A Novel Fuzzy Clustering Approach for Gene Classification

Meskat Jahan[1], Mahmudul Hasan[2]
Computer Science and Engineering, Comilla University, Cumilla, Bangladesh[1, 2]
Information and Computer Sciences, Saitama University, Japan[2]

***Abstract*—Automatic cluster detection is crucial for real-time gene expression data where the quantity of missing values and noise ratio is relatively high. In this paper, algorithms of dynamical determination of the number of cluster and clustering have been proposed without any pre and post clustering assumptions. Proposed fuzzy Meskat-Hasan (MH) clustering provides solutions for sophisticated datasets. MH clustering extracts the hidden information of the unknown datasets. Based on the findings, it determines the number of clusters and performs seed based clustering dynamically. MH Extended K-Means cluster algorithm which is a nonparametric extension of the traditional K-Means algorithm and provides the solution for automatic cluster detection including runtime cluster selection. To ensure the accuracy and optimum partitioning, seven validation techniques were used for cluster evaluation. Four well known datasets were used for validation purposes. In the end, MH clustering and MH Extended K-Means clustering algorithms were found as a triumph over traditional algorithms.**

***Keywords*—Meskat-Hasan clustering (MH clustering); MH Extended K-Means clustering; K-Means; fuzzy clustering**

I. Introduction

Clustering divides the dataset based on data's attributes or characteristics. The fundamental purpose of clustering is to categorize the data based on their distinguishable attributes. For partitioning clustering, there are few established soft and hard versions of algorithms. Popular versions of hardcore clustering are K-Means, K-Medians, K-Modes, Forgy's algorithm and soft clustering are Fuzzy C-Means, Fuzzy K-Means (Fuzzy clustering). The variances of the Fuzzy C-Means, such as Gath-Geva (GG) clustering and Gustafson-Kessel (GK) clustering algorithms are used. Extended versions of the fuzzy clustering algorithm like E-FCM, Extended GK cluster exist. There are some other versions of the algorithm like Fuzzy K-NN algorithm and Fuzzy Local Information C-Means clustering algorithm. GrFPCM select features in the preprocessing step while FPCM and Granular Computing used for outlier detection and features selection [1]. To defeat high – dimensionality the problem, an ant-based algorithm used in the bioinformatics domain which enhanced with the use of FCM and heaps merging heuristic [2]. Gene ontology annotations based GO-FRC algorithm used biological data for gene clustering and this method may assign one gene into multiple clusters [3]. PSO clustering method established on fuzzy point symmetry used for gene expression classification [4]. FWCMR merge the sub clusters to form a final cluster which is implemented on the parallel and distributed environment [5]. WGFCM used entropy based weight vector calculation to appropriately measure the distance [6]. Immune system behavior based MCSOA used a new fast convergence mechanism for optimum solutions where the number of clusters varies in a certain range [7]. Dynamic Time Wrapping distance technique is useful in shaped based clustering while grouping time series GE data [8]. Fuzzy decision tree algorithm outperforms over classical decision tree algorithm in analyzing cancer GE data [9]. Also, there are techniques for determining the number of clusters like FLAME clustering. These algorithms help to find the behavior of the dataset to reveal the underlying hidden pattern by grouping similar categories of data based on characteristics and most of them need a good initial guess of the number of clusters to perform clustering. Predicting the accurate number of cluster is challenging task. Lots of algorithms are developed but depend on some predefined or prior knowledge. For example, K-NN, K-Means, FCM, etc. are all need robust initial guess.

It has been founded that previously this scenario was solved by applying some post cluster analysis to predict and selecting the number of clusters that require time and cost. Therefore, a method to determine the number of clusters dynamically is required. However, we develop two new algorithms, named Meskat-Hasan clustering (MH clustering) algorithm and MH-Extended K-Means clustering algorithm are proposed for automatic cluster detection, dynamic cluster selection and partitions. Moreover, post cluster enhancement is not required for them. So, they minimize time and cost complexity. Performance of these methods was validated using seven validation criteria Separation Index(S), Partition Coefficient (PC), Dunn's Index(DI), Alternative Dunn Index(ADI), Classification Entropy(CE), Partition Index(SC), Xie and Beni's Index(XB) and comparing results with existing clustering techniques like Fuzzy C-Means, Gath-Geva clustering (GG), Gustafson- Kessel algorithm (GK), K-Means and K- Median. Four different datasets (Wisconsin Breast Cancer, Leukemia, Irises and Motor Cycle [10]) were used to evaluate the performance of the algorithms.

Finally, the proposed algorithms are performed better than other existing literature.

II. Literature Reviews

Patrik D'haeseleer [11], described the working principles of gene expression clustering and suggested to use more than one clustering algorithms. According to Jain and Dubes [12], there is no single criterion to define a good clustering algorithm. Clusters are of arbitrary size and shapes in a multidimensional pattern space and clustering quality may be evaluated based on

internal criteria or external criteria. James C. Bezdek, Robert Ehrlich and William Full [13], proposed a program for Fuzzy C-Means (FCM) clustering algorithms to generate prototypes and fuzzy partitions for the numerical datasets. This fuzzy partition is useful for suggesting definite substructure for the raw datasets. Representing the similarity of a point is shared among the nearest clusters with the help of membership function whose values ranges from one to zero, is the idea given by Zadeh [14]. In FMLE algorithm good initial seed points are required because of the exponential distance helps to converge the algorithm to a local optimum rather in a narrow region. Except for this limitation, the FMLE algorithm is better than Gustafson and Kessel algorithm [15]. D. E. Gustafson and W. C. Kessel [15], developed a fuzzy clustering algorithm using fuzzy covariance matrix to prove the argument that in fuzzy clustering, fuzzy covariance has a natural approach. An expression of the interpretation of the membership functions was proposed by Ruspini [16]. This relationship denotes the similarity between samples where a fuzziness parameter is used, whose increasing value trend to indicate the more fuzziness of the clustering process. Jiye Liang, Xingwang Zhao, Deyu Li, Fuyuan Cao and Chuangyin Dang [17], proposed a clustering algorithm for special datasets like mixed datasets containing both numeric and categorical attributes. They presented a mechanism to characterize the data within the cluster and between cluster entropies and to detect the worst cluster in that particular dataset. Kaiser [18] proposed eigenvalues greater than one rule, which is now a commonly used criterion for finding the number of factors. It strongly states that the number of reliable factors is equal to the number of eigenvalues greater than one. As negative eigenvalues have negative reliability, the respective composite score should be reliable. In the internal consistency, it must have some positive reliability. Norman [19] concluded this by suggesting that more reliable components there will be that those are indicated in the eigenvalues greater than unity rules. `The convergence rate of evolutionary clustering methods is high enough than partition clustering methods [20]. This [21] comparative dissection paves the way of choosing the desirable clustering algorithm for some particular dataset.

III. Proposed Algorithms

Clustering is an unsupervised technique for categorizing the data elements. Previously, it was not possible to predict the accurate number of a cluster without conducting pre cluster analysis. Proposed techniques have been developed based on the principal component analysis. Then, these techniques were applied to the two types of clustering (Fuzzy and hard-core) algorithms. Finally, validation of the approaches was done and in the next sections, proposed two algorithms based on fuzzy clustering and hard clustering algorithms are shown.

A. Meskat-Hasan Clustering (MH Clustering) Algorithm

MH clustering is a fuzzy approach for data clustering. It is an integrated package of all the tasks to perform clustering including the determination of the number of cluster and clustering. Most of the established clustering algorithm has some kinds of dependency or need some data related prediction for knowing the behavior or characteristics of the dataset. To overcome this, Meskat-Hasan clustering (MH clustering) dynamically determine the clusters number. MH clustering has the following four steps:

Step-1: Normalize dataset.

Step-2: Extract underlying structure.

Step-3: Run time cluster number determination.

Step-4: Clustering using fuzzy c-means algorithm.

The generalized version of the algorithm is given below:

Meskat-Hasan (MH) clustering algorithm pseudocode

1. Let, $X = \{x_1, x_2, x_3 \ldots\ldots\ldots x_n\}$ be the dataset
2. Scale the data set,
$$X_{i=\{1\ldots.n\}} = \frac{Xi - Xmin}{X\max - Xmin}$$
3. Apply PCA and get variance.
4. Determine num_elements = count(variance)
5. Initialize i=1 and cumulative summation,
cum_sum=0.
6. Initialize k=0 and sum=0;
7. For i=1 to num_elements{
cum_sum = cum_sum + $\frac{variance\,(i)}{sum(variance)} * \mathbf{100}$;
}
8. Determine σ based on cum_sum.
9. Do {
Calculate cumulative summation of variance;
k=k+1;
}
Until (cum_sum≤ σ)
End Do
10. Assign the number of clusters, c = k;
11. Determine the fuzzy membership [3],
$\mu_{ij}^{(l)} = 1 / \sum_{p=1}^{c} (D_{ij} - D_{ip})^{(2/m-1)}$;
12. For j=1 to c, Calculate the fuzzy
centers [3],

$$V_{ij} = \left(\sum_{i=1}^{n} (\mathbf{1} * \mu_{ij})^m x_i\right) / \left(\sum_{i=1}^{n} (\mathbf{1} * \mu_{ij}^m\right);$$

13. Iterate step 11 and line 12 until $\| U^{(l)} - U^{(l-1)} \| < €$.

where € =1x10^{-6} and m=1 to ∞ is the fuzziness parameter.

In steps 1 to 10 determine the desired number of clusters and steps 11 to 13 clustering is performed.

B. MH Extended K-Means Clustering Algorithm

The MH Extended K-Means clustering algorithm is an extension of the K-Means algorithm. In the MH Extended K-Means algorithm, we will apply the techniques of determining the number of the clusters dynamically along with the original k-means algorithm. That means, to implement it in a hardcore clustering process. Main steps of this algorithm are as follows:

Step-1: Scale dataset

Step-2: Extract underlying structure

Step-3: Approaches to determine the cluster number dynamically

Step-4: Perform clustering by K-Means clustering procedure

The generalized version of the algorithm is given below:

MH Extended K-Means Clustering Algorithm Pseudocode

1. Let, X = {$x_1, x_2, x_3 \ldots\ldots\ldots x_n$} be the dataset.

2. Scaling of the dataset,

$$X_{i=\{1\ldots n\}} = \frac{Xi - Xmin}{Xmax - Xmin};$$

3. Apply PCA and get variance.

4. Count num_elements= count(variance)

5. Initialize i=1 and cum_sum=0

6. Set k=0 and sum=0;

7. For i=1 to num_elements{

$$cum\_sum = cum\_sum + \frac{variance\ (i)}{sum(variance)} * \mathbf{100};$$

}

8. Determine σ based on cum_sum.

9. Do {Calculate cumulative summation of variance;

k=k+1;

}

Until (cum_sum≤ σ)

End Do

10. Assign clusters, c = k;

11. Initially, select c centres randomly.

12. Compute minimum distances between data elements and cluster centres.

13. Update c using mean values of elements

14. Repeat steps 12 to 13 until no data elements won't be reassigned.

MH Extended K-Means is designed to determine the number of clusters for the well-separated dataset. It is a combined method of dynamically determining the number of the cluster with the traditional K-means algorithm. K-means works for well-separated dataset but, it needs a strong initial assumption to the number of clusters. To defeat limitation, MH Extended K-Means is designed. It accurately determines the number of the cluster for the well-separated dataset.

C. Predicting the Number of Cluster-based on σ Value

Non-parametric value σ is the limiting criterion indicating the percentage of variation of the dataset. σ value is determined from the principal component analysis technique. The cumulative summation of main components variances of the dataset is the limiting value of σ. If the value of σ is high, then the number of the cluster becomes less and vice versa. For example, in leukaemia dataset, the summation of three main principal components is 66.3% indicating 66.3% variation of the total dataset. So for leukaemia dataset, we set σ = 66.3 and finally it concludes to set 2 as the number of clusters. For Wisconsin Breast Cancer (WBC) dataset the outcome of MH clustering algorithm is three number of cluster and σ value is 86.7 and so three cluster hold at least 86.7% data. For Leukemia dataset number of cluster is two and σ value is 66.3, so two clusters have 66.3% data. For Irises dataset the number of clusters is two and σ value is 92.46, therefore, two clusters keep 92%. For Motor Cycle dataset the number of clusters is three and σ value is 85.3448 thus three clusters contain 85%. In MH Extended K-Means clustering algorithm number of cluster for WBC, Leukemia, Irises and Motor Cycle dataset is three, two, two and three respectively. MH Extended K-Means clustering algorithm brings 86%, 85%, 92% and 85% underlying data into consideration for dynamically determination of cluster number.

IV. A Comparative Study among Clustering Algorithms

Performance of algorithms is compared based on average execution time vs. no. of clusters. Considering a certain number of clusters, execution times of the corresponding algorithm is obtained and comparative study is as below:

A. Time Comparison among Clustering Algorithms

Fig. 1, the proposed MH clustering algorithm takes the lowest time for clustering and MH Extended K-means clustering takes highest times. WBC dataset is not well separated and for this nature of datasets proposed fuzzy algorithm takes less time comparatively. Fig. 2, MH clustering algorithm takes the lowest time to perform clustering compared to other fuzzy clustering algorithms. Hard-core clustering like K-Means, K- Mediods takes comparatively longer time than MH Extended K-Means clustering algorithm. Leukemia dataset is not well separated and for these nature of dataset proposed fuzzy algorithm takes comparatively much less time. Fig. 3. Meskat-Hasan (MH) clustering algorithm takes the lowest time. Others fuzzy clustering algorithm like FCM, GG and GK takes almost the same times to perform executions. Besides, hard-core clustering likes K-Means, K-Mediods takes comparatively more times than MH Extended K-means clustering. In Fig. 4. Meskat-Hasan (MH) clustering algorithm takes the lowest time and MH Extended K-Means clustering takes highest times to execute.

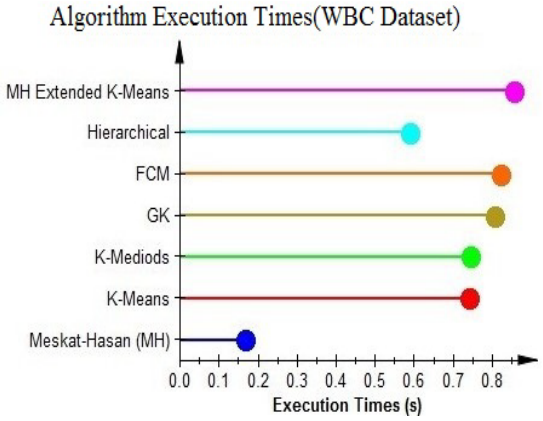

Fig. 1. Algorithms Performance Comparison on WBC Dataset.

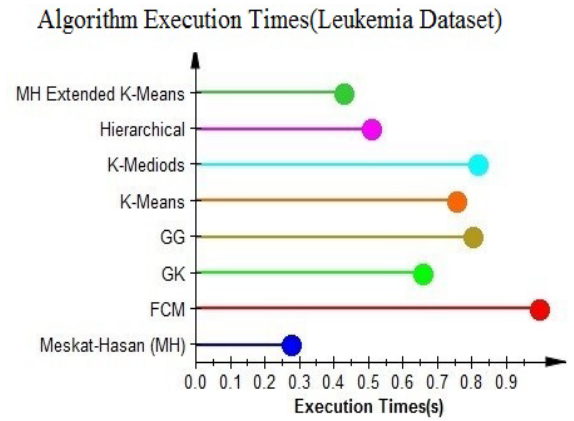

Fig. 2. Algorithms Performance Comparison on Leukemia Dataset.

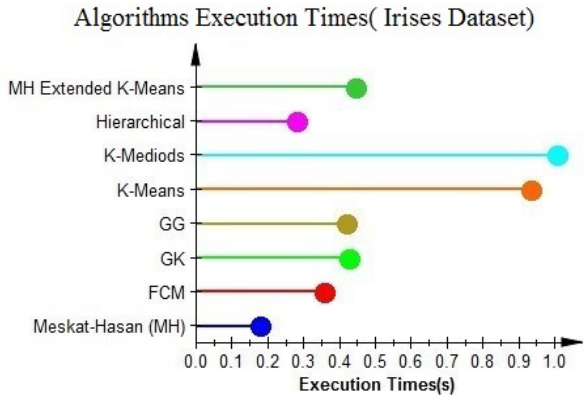

Fig. 3. Algorithms Performance Comparison on Irises Dataset.

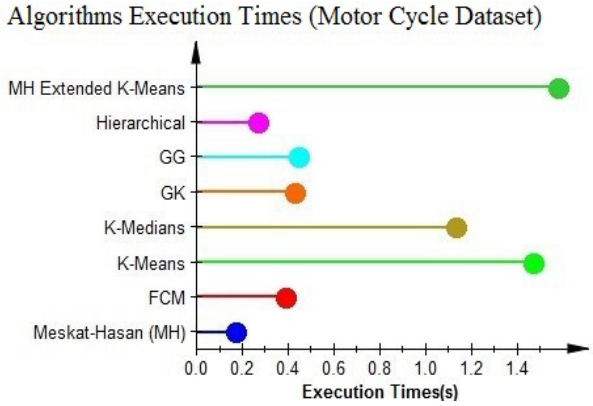

Fig. 4. Algorithms Performance Comparison on Motor Cycle Dataset.

B. MH Over MH Extended K-Means Clustering Algorithm

MH uses the fuzzy approach in cluster implementation whereas MH Extended K-Means use hard clustering approach. MH Extended K-Means perform better when the input dataset is well separated and MH performs better for a non-linearly separable dataset.

V. Cluster Validation

To evaluate cluster performance analysis validation played an important role. For validating partition Separation index (S) takes the minimum separation distance where the smallest values of S indicate a valid optimal partition. To measure the amount of overlapping between cluster partition coefficient (PC) indexing is used and its high value provide cluster accuracy. Dunn Index (DI) is internal evaluation criteria which identify the well separate cluster. Higher DI and ADI values indicate better clustering results. CE measures the cluster partitions fuzziness. Low values of CE and SC reflect good performance. XB index recognizes whole cluster compactness and smallest value site the optimum number of cluster. Validation result of the MH clustering is organized in Table I.

Table II reveals the performance of the MH Extended K-Means algorithm over the four datasets. As PC indicates the amount of overlapping between cluster regions, so it provides constant values in MH Extended K-Means which is a hard clustering algorithm.

TABLE I. MH Clustering Algorithm Validation

Dataset	Cluster Number	PC	CE	SC	S	XB	DI	ADI
WBC	3	0.8	0.3	0.3	0.0	1.9	0.1	0.0
Leukemia	2	0.9	0.1	0.5	0.01	4.1	0.4	0.10
Irises	2	0.9	0.2	0.8	0.01	6.3	0.1	0.01
Motor Cycle	3	0.7	0.5	1.9	0.02	5.4	0.01	0.01

TABLE II. MH Extended K-Means Clustering Algorithm Validation

Dataset	Cluster Number	PC	CE	SC	S	XB	DI	ADI
WBC	3	1	-	0.31	0.001	3.2	0.17	0.004
Leukemia	2	1	-	0.56	0.007	4.7	0.38	0.03
Irises	2	1	-	0.62	0.004	41	0.1	0.002
Motor Cycle	3	1	-	1.37	0.012	16	0.01	2e-02

A. Analyzing Total Outcomes

All outcomes of both proposed and existing algorithms are compared based on the evaluation criteria for each dataset. Table III is the outcomes of a comparative view of the performance of the algorithm of WBC dataset. The values of PC & CE are respectively constant for hard clustering. The lowest values of SC, S & XB are provided by the MH clustering algorithm. Highest values of the DI and the lowest value of ADI are provided by the MH Extended K-Means clustering algorithms. Hence, MH clustering algorithm gives better results. Table IV shows the algorithms performance of Leukemia dataset. Based on values of PC, CE, SC, S, XB, DI and ADI it can be concluded that for Leukemia dataset GK and MH clustering algorithm gives better results. Table V organized the validation result of Irises dataset. The values of PC & CE are respectively constant and not working for hard clustering. The lowest value of SC & S is provided by MH Extended K-Means clustering algorithm. The low value of XB provides by GG. But, the low value of the DI provided by the GG and GK and the lowest value of ADI are provided by the K-Mediods algorithm. Table VI performs comparative studies of Motor Cycle dataset. MH clustering algorithm takes lowest times. The values of PC & CE are not working for hard

clustering. The lowest value of SC, S & ADI is provided by the MH Extended K-Means clustering algorithm. By considering above stated measurement MH Extended K-Means clustering algorithm gives a better result for the motorcycle dataset.

TABLE III. ALGORITHMS PERFORMANCE COMPARISON ON WBC DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.3	0.8	6E-4	0.3	0.1	1.9	0
FCM	0.7	0.8	1E-3	0.3	0.1	2.6	0
GK	0.6	0.9	1E-3	0.2	0.1	2.2	0
K-Means	0.5	1.0	9E-4	NA	0.1	2.7	0
K-Mediods	0.5	1.0	1E-3	NA	0.1	Inf	0
MH Extended K-Means	0.3	1.0	7E-4	NA	0.2	3.2	0

TABLE IV. ALGORITHMS PERFORMANCE COMPARISON ON LEUKEMIA DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.5	0.9	7E-3	0.1	0.4	4.1	0.11
FCM	0.6	0.9	9E-3	0.2	0.4	5.2	0.03
GK	0.5	0.9	7E-3	0.1	0.4	4.1	0.11
GG	1.7	1.0	2E-2	0.0	0.4	2.7	0.05
K-Means	0.5	1.0	7E-3	NA	0.4	4.7	0.03
K-Mediods	0.6	1.0	1E-2	NA	0.4	Inf	0.02
MH Extended K-Means	0.6	1.0	8E-2	NA	0.4	4.7	0.03

TABLE V. ALGORITHMS PERFORMANCE COMPARISON ON IRISES DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.8	0.9	0.0	0.2	0.1	6.3	0.01
FCM	0.8	0.9	0.0	0.2	0.1	6.3	0.01
GK	0.6	0.9	0.0	0.2	0.2	16.6	0.01
GG	1.0	1.0	0.0	4E-4	0.2	1.9	0.0
K-Means	0.6	1	0.0	NA	0.1	41.2	0.0
K- Mediods	0.6	1	0.0	NA	0.1	Inf	0.0
MH Extended K-Means	0.6	1	0.0	NA	0.1	41.2	0.0

TABLE VI. ALGORITHMS PERFORMANCE COMPARISON ON MOTOR CYCLE DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	1.9	0.7	0.02	0.5	0.01	5.3	0.01
FCM	1.9	0.6	0.02	0.5	0.0	5.4	0.01
GK	2.0	0.6	0.02	0.6	0.0	3.7	0.02
GG	4.9	0.9	0.05	0.1	0.03	1.9	0.06
K-Means	1.4	1	0.01	NA	0.01	14	0.02
K-Mediods	1.5	1	0.01	NA	0.02	Inf	0.0
MH Extended K-Means	1.4	1	0.01	NA	0.01	16	0.0

VI. RESULT AND DISCUSSION

MH clustering algorithm provides solutions for automatic clusters number detection, run time cluster selection, and performs fuzzy clustering accordingly. It appropriately determines clusters number and produces proper partitioning. MH clustering algorithm works well for the non-linear dataset. MH Extended K-Means algorithm performs hard clustering on the basis of dynamic cluster number determination. It works well for a clearly separable dataset. By analyzing all results, both MH and MH Extended K-Means clustering algorithms select the precise cluster number and produce optimum partitioning and perform clustering accordingly. MH clustering algorithm takes comparatively less time to execute than other algorithms. Based on validation techniques, MH clustering algorithm performance is quite better than the other established literature. MH clustering algorithm meets the objective of automatic cluster number detection without using post cluster analysis and performs clustering accurately and satisfy the time complexity. Though MH Extended K-Means clustering algorithm takes more time than the MH clustering algorithm but performs better than other hard clustering algorithms. Evaluating validation results MH and MH Extended K-Means clustering algorithm are acceptable.

VII. CONCLUSION

The idea of MH clustering algorithm and MH Extended K-Means clustering algorithm comes to solve the problem of the exact number of cluster detection automatically and perform clustering accordingly. MH and MH Extended K-Means were applied on both linear and non–linearly partitioned dataset. Performance of the proposed algorithms was compared with other selected algorithms by validating the cluster and performance evaluation. The comparison was done based on execution time and validation indexes and it provides an effective way of selecting an efficient clustering algorithm for the particular dataset. For linearly separable dataset performance of MH Extended K-Means clustering algorithm and non-linearly separable dataset, MH clustering algorithm is better. Both MH and MH Extended K-Means clustering algorithm meet the desired needs of dynamically determining the number of clusters accurately and provides better and efficient results than the selected clustering algorithms.

Here we work on gene expression datasets and algorithms are tested in standalone systems. In the future, we want to work with real-time microarray gene expression dataset and implement in parallel and distributed system and will upgrades the algorithms accordingly. So that classification of microarray gene expression data can take computational benefit from cloud infrastructure.

REFERENCES

[1] Truong, H.Q., Ngo, L.T. and Pedrycz, W., 2017. Granular fuzzy possibilistic C-means clustering approach to DNA microarray problem. Knowledge-Based Systems, 133, pp.53-65.

[2] Bulut, H., Onan, A. and Korukoğlu, S., 2020. An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data. Sādhanā, 45(1), pp.1-17.

[3] Paul, A.K. and Shill, P.C., 2018. Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. Biosystems, 163, pp.1-10.

[4] Das, R. and Saha, S., 2015, November. Gene expression classification using a fuzzy point symmetry based PSO clustering technique. In 2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI) (pp. 69-73). IEEE.

[5] Hosseini, B. and Kiani, K., 2018. FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression. Expert Systems with Applications, 91, pp.198-210.

[6] Jiang, Z., Li, T., Min, W., Qi, Z. and Rao, Y., 2017. Fuzzy c-means clustering based on weights and gene expression programming. Pattern Recognition Letters, 90, pp.1-7.

[7] Zareizadeh, Z., Helfroush, M.S., Rahideh, A. and Kazemi, K., 2018. A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework. Expert Systems with Applications, 113, pp.301-314.

[8] Izakian, H., Pedrycz, W. and Jamal, I., 2015. Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, pp.235-244.

[9] Ludwig, S.A., Picek, S. and Jakobovic, D., 2018. Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm. In Operations Research Applications in Health Care Management (pp. 327-347). Springer, Cham.

[10] Asuncion, A., & Newman, D.: UCI machine learning repository. (2007).

[11] D'haeseleer, P.: How does gene expression clustering work? Nature biotechnology, 23(12), 1499-1501. (2005).

[12] Jain, A. K., & Dubes, R. C.: Algorithms for clustering data. Prentice-Hall, Inc... (1988).

[13] Bezdek, J. C., Ehrlich, R., & Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203. (1984).

[14] Lee, E. T. L., A. ZADEH: Note on Fuzzy Languages. Inform. Sciences, 1(4). (1969).

[15] Gustafson, D. E., & Kessel, W. C.: Fuzzy clustering with a fuzzy covariance matrix. In 1978 IEEE conference on decision and control including the 17th symposium on adaptive processes (pp. 761-766). IEEE. (1979, January).

[16] Ruspini, E. H.: A new approach to clustering. Information and control, 15(1), 22-32. (1969).

[17] Liang, J., Zhao, X., Li, D., Cao, F., & Dang, C.: Determining the number of clusters using information entropy for mixed data. Pattern Recognition, 45(6), 2251-2265. (2012).

[18] Kaiser, H. F.: The application of electronic computers to factor analysis. Educational and psychological measurement, 20(1), 141-151. (1960).

[19] Cliff, N.: The eigenvalues-greater-than-one rule and the reliability of components. Psychological bulletin, 103(2), 276. (1988).

[20] Patibandla, R. L., & Veeranjaneyulu, N.: Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria. Arabian Journal for Science and Engineering, 43(8), 4379-4390. (2018).

[21] Jahan, M., & Hasan, M.: Performance Analysis and Benchmarking of Clustering Algorithms with gene datasets. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) IEEE. (pp. 1-5). (2019, May).