A new approach for improving clustering algorithms performance

Anfal F. N. Alrammahi, Kadhim B. S. Aljanabi
Faculty of Computer Science and Mathematics, University of Kufa, Iraq

Article Info

Article history:

Received Jan 14, 2020
Revised Feb 3, 2020
Accepted Feb 20, 2020

Keywords:

Centroids
Clustering
Euclidian Distance

ABSTRACT

Clustering represents one of the most popular and used Data Mining techniques due to its usefulness and the wide variations of the applications in real world. Defining the number of the clusters required is an application oriented context, this means that the number of clusters k is an input to the whole clustering process. The proposed approach represents a solution for estimating the optimum number of clusters. It is based on the use of iterative K-means clustering under three different criteria; centroids convergence, total distance between the objects and the cluster centroid and the number of migrated objects which can be used effectively to ensure better clustering accuracy and performance. A total of 20000 records available on the internet were used in the proposed approach to test the approach. The results obtained from the approach showed good improvement on clustering accuracy and algorithm performance over the other techniques where centroids convergence represents a major clustering criteria. C# and Microsoft Excel were the software used in the approach.

Corresponding Author:

Anfal Falah Najeeb Alrammahi,
Department of Computer Science,
University of Kufa, Iraq.
Email: anfal9abd@gmail.com

1. INTRODUCTION

Clustering among the different Data Mining techniques is the most widely used DM technique. Clustering in Data Mining (DM) is the method of grouping data objects into clusters, where data objects are similar to each other within a cluster and dissimilar to other data objects in other clusters. Similarities/Dissimilarities between data objects are estimated based on the attribute values by using distance measures. The main goal of data clustering to find the groups of a set of points, patterns or objects.other purposes to data clustering can be used to gain insigt into data, discover anomalies, identify features of the data, find the degree of similarity and organize and summarize the data. In some applications, Clustering can be called as data segmentation as a result of partitioning large data sets into groups based to their similarity. Clustering can be utilized for outlier detection and the detection of credit card fraud. Clustering is a task in exploratory DM and a technique used in many fields including biology, statistics, pattern recognition,information retrieval , bioinformatics and machine learning [1-3].

Clustering broadly divided into two types hierarchical and partitioning based on the cluster structure which they produce. Hierarchical clustering which use a hierarchy of clusters or nodes to cluster data objects. Advantage of this method is didn't require any predefined information about the number of clusters. But it could not to get back to a previously finished work. The hierarchical methods collect training data into a tree structure of clusters, which this tree called dendrogram. It represents a sequence of nested cluster which constructed top-down or bottom-up. The root of the dendrogram tree represented by one cluster, including all data points, while the remaining n clusters represents the leaves of the tree, each cluster include one data point. In order to Cluster the data points into disjoint groups by cutting the tree at a desired level. Heirarchical clusters data in agglomerative or divisive mode. First mode, the clustering process start with representing each data object as a single cluster. After that, the clustering proceed by merging similar clusters recursively. The second mode, the clustering process consider all data objects are a single cluster and then dividing this cluster into different clusters recursively [4, 5].

The second type is partitioning clustering that work on partitioning objects into different clusters. In this method, each cluster contain the data objects of a similar characteristics while different clusters has a dissimilar data objects. A partition clustering algorithm objective is to split the data points into K partitions. Each partition will represent one cluster. Partition technique depends upon specific objective functions. The weakness of a partition algorithm is whenever the distance between the two data points from the center are close to another cluster because the result becomes misleading due convergence of the data points. K-means algorithm most known clustering algorithm used in this paper. This simple iterative method that partition a given data set into a number of clusters.This algorithm suffers from the diffcullty to find the optimal number of k. Alongside the quality of clustering results depends on the selection of initial centroids [6-9].

In some clustering applications, the number of clusters is given as input (e.g. grouping a set of students according to their overall average).Whereas some other applications, the goal is to find out the optimum number of clusters for such applications (what is the most suitable number of clusters required to group families according to the standard of living). The goal of this research paper to propose an approach to find out the optimum number of clusters suitable for grouping some data sets. For this issue, many attempts have been proposed methods to find out the optimal number of clusters. Cohen-addad, *et al.* [10] optimize a specific objective for hierarchical clustering and analysis the performance of both similarity and dissimilarity based on hierarchical clustering. Nguyen, *et al.* [11] uses a bi-level hierarchical clustering model that it is formulations suffers from a discrete optimization problem. Dey, *et al.* [12] proposes a three optimization parameters which represents a solution to temporal clustering problems. This lead to develop new algorithms that performs compromise between these three parameters. Levin [13] describes a various combinatorial balancing problems in clustering and a new balance indices suggested for clustering solutions.

Śmieja *et al.* [14] propose a model to clustering the sparse high dimensional binary data called SparseMix. This model constructs a highly compatible partition. Howe [15] improve a new k-means method called Augmented k-means. This method clustering data sets accurately and in fewer iterations than the original k-means. Azab and Hefny [16] propose a local model of PSO (particle swarm optimization) for partitioning clustering to get rid off the disadvantages of this model. That led to optimize the location of the centroid of the cluster. Zhou *et al.* [17] propose a method to determine the optimal number of clusters based on an agglomerative hierarchical clustering algorithm and a new clustering validity index to evaluate the results produced by the method. Khanchouch, *et al.* [18] focus on multi-SOM clustering algorithm and test this algorithm using real data sets with two evalution criteria to extract the optimal number of clusters. This method takes less iterations steps but considered insufficient to define the boundaries of each cluster. Kiani, *et al.* [19] propose a method by applying a model depended on data mining techniques and optimize the clustering technique by assigning weights to features and they also use GA in order to improve outlier detection. Sekula [20] propose the R package optCluster as method to determine the best number of clusters with the most suitable algorithm for a given set of data. This method evaluate data with ten clustering algorithms and then the selected algorithms are evaluated using nine validation measures which classified as “biological”, “internal”, or “stability”. Muca, *et al.* [21] proposed a method for determing the optimal number of clusters using cluster validation measures.results shows that the method sensitive to the initial selection of centroids and takes more computational time if used with large data. Subbalakshmi, *et al.* [22] proposed a method to solve the optimal number of clusters based on fuzzy silhouette on dynamic data.

This method suffer from a large time complexity. Liang, *et al.* [23] propose an intializtion method that find both the cluster centroids and number of clusters for categorical data. Chiang and Mirkin [24] propose an adjusted method of k-means to find the right number of clusers based on the comparison with seven different approachs under different cluster spread-shape parameters.they get best results of one method that reproduce the right number of clusters but not good results within clusters or centroids recovery. Unlike traditional clustering applications in which number of clusters (k) represent one of the inputs to the algorithms, some other applications tend to deal with k as a parameter to be estimated. This paper proposes a novel improving method of clustering to tackle the issue of determining the optimum number of clusters in such applications.

2. THE PROPOSED APPROACH

The data set used to test the algorithm from UCI repository that consists of two attributes and 20000 reocords as shown Table 1 a sample of 25 reocrds:

Table 1. A sample of data set

	X1	Y1		X1	Y1
1	664159	550946	26	601182	582584
2	665845	557965	27	562704	570596
3	597173	575538	28	605107	563429
4	618600	551446	29	607214	575069
5	635690	608046	30	568824	570203
6	588100	557588	31	612485	518009
7	582015	546191	32	589244	573777
8	604678	574577	33	625579	551084
9	572029	518313	34	560237	500154
10	604737	574591	35	626224	569687
11	577728	587566	36	610666	551701
12	602013	574722	37	597428	569940
13	627968	574625	38	600582	599535
14	607269	536961	39	604168	555003
15	603145	574795	40	613871	550423
16	671919	571761	41	617310	551945
17	612184	570393	42	625728	579460
18	600032	575310	43	606300	566708
19	627912	593892	44	638559	558807
20	601967	604428	45	582176	630383
21	591851	569051	46	544056	577786
22	601444	572693	47	631297	578351
23	629718	558104	48	561574	621747
24	661430	603567	49	604973	574773
25	597551	556737	50	664159	550946

The proposed algorithm used the output of the initializing centroids in k-means clustering algorithm. The algorithm presents an approach for estimating the starting intial centroids through three processes including density based, normalization and smoothing ideas [25]. The proposed algorithm clusters a data set X into k clusters in k-means using Euclidean distance between data objects and centroids. The algorithm calls initializing centroids algorithm once to initialize the centroids for a given data set with k value starting with k=2. After that, iterative K-means algorithm is use and the convergence criteria is applied to find the optimal k. A convergence criterion consists of the following:

- Using Euclidean distance given in (1) between the objects of a given cluster Ci and its centroid.
- Calculating the difference between two iterative centroids of a given cluster.
- Calculating the number of migrated objects (objects migrant from one cluster to another).

After calling the first algorithm then start for loop to accomplish the k-means clustering of the data set with current estimated centroids of k value as in Euclidean distance (1) below:

$$d(X,C) = \sqrt{(X_{i1} - C_{i1})^2 + (X_{i2} - C_{i2})^2 + \cdots + (X_{iL} - C_{iL})^2} \qquad (1)$$

where i is the row number that both data object and centroid value belong to L is the column number start from (1, 2…, L).

Calculate the distance between the data objects and centroids for same row with L columns. Calculate distance of each data object to all centroids then include that object to the cluster of the minimum distance calculated between data and it is centroid. In the same iteration with current k value and centroids values, estimate the mean values of the objects for each cluster Ci=1..., k resulted from clustering of the previous step and the mean values represents a new centroid for the next iteration with current k value as in (2):

$$mean = \frac{X1C_i + X2C_i + \cdots + XtC_i}{t} \qquad (2)$$

where t is total number of objects in cluster Ci

In every iteration, check if the centroids of the n iteration and n-1 iteration are totally similar. Perform the centroid convergence between the two centroids of current and previous iterations with current k value, by using percentage difference of the two centroids as follow as in (3):

$$difference = \frac{abs\ (C_{(n-1)} - C_{(n)})}{C_{(n-1)}} \tag{3}$$

where n-1 is the previous iteration and n is the current iteration

If the condition in the previous step is performed, then terminate the loop and perform the other two criteria to current k value and increment k value bisectionally as: k=k*2 and start over again for loop from 0 to N iterations and clustering data set X with initial centroids generated by calling algorithm one for new k value. Otherwise, if the condition in the previous step isn’t performed, then move to the next iteration with current k value and cluster data set X with mean values as new centroids and compute a new means until perform the centroids convergence criteria. In case the centroids criteria performed and loop terminated then perform the two other criteria, total distance for each k value by calculating distance for each cluster then sum the distance values as total distance of k value as in (4):

$$d(i,j) = \sqrt{(Xc_{i1} - Cc_{i1})^2 + (Xc_{i2} - Cc_{i2})^2 + \cdots + (Xc_{iL} - Cc_{iL})^2} \tag{4}$$

where X is data object in cluster number ci (ci=1…k) and C is cluster centroid for the same ci number for each column.

Repeat this process for each cluster separately and calculate the sum of the distances of all clusters with current k value to represent the total distances of specific k value. The third criterion is the number of migrated objects in each k and it is percentage to total number of objects. In the last two iterations (n and n-1) when the loop stopped (after the centroids convergence performed), a counter value start with 0 then incremented if each data object of a cluster Ci in iteration n-1 moved to another cluster number in iteration n (clusters of current k value) and that counter value represent number of migrated objects in k value as in (5):

$$percentage = \frac{m}{D} \tag{5}$$

where m is the calculated number of migrated objects in k value and D is the total number of objects in data set X.The following is the proposed algorithm.

```
Input: Given a data set X with L columns and k clusters
Output: initializing the optimal k cluster for a given data set
Step 1: start k=2
Step 2: Call algorithm one (X, k) as C
Step 3: for n =0 to N (where N is the number of iterations) perform the steps from 4 to 6:
Step 4: cluster X with C by calculating Euclidean distance based on equation (1)
Step 5: Estimate the mean values based on equation (2)
Step 6: Apply the centroid convergence between the centroids of current and previous iterations with current k value, by using percentage difference equation (3)
Step 7: If the condition in step 6 is fulfilled, then go to step 8 otherwise, increment k=k*2 and repeat from step 2.
Step 8: Perform total distance criteria for each k value by calculating distance equation (4)
Step 9: Perform the third criteria is the number of migrated objects in each k using equation (5)
Step 10: Starting from maximum k value received from step 7 and applying the number of migrated objects, where the convergence criteria is achieved if the number of migrated objects < 10% of the total objects in the dataset. Else, k=k-2 and repeat from step 2 to 9 until k value reach ki /2 (ki represent the last incremented k value been reached).
If a minimm total distance for each decremented k values criteria is achieved then the solution is found and that k (any of decremented k) represent the optimal k for X. else increment k and repeat from step 2.
Step 11: End
```

3. RESULTS AND ANALYSIS

In the proposed approach, three criteria were used to find the optimal k value for a given data set. Depending of the dataset, its application and the clustering domain the centroid convergence criteria is not sufficient to get the optimum number of clusters, and hence two other criteria were applied.

3.1. Centroids convergence criteria:

The centroid criteria is a critical condition in all clustering techniques. In the proposed approach, the two other conditions were added to it after it is performed. This condition tests the similarity of the centroids values between two iterations. Inside iterations with each k value, test if the centroids of the current and previous iteration are similar. Each time check if the percentage difference of the centroid values of each column less than 2%. The selected percentage value 2% represents 400 training from the total number of the data set (20000) to create a model and reach a solution.

3.2. Total distance

This condition is applied for each k value by calculating the distance of each cluster using Euclidean distance and summing the distances to represents the total distance for k. this condition shows the distance of each data object to the centroid in each cluster then summing the calculated distances of all objects belong to the same cluster. At the end sum the distances of all clusters of k value. Results shows in most cases that the total distance of each k gets minimize while k value incremented each time as shown in Figure 2.

3.3. Number of migrated objects

Clustering process is ended when no objects (or percentage of the total objects) migrated between clusters. The third condition performed on the clusters of the last two iterations (n and n-1) to count the number of objects moved between clusters in a k value. The percentage of the migrated objects to the total number of data objects. That percentage must accomplish value less than 10%. 10 % represent 2000 training from the data set size 20000 then it is need to 18000 to build a model. Sometimes the value of the number of migrated objects increased each time the k value incremented when the centroids values be very convergent to each other as shown in Figure 1.

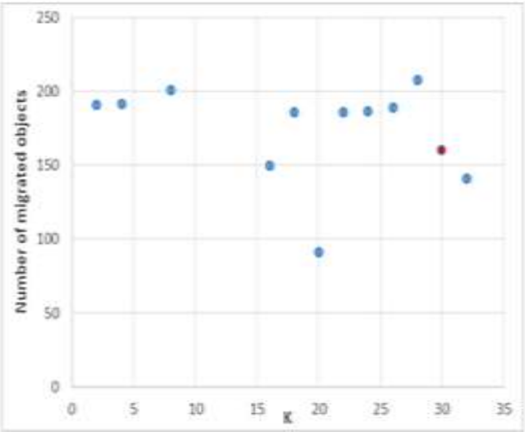

Figure 1. Variations of number of migrated objects with number of clusters K

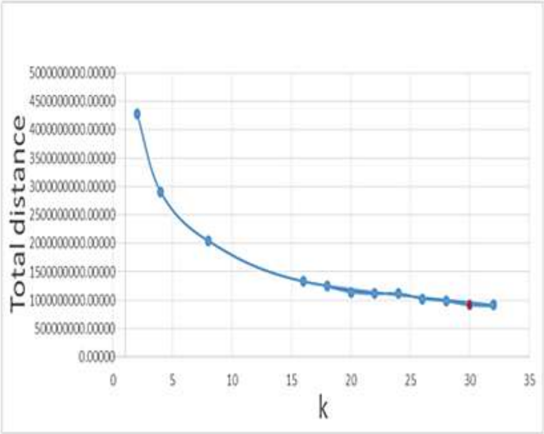

Figure 2. Variations of total distance with number of clusters K

The following tables represents the results of using the proposed approach using the three different criteria. We modify this proposed approach by making the stop condition is the percentage to the number of migrated objects instead of centroids criteria in case it performed a value less than 0.01 then perform the other two criteria and increment k value to start over again. The results in this case shows that it need more iterations to accomplish both percentages of migrated and centroids to increment k value. If the migrated criteria is performed and when moved to the centroids criteria did not performed a value less than 0.02, so the loop continue (of the current k value) to accomplish a stop condition and centroids criteria for the same iteration. So at the end, the result of both cases (the proposed approach and modified one) shows same optimal k value for the used data set but notice an increase of number of iterations to reach a solution in the modified approach and a decrease in the number migrated objects as showed in the following Table 3. The results shows that the proposed approach has high efficiency and better algorithm performance. Depending on the clustering application behavior and characteristics, all or parts of these criteria can be used to get better clustering performance.

Table 2, Table 3, Figure 1, Figure 2 and Figure 3 show that the optimal number of clusters for the data set been used in this paper can be achieved at k=30 with total distance=912780284.928115. The overall complexity of the proposed approach is O(I K N+ K N+ K L) Where K is the number of clusters, I is the number of iterations, N is the number of objects in a dataset and L is the number of attributes in data set.

Table 2. Results of the proposed algorithm

Iteration	K	Number of migrated objects	Percentage of migrated objects	Total distance
5	2	191	0.00955	4286606683.03855
9	4	192	0.0096	2902839691.63797
13	8	201	0.01005	2045256657.51508
10	16	150	0.0075	1331491345.24666
16	32	141	0.00705	914112864.505519
14	30	160	0.008	912780284.928115
10	28	208	0.0104	986674137.522737
18	26	189	0.00945	1026804521.38925
8	24	187	0.00935	1119842242.43811
16	22	186	0.0093	1111071182.45076
20	20	91	0.00455	1143500945.08777
16	18	186	0.0093	1244551702.78833

Table 3. Results of the modified to the proposed algorithm

Iteration	K	Number of migrated objects	Percentage of migrated objects	Total distance
5	2	191	0.00955	4286606683.03855
9	4	192	0.0096	2902839691.63797
14	8	157	0.00785	2044153529.13981
10	16	150	0.0075	1331491345.24666
16	32	141	0.00705	914112864.505519
14	30	160	0.008	912780284.928115
13	28	130	0.0065	984910742.436268
18	26	189	0.00945	1026804521.38925
8	24	187	0.00935	1119842242.43811
16	22	186	0.0093	1111071182.45076
20	20	91	0.00455	1143500945.08777
16	18	186	0.0093	1244551702.78833

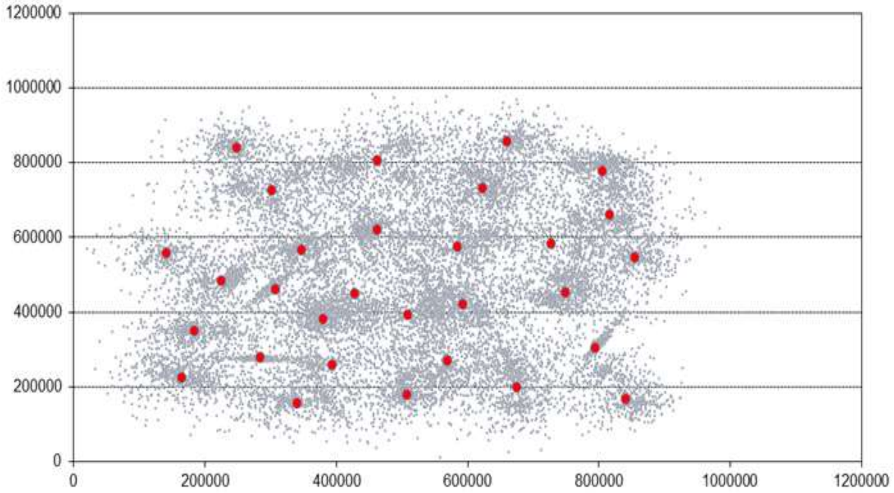

Figure 3. Data set scattered with the final centroids k=30 that represent the optimal solution found by the proposed approach

4. CONCLUSION

A proposed approach of a set of algorithms has been suggested in order to determine the optimal value of the number of clusters taking into consideration optimization of the time complexity of the proposed algorithms and the convergence criteria of clustering. Total distance, centroids convergence and number of migrated objects are the core criteria of the proposed approach. Iterative k-means clustering is an effective technique for grouping the dataset and an efficient algorithm for initializing centroids was used to improve the overall performance. The results given in Tables 2 and 3 and Figures 1 and 2, show that the proposed approach in which three different convergence criteria were used gives better and reliable convergence than using each criteria individually.

REFERENCES

[1] K. Jiawei Han, “Data Mining: Concepts and Techniques,” *Elsevier*, vol. 12, 2011.

[2] B. Gondaliya, “Review paper on clustering techniques,” *Internatonal Journal of Engineering Technology*, vol. 2, no. 7, pp. 234–237, 2014.

[3] Anil K. Jain, "Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters, Elsevier*, vol. 31, no. 8, pp. 651-666, 2010.

[4] K. Kameshwaran & K. Malarvizhi, “Survey on Clustering Techniques in Data Mining,” *International Journal of Computer Sceince and Information Technologies*, vol. 5, no. 2, 2014.

[5] M. K. Rafsanjani, Z. A. Varzaneh, & N. E. Chukanl, “A survey of hierarchical clustering algorithms,” *Journal of Mathematics and Computer Science*, vol. 5, no. 3, pp. 229–240, 2012.

[6] G. Gandhi, “Review Paper : A Comparative Study on Partitioning Techniques of Clustering Algorithms,” *International Journal of Computer Applications*, vol. 87, no. 9, pp. 10–13, 2014.

[7] S. S. J & S. Pandya, "An Overview of Partitioning Algorithms in Clustering Techniques,” *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 5, no. 6, pp. 1943–1946, 2016.

[8] Wu, Xindong, et al., "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1-37,2008.

[9] Dhanachandra, et al., "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, vol. 54, pp. 764-771, 2015.

[10] V. Cohen-addad, et al., “Hierarchical Clustering : Objective Functions and Algorithms,” *Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[11] M. N. Nguyen, et al., “Nesterov’s Smoothing Technique and Minimizing Differences of Convex Functions for Hierarchical Clustering,” *Optimization Letters*, vol. 12, no. 3, pp. 455-457, 2018.

[12] T. K. Dey, A. Rossi, & A. Idiropoulos, “Temporal Clustering,” *25th Annual European Symposium on Algorithms (ESA 2017)., Leibniz International Proceedings in Informatics*, Article no. 35, pp. 35:1–35:14 arXiv by Cornell University, arXiv Prepr. arXiv1704.05964, 2017.

[13] M. S. Levin, “Towards balanced clustering-part 1 (preliminaries),” *distributed computing*, vol. 12, pp. 31-145 arXiv by Cornell University, arXiv Prepr. arXiv1706.03065 , 2017.

[14] M. Śmieja, K. Hajto, & J. Tabor, “Efficient mixture model for clustering of sparse high dimensional binary data,” *Data Mining and Knowledge Discovery*, vol. 33, pp. 1583-1624, 2019.

[15] J. A. Howe, “Improved Clustering with Augmented k-means,” stat 1050 (2017): 22 arXiv by Cornell University ,arXiv Prepr. arXiv1705.07592, 2017.

[16] S. S. Azab & H. A. Hefny, “Center of Gravity PSO for Partitioning Clustering,” arXiv by Cornell University, arXiv Prepr. arXiv1706.00997), 2017.

[17] S. Zhou, Z. Xu and F. Liu, "Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 3007-3017, Dec. 2017. doi: 10.1109/TNNLS.2016.2608001.

[18] I. Khanchouch, M. Charrad, and M. Limam, “An Improved Multi-SOM Algorithm for Determining the Optimal Number of Clusters,” *Computer and Information Science 2015, Springer*, pp. 189–201, 2015.

[19] R. Kiani, S. Mahdavi, and A. Keshavarzi, “Analysis and prediction of crimes by clustering and classification,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 8, 2015.

[20] M. N. Sekula, “OptCluster: an R package for determining the optimal clustering algorithm and optimal number of clusters,” *Electronic Theses and Dissertations*, 2015.

[21] Muca, et al., "A proposed algorithm for determining the optimal number of clusters," *European Scientific Journal, ESJ*, vol. 11, no. 36. pp 112-120, 2015.

[22] C. Subbalakshmi, et al., "A method to find optimum number of clusters based on Fuzzy Silhouette on dynamic dataset," *Procedia Comput. Sci*., vol. 46, pp. 346-353, 2015.

[23] Bai, Liang, Jiye Liang, and Chuangyin Dang, "An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data," *Knowledge-Based Systems*, vol. 24, no. 6, pp. 785-795, 2011.

[24] Chiang, Mark Ming-Tso, and Boris Mirkin, "Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads," *Journal of classification* vol. 27, pp. 3-40, 2010.

[25] A. H. Aliwy and K. B. S. Aljanabi, “An Efficient Algorithm for Initializing Centroids in K-means Clustering,” مجلة الكوفة للرياضيات الحاسبات| *J. Kufa Math. Comput.*, vol. 2, no. 2, 2016.