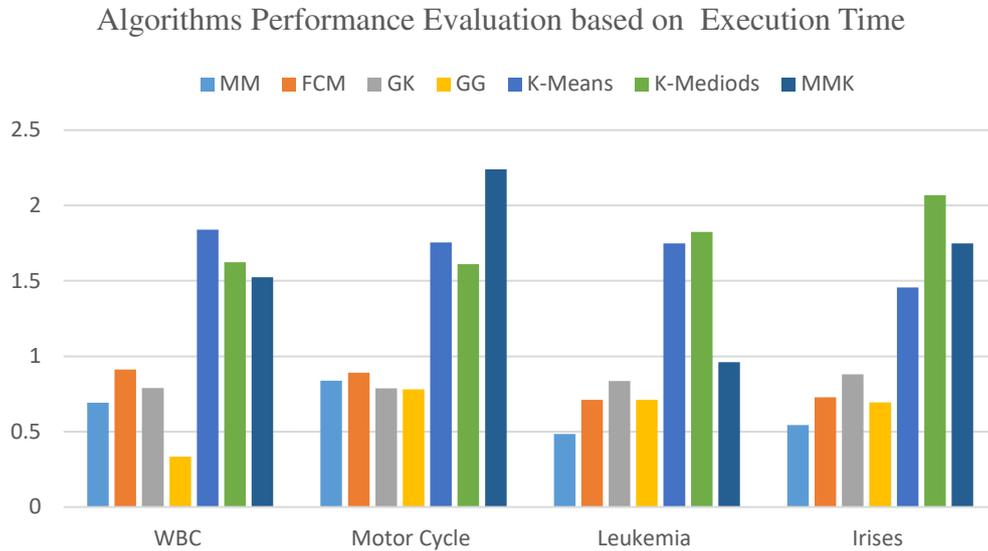
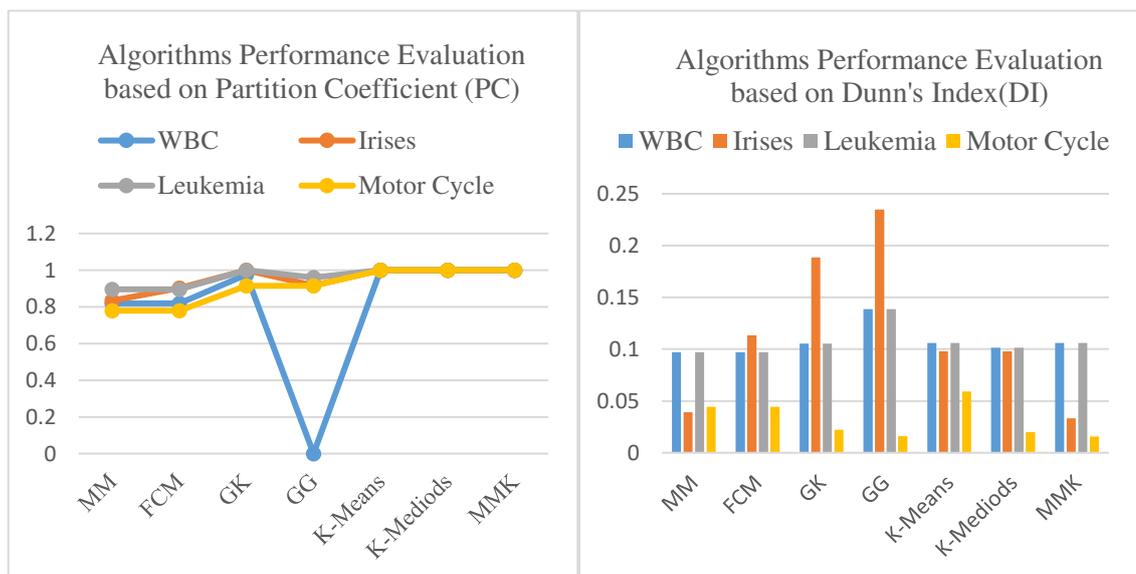


Fig 5. Comparative Study of Algorithms based on Execution Time

4.2 Algorithms Performance Evaluation based on Validation Indexes

Validation played an important role in assessing cluster efficiency. The efficiency of clusters has been tested by using screening indicators: Classification Entropy (CE), Partition



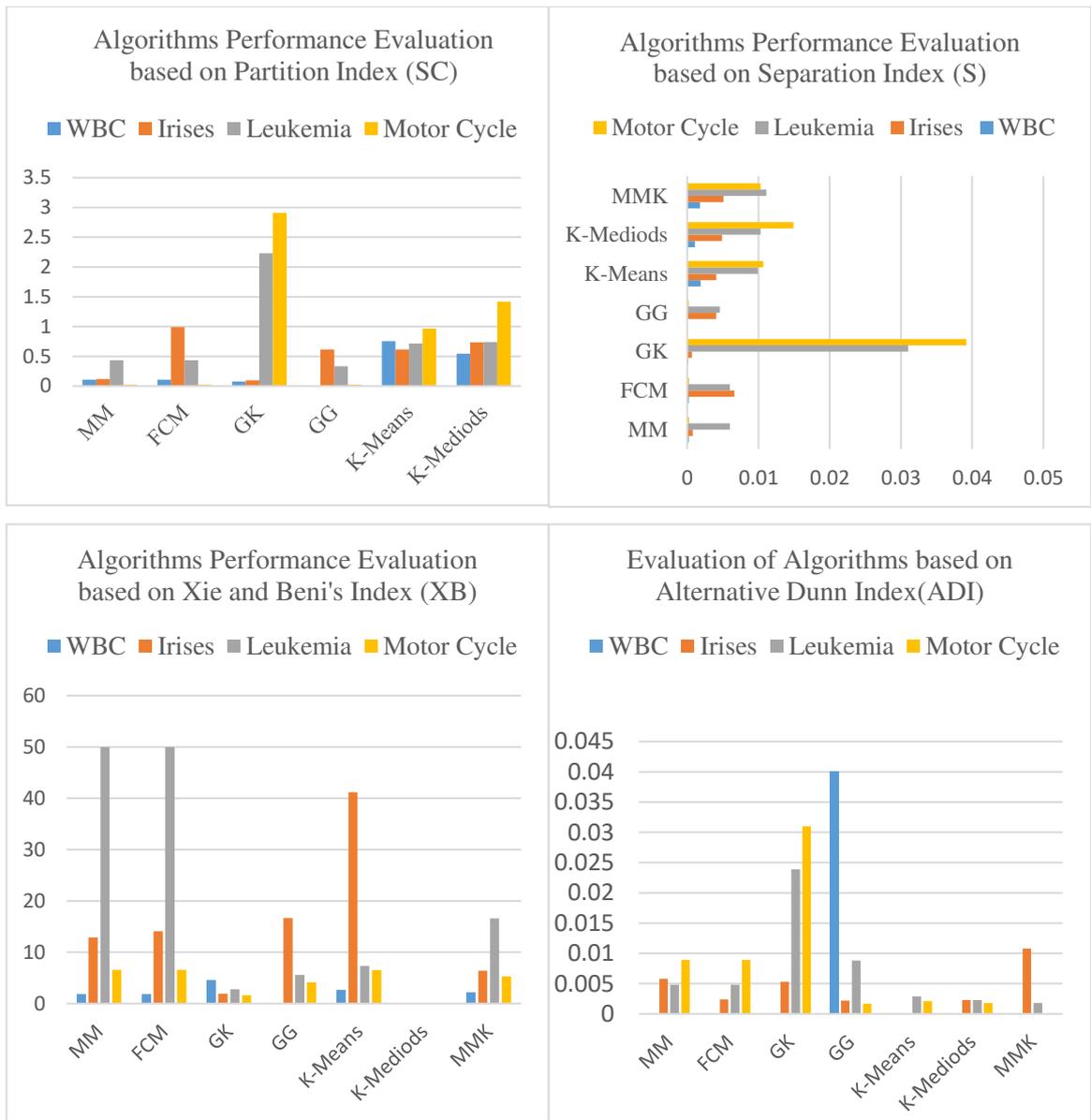


Fig 6. Evaluation of Algorithms based on PC, CE, SC, S, XB, DI, ADI for the selected dataset

Coefficient (PC), Xie and Beni's Index (XB), Partition Index (SC), and Separation Index (S), Dunn's Index (DI), Alternative Dunn Index (ADI).

For evaluation of cluster partitions, the minimum separation measurement done by the separation index (S), with the smallest values indicating a true optimal partition. Cluster partition coefficient (PC) indexing used to assess the quantity of overlapping between clusters and its high value ensures cluster accuracy. The Dunn Index (DI) is also an internal assessment parameter for evaluating which cluster well separated. Better clustering results demonstrated by higher DI and ADI values. The fuzziness of cluster partitions measured by CE. The strong output indicated by low CE and SC values. The XB index acknowledges the compactness of the whole cluster, with the smallest value suggesting the optimal number of clusters. Fig 6 demonstrate the algorithms performance evaluation based on the above-selected approaches.

5 RESULT & DISCUSSION

Validating the clustering algorithm is crucial for the gene expression dataset. Cluster validation is an important criterion for assessing a cluster's quality. Here, we analyzed the total results of the algorithms for the selected four datasets. Fig 7 illustrate the validation results for all selected algorithms for Wisconsin Breast Cancer (WBC) dataset. The upper limit of PC[1] indexing is one and the lower limit is $1/k$, where k is the cluster number. Therefore, the high value of PC is always desirable and for the WBC dataset K-Means, K-Medoids and MM Extended K-Means algorithms show the highest value and the rest of them gives high value. The small value of CE in MM and FCM algorithms indicates fuzzy partitions and others algorithm gives zero value, which recommended hard partitions in clustering algorithms. Low SC value of MM, FCM, GG and GK point out good clustering algorithms. Similarly, the low S index of FCM, MM, GG, GK algorithms stipulates satisfactory cluster partitioning. Although DI and ADI value remains same for all soft and hard algorithms. Upon analyzing the above-stated performance, we concluded that MM and MMK algorithms performance are quite satisfactory.

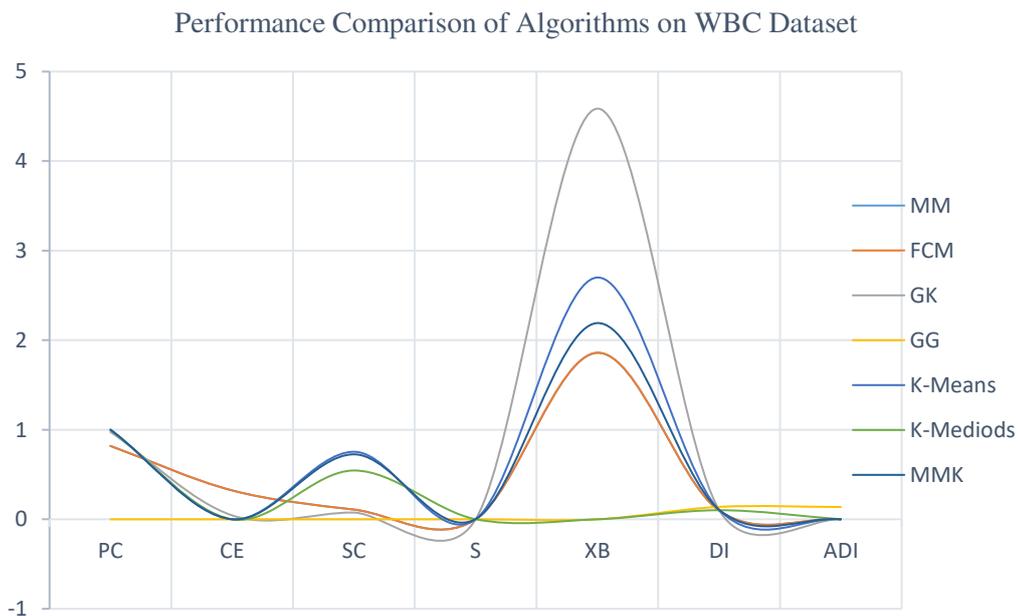


Fig 7. Algorithms Analysis on WBC Dataset

Performance Comparison of Algorithms on Irises Dataset

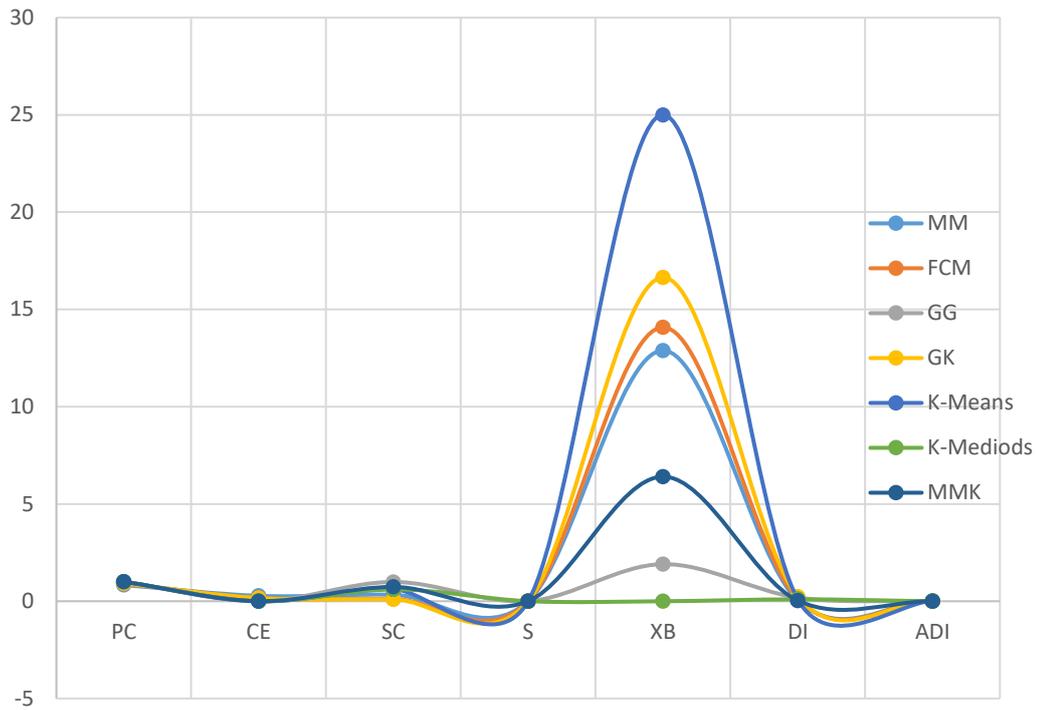


Fig 8. Algorithms Analysis on Irises Dataset

Fig 8 shows the validation results at a glance of the algorithms based on validation indexes for the Irises dataset. The PC indexing values of K-Means, K-Medoids and MM Extended K-Means algorithms are maximum which points to the rigid cluster partitioning and the average value of PC in FCM, GG and MM clustering algorithms represents fuzzy partitioning. As MM is a fuzzy clustering algorithm, it secures its fuzzy partitioning according to CE indexing along with FCM, GG and GK algorithms. Low values of SC, S and XB values provided by FCM, MM and GG algorithms. The highest value of ADI and DI provided by MMK, GG and MM clustering algorithms. We summarized that two developed algorithms (MM and MMK) performance are satisfiable in comparing other literature for the Irises dataset.

Fig 9. summarizes the algorithms performance analysis on the leukemia dataset. It shows that the performance of MM and MMK algorithms are meet the requirement of standard literature. Where MM algorithm secures its position by satisfying low value of CE, SC and S indexing. On the other hand, the MMK algorithm provides maximum value in PC indexing and lowest value of S indexing. GK algorithm gives the upper limit of PC, DI and ADI values. K-Means algorithm shows the highest PC value and lowest S values.

Performance Comparison of Algorithms on Leukemia Dataset

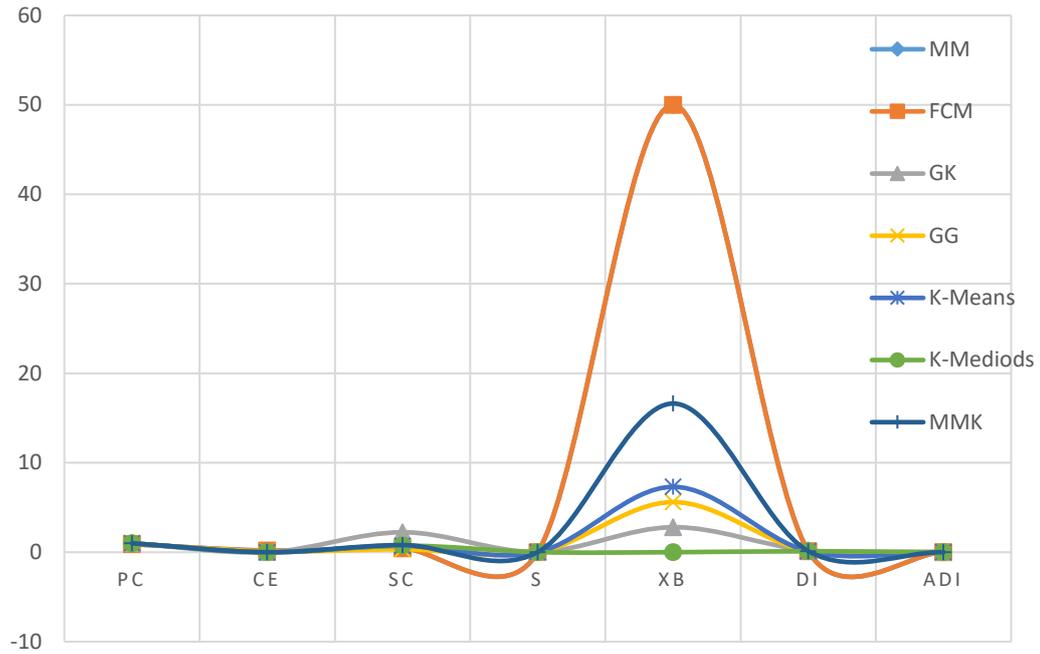


Fig 9. Algorithms Analysis on Leukemia Dataset

In Fig 10, MM algorithms give low value in SC and S indexing and top value in ADI indexing. GK algorithms show the lowest value of XB and CE indexing. K-Means, K-Medoids and MMK algorithms provide unity values in PC indexing.

Performance Comparison of Algorithms on Motor Cycle Dataset

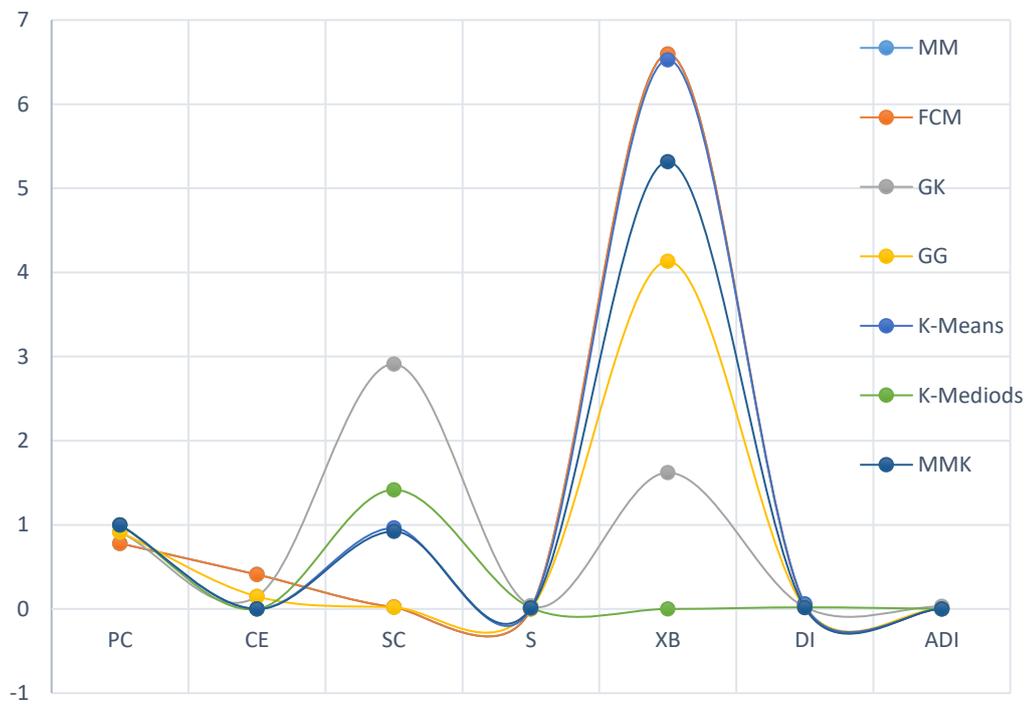


Fig 10. Algorithms Analysis on Motor Cycle Dataset

The MM clustering algorithm solves the problems include automatic cluster number identification, cluster placement at runtime and fuzzy partitioning. It specifies the number of clusters and separates them appropriately. For non-linear samples, the MM clustering methodology operates well. The MM Extended K-Means approach employs a dynamic cluster number decision to perform hard clustering. It fits well on a dataset that is distinguishable. Both the MM and MM Extended K-Means clustering algorithms choose the precise cluster number and deliver optimal partitioning and clustering based on all of the data. The MM clustering algorithm takes a lot less time to run. The MM clustering algorithm achieves the goal of automated cluster numbering prediction without the use of post cluster analysis, as well as efficient clustering and higher accuracy. While it takes longer than the MM clustering algorithm, the MM Extended K-Means clustering algorithm outperforms other hard clustering algorithms. MM and MM Extended K-Means clustering algorithms are acceptable by considering validation performance.

In end, the MM clustering algorithm takes the least amount of time compared to other standardized algorithms and the MMK algorithm finishes the job faster than other hard clustering algorithms. In all four datasets, MM clustering and MMK clustering approaches work well.

6 CONCLUSION

The philosophy behind the MM clustering algorithm and the MM Extended K-Means clustering algorithm is to naturally detect the exact clusters and classify them correspondingly. On both linear and non-linearly separated datasets, MM and MM Extended K-Means algorithms used. By validating the cluster and conducting performance analysis, the developed algorithms results compared to that of the other avant-garde algorithms. The distinction focused on execution time and validation indexes and this is a good way to measure which clustering algorithm is ideal for this type of dataset. The outcome of the MM Extended K-Means clustering algorithm on discriminant datasets is stronger than the MM clustering algorithm and on non-linearly separable datasets, MM algorithms triumph over others. Both the MM and MM Extended K-Means clustering algorithms achieve the particular requirements of correctly selecting the number of clusters efficiently and they provide faster and more reliable results than most of the other clustering algorithms.

We concentrate above on gene expression datasets and the methodologies put to the test in a monolithic architecture. We expect to work with large and real-time micro-array gene expression dataset in the future and incorporate them in distributed and parallel systems and perhaps even upgrade the algorithms. Consequently, innovation will be used to aid in the classification of micro-array gene expression information.

Acknowledgements

The researchers are thankful to the editors and reviewers for their insightful feedback and comments.

Declaration

Compliance with Ethical Standards:

Funding: N/A

Conflict of Interest: N/A.

Ethical approval: N/A

Reference

1. Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. *Comput Geosci* 10:191–203. doi: 10.1016/0098-3004(84)90020-7
2. Cai J, Wei H, Yang H, Zhao X (2020) A Novel Clustering Algorithm Based on DPC and PSO. *IEEE Access* 8:88200–88214. doi: 10.1109/ACCESS.2020.2992903
3. Chao Xu, Peilin Zhang, Bing Li, Dinghai Wu HF (2013) Vague C-means clustering algorithm - *ScienceDirect. Pattern Recognit Lett* 34:505–510
4. Chen JY, He HH (2015) Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Zidonghua Xuebao/Acta Autom Sin.* doi: 10.16383/j.aas.2015.c150062
5. Dhanachandra N, Manglem K, Chanu YJ (2015) Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm. In: *Procedia Computer Science*. Elsevier, pp 764–771
6. Dhanachandra N, Manglem K, Chanu YJ (2015) Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Comput Sci* 54:764–771. doi: 10.1016/j.procs.2015.06.090
7. Dua D, Graff C (2017) {UCI} Machine Learning Repository
8. Eberhart R, Kennedy J (1995) New optimizer using particle swarm theory. In: *Proceedings of the International Symposium on Micro Machine and Human Science*
9. Ester M, Kriegel H-P, Sander J, Xu X (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*
10. Gentle JE, Kaufman L, Rousseuw PJ (1991) *Finding Groups in Data: An Introduction to Cluster Analysis*.
11. Han J, Kamber M, Pei J (2012) *Data Mining: Concepts and Techniques*
12. Hou J, Gao H, Li X (2016) DSets-DBSCAN: A Parameter-Free Clustering Algorithm. *IEEE Trans Image Process.* doi: 10.1109/TIP.2016.2559803
13. Jahan M, Hasan M (2019) Performance Analysis and Benchmarking of Clustering Algorithms with gene datasets. In: *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*
14. Jahan M, Hasan M (2020) A novel fuzzy clustering approach for gene classification. *Int J Adv Comput Sci Appl.* doi: 10.14569/IJACSA.2020.0110809

15. Janmenjoy Nayak BN and HSB (2015) Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014 Janmenjoy. *Comput Intell Data Min* 2:133–149. doi: 10.1007/978-81-322-2208-8
16. Jinyin C, Xiang L, Haibing Z, Xintong B (2017) A novel cluster center fast determination clustering algorithm. *Appl Soft Comput J* 57:539–555. doi: 10.1016/j.asoc.2017.04.031
17. Lei T, Jia X, Zhang Y, He L, Meng H, Nandi AK (2018) Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering. *IEEE Trans Fuzzy Syst* 26:3027–3041. doi: 10.1109/TFUZZ.2018.2796074
18. Ng RT, Han J (1994) Efficient and Effective Clustering Methods for Spatial Data Mining. *Proc 20th Int Conf Very Large Data Bases* 144–155
19. Reddy GT, Khare N (2017) Hybrid Firefly-Bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis. *Int J Intell Eng Syst* 10:18–27. doi: 10.22266/ijies2017.0831.03
20. Reddy GT, Khare N (2017) An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model. *J Circuits, Syst Comput* 26:1–21. doi: 10.1142/S021812661750061X
21. Reddy GT, Reddy MPK, Lakshmana K, Rajput DS, Kaluri R, Srivastava G (2020) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evol Intell* 13:185–196. doi: 10.1007/s12065-019-00327-1
22. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. *Science* (80-). doi: 10.1126/science.1242072
23. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20:53–65. doi: 10.1016/0377-0427(87)90125-7
24. Sinaga KP, Yang MS (2020) Unsupervised K-means clustering algorithm. *IEEE Access* 8:80716–80727. doi: 10.1109/ACCESS.2020.2988796
25. Tilson L V., Excell PS, Green RJ (1988) A generalisation of the Fuzzy c-Means clustering algorithm. *Remote sensing Proc IGARSS '88 Symp Edinburgh, 1988 Vol 3* 10:1783–1784. doi: 10.1109/igarss.1988.569600