K.Mohana Prasad et al /J. Pharm. Sci. & Res. Vol. 9(1), 2017, 70-73

ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

Formulation of Clusters with Minimum Limitations Using AKM and LIC Techniques

K.Mohana Prasad[1], Dr.R.Sabitha[2]

[1]*Research Scholar, Faculty of Computer Science and Engineering, Sathyabama University, Chennai, Tamil Nadu 600119, India*
[2]*Professor, Dept of IT, Jeppiaar Engineering College, Chennai, Tamil Nadu 600119, India*

Abstract: *K-means* clustering is a procedure of cluster creation aims to split *n* observations on *k* clusters in which each one observation fits to particular cluster. The procedure k means could diminish the space between cluster entities, but it does not surety that outcome has a global minimum optimum solution. The k means is a most popular clustering algorithm taken centroid distance to group similar items in to different possible clusters. The distance measure is the important thing in k mean clustering for that there are several distance metrics are available to validate the similar items in clustering some of the most using distance metrics are Euclidean, Cosine similarity etc. In this concept we can improves the performance of the clustering by using AKM (Accelerate k means) and LIC (Least iteration Clustering)to improve the performance by minimizing the iteration count. There are many more clustering algorithms rather than k mean are available differed each algorithm on their perspective performance values.

Keywords: K-Means, AKM, LIC, GSA, Fuzzy C-Mean,ACM

1. INTRODUCTION:

The clustering is obtained from two categories Hierarchy and partitioned based clustering [1-3]. In hierarchy method again two different types are there which are Agglomerative and Divisive [4].

Agglomerative method might start with each data entity in a discrete cluster and goes to merging the most similar pairs until adequacy of the end process [5]. The execution which called as divisive start positioning all the elements in single cluster and goes to divide each cluster into different smaller number of clusters, the clustering continues until the adequacy of termination criteria reached [6].

In partitioned one the popular clustering called centre based clustering in that k means is the most useful and popular algorithm used for clustering.[7]

The clustering is natural in many more popular domains of data mining, biomedical etc there is a problem of bulk of data to get the finite similar items which user try to retrieve the clustering can be used to group the bulk data in to relevant group items[8]. several methods based on clustering technique are current to use for the development of clusters the k- means is the generally used algorithm for the formation of clusters but it has some of the boundaries which is selection of preliminary number of clusters and each cluster centres which is also called initialization problem that drains the performance of the algorithm and the solutions which we get are not finest. [9]

Clustering is an significant data retrieving task that refers to a progression of formation of numerous data collections and made a set of observations in those groups and correct the group members for each repetition and check the similarity of entities in that group if not similar repetition continues until the formation of un similar groups with similar data in each group[10].

The formulation of clusters is made by calculating the distance from assigned point value to each item to another and groups the items of point values which will be closer ones[11].

The methods which we used for clustering are varied on the performance each clustering method shows various performances [2]. The Ant Colony Optimization, Artificial Bee Colony optimization, k harmonic those are different in their performances.[12]

These methods could be able to control difficulty of the clustering on finding a good measure for the level of separation among clusters [13]. The better performance of the methods can be depends on the factors of the values of inter cluster distance and the standard deviation decrease in those values causing the increasing performance of those algorithms [14].

2. EXISTING SYSEM:

There are many existing clustering algorithms varied in their performance some of the existing ones which we used for clustering are ant colony optimization, artificial bee colony optimization, k harmonic mean etc.

Ant colony optimization is a clustering technique based on general behaviour of ants updating through pheromone.

The general behaviour of the ants should be used for performing cluster formation.

For adapting the ACO algorithm to the problems of scheduling some parameters has to be followed.

1. The number of ants in the colony
2. Vapouring factor of pheromone

The ABC clustering is also one of the existing which used the swarm intelligence technique for finding the solutions

In this algorithm three types are bees are using for clustering employee bees, scout bees , onlooker bees the employee bees are the solutions of selecting food sources if the vector quantity level decreased it changed to scout bees and look for another food source the onlooker bees are those searching for a food from employee bees.

Like this many clustering algorithms are existing differ in their performances some of the other existing clustering algorithms are k harmonic mean, fuzzy c mean, gravitational search, genetic etc each one having difference on their performances.

The most popularly well known clustering algorithms used widely is k mean clustering which is very best one for performing clustering

The selection of random clusters centres as centroid and calculating the effective distance from that point to all the nearest points the points which are closer to the centroid can be grouped as one cluster like that we can construct several other clusters.

The main important thing in this k means clustering is the distance metric that used to find the similarity between the points the basic aim of clustering is to minimize the distance between points for increasing the performance of clustering.

The performance can be increased when the iteration limit is minimized it should based on corresponding distance metric used.

K means is very popular with the use of Euclidean distance metric for clustering which is existing and used by most of the people.

The Euclidean distance should be as follows.

$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The points which indicate the attribute values can taken as numerical values the distance calculated by using this formula and find the similar pairs in that group.
Some of the other distance metric used in existing system are as follows
The Manhattan distance is one distance metric

$$d(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

The Minkowski distance as follows

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^{p}.$$

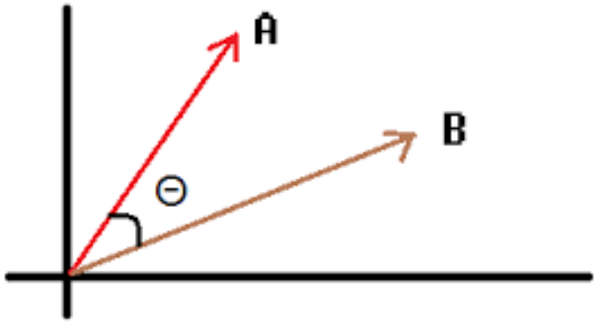
Like this there are many more distance metrics are used for calculating distance each one is differ in their obtaining distance values.

3. PROPOSED SYSTEM:

The proposed method here which we named as Modified K Means can be implemented by targeting the performance which navigates to incrementing when compares to previous existing methods.
In this proposed method we are going to implementing the k means algorithm by using an efficient distance metric which called as cosine similarity metric.
The following is the cosine similarity metric which we used in this system

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

We can discuss something about cosine similarity metric
The cosine similarity is the cosine of angle subtended at the origin between two documents in the frequency distribution space.
A document can be represented by thousands of attributes each regarding the frequency of particular word such as keywords
Following fig shows the illustration of calculating distance between two points using cosine similarity metric.



If you look at the visual with the 2 axis and 2 points, we need the cosine of the angle *theta* that's between the vectors associated with our 2 points. And for our experiment it does give better results.
In the above graph we have to take two point vectors or documents and calculating the similarity between those two vectors A and B using cosine similarity metric.
It is an angular metric those values bounded within the interval [0,1]. If the cosine value should be indicate as -1 means that the two items are totally opposites if it is 0 means that the two items are independent to each other if it is 1 it means that the two items are very similar.

Cosine measure: If d1and d2 are two vectors, then
Cos (d1, d2) = (d1 □ d2) /||d1|| ||d2||,
Where □ indicates vector dot product,
||d||: the length of vector d
Example:
d1 = (3 2 0 5 0 0 0 2 0 0)
d2 = (1 0 0 0 0 0 0 1 0 2)

d1□d2 =3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5
||d1||= (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)
0.5=(42)0.5
= 6.481
||d2||= (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)
0.5= (6)0.5
= 2.245
Cos (d1, d2) =0.3150

3.1 The steps that need to be followed in the proposed approach:

To resolve the drawbacks of current methods can ongoing by a two-step process.
Step1: To increase the performance of clustering by minimizing the limitations can be done by using additional algorithms with k-means like gravitations search algorithms and genetic algorithms etc.
Step2: after the step 1 finished modified k means with a consistent mathematical distance metric that overwhelms all the other limitations which we seen.
The below steps to be followed which is similar to a existing k-mean algorithm
The each step should be described as follows

1. The data entities that taken like X = {x_1, x2, x3,……..,x_n} should be assembled in to dissimilar clusters.
2. The relationship of the data objects should be patterned bycompute the distance between each data point and cluster midpointsusing a best distance metric which has taken from the current metrics.

- The cosine similarity distance metric is used to increase the performance of clustering.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

The assortment of preliminary value of cluster center ought to be overcome

3. Disperse the data point to the cluster midpoint whose space from the cluster midpoint is minutes of all the cluster midpoints.
4. The end of repetition checks the object resemblance if not analogous continues up to the nth iteration which the formation of parallel data objects is presented. Recalculate the detachment between each data point and new attained cluster midpoints.
5. If no data point was reallocated then stop, else recap rehearsals.

3.2 Proposed Iterations Algorithm (LIC)

Least Iteration clustering (LIC) is one of the simple and easy ways to classify a given dataset by the certain number of clusters. The main principle of the proposed iteration clustering algorithm is to describe k-centers, one for each cluster. Assign the random weights for all the data points. Then the centers are selected based on the distance, the data point that is having less distance with other data points of different locations. The next step with selected center to a given dataset and combine it to the nearby center. When no data point is missed during the clustering, then first step is calculated. Afterwards, need to re-calculate k new centroids cluster resulting. As a result k-centers will change their

location until no more changes are done in another way can say that centers couldn't move anymore. Finally the proposed iterations algorithm minimizes the squared error function.

Algorithm: Least *Iteration Clustering Algorithm*

Step 1: Assign some random weight to each data point (weight varies from 0 to1)

$(x_{Rw}, y_{Rw}) = Rw(x, y)$ → 3.1

Where,

Rw→ Random weight

(x,y) → data point

Step 2: Choose the number of cluster; the cluster is selected based on the maximum length of the data points

Choose the random cluster= $L_d / r<L_d$ →3.2

Step3: Choose the centroids in each cluster i.e. location

$$Xc = \sum_{i=1}^{n_i} \sum_{\substack{j=1 \\ \arg\min}}^{u_j} (xij \quad xjn) \rightarrow 3.3$$

Where,

i→Represents the number of cluster

j→ Represents the number of data points

Xc → the centroids in each cluster group

Step 4: The data point that is having less distance is chosen as centroids in each group

Xc <Xn

Where,

X_c →Represents the cluster centroids

Xn → Represents the number of cluster points in each cluster group.

Step 5: iterate the group until the criteria meet the condition

Step 6: Re-cluster the group

$$Recluster\ Xr = \sum_{k=1}^{kn} \sum_{\substack{l=1 \\ \arg\min}}^{ln} (xkl - xkn)$$

4. EXPERIMENTAL RESULTS

Table 4.1 Dataset Used in the Experiments and Their Characteristics

Dataset	Objects	Classes
Whole sale customers data	8	440
Seeds-Data	7	210
Student evolution	33	5820
Water-Data	38	527
Dresses-Attributes	14	501
Bank Data	4	434874
Sponge Data Set	45	76
Turkey Student Evaluation on R-Specific	7	10421

Table 4.2 Parameters Used in the Clustering Algorithm

Algorithm	Parameters	Values
Iteration Clustering	Number of iterations	10
SC-FCM	Number of stems cells	20
FCM	Number of iterations	1000
PSO	Number of swarm	100
ABC	Number of bees	20
VGAPS	population	20
GCUK	population	100
K-Means	Number of iterations	1000

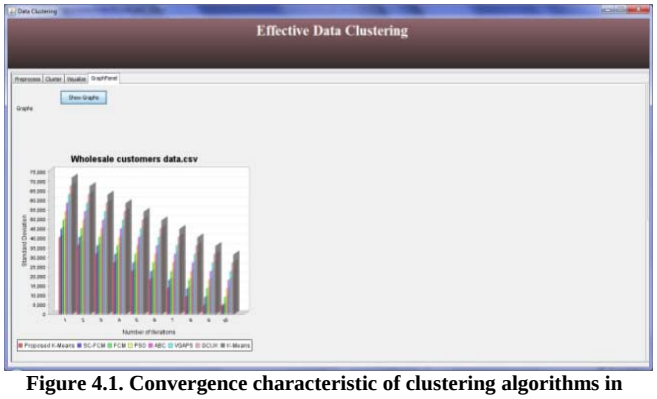


Figure 4.1. Convergence characteristic of clustering algorithms in reaching best solution on whole sale customer's data

Figure 4.2. Convergence characteristic of clustering algorithms in reaching best solution on Seeds-Data

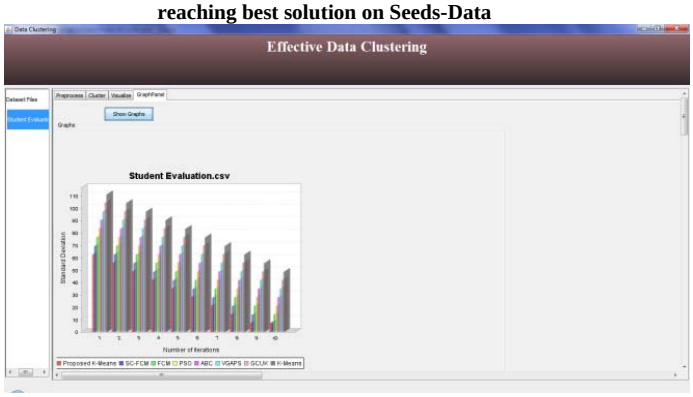


Figure 4.3. Convergence characteristic of clustering algorithms in reaching best solution on Student evolution

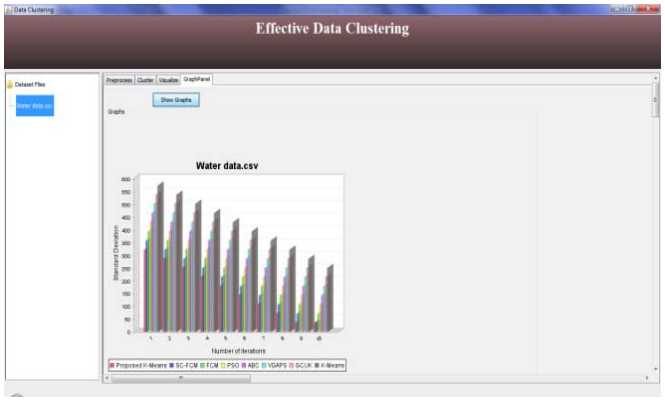


Figure 4.4. Convergence characteristic of clustering algorithms in reaching best solution on Water-Data

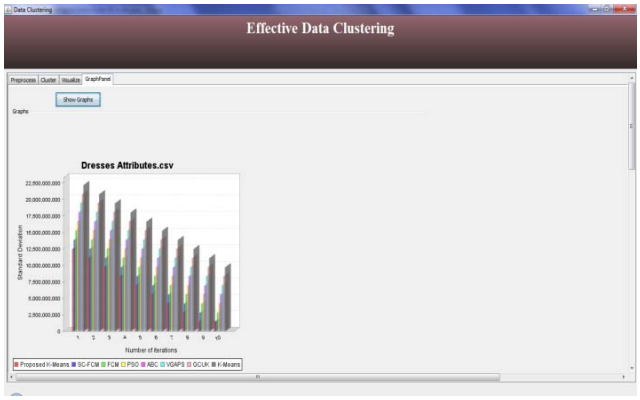


Figure 4.5. Convergence characteristic of clustering algorithms in reaching best solution on Dresses-Attributes

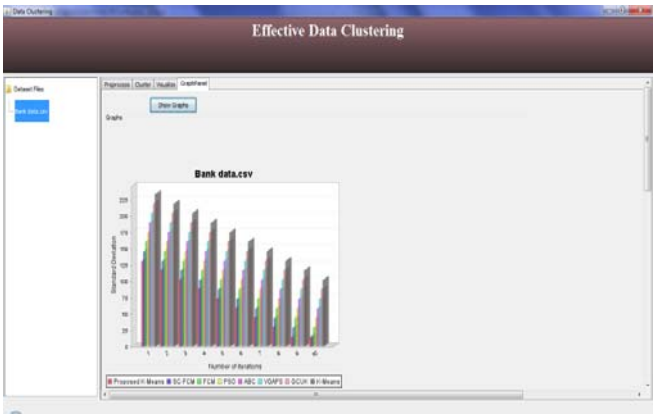


Figure 4.6. Convergence characteristic of clustering algorithms in reaching best solution on Bank Data

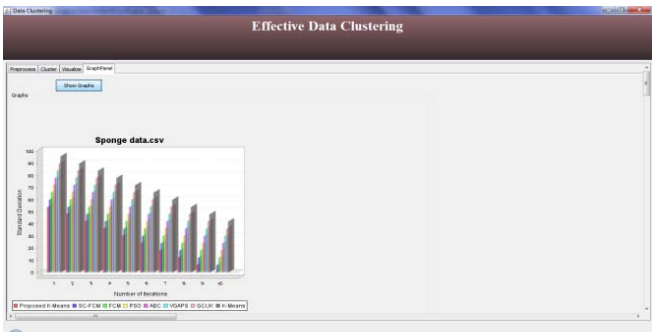


Figure 4.7. Convergence characteristic of clustering algorithms in reaching best solution on Sponge Data Set

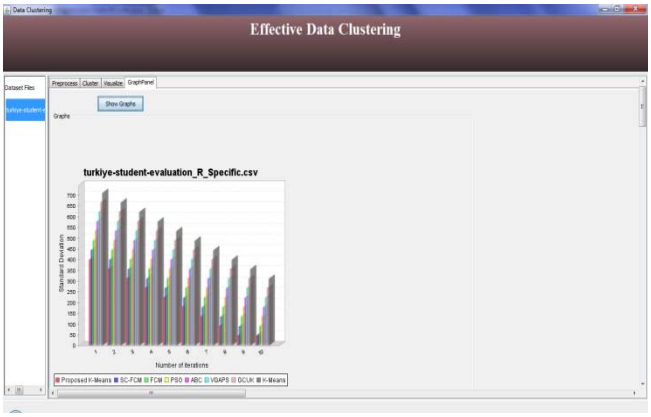


Figure 4.8. Convergence characteristic of clustering algorithms in reaching best solution on Turkey Student Evaluation

5. CONCLUSION:

K means is the most popular and best clustering algorithm but some times the performance can be depleted with some distance metrics and others we have to evaluate which one is suited and better for the clustering to increase performance measure with the rule of data dispensation well liked and lucrative rule as a result of its fewer convolutions in their operations. However some disadvantages throughout this rule cause some problem. For solving those negatives we use this rule in better mining of facts from data sets.

REFERENCES:

1. Waltman, L., & Van Eck, N.J. (2013). A unified approach to mapping and clustering of Bibliometric networks.
2. Melnykov, V., Chen, W.C., Maitra, R., 2012. MixSim: an R package for simulating data to study performance of clustering algorithms. Journal of StatisticalSoftware 51, 1–25.
3. Grabusts, P., 2011. The choice of metrics for clustering algorithms, in: Proceedings of the 8th International Scientific and Practical Conference, pp. 70–76.
4.Erisoglu, M., Calis, N., Sakallioglu, S., 2011. A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Letters 32, 1701–1705.
5. R., Melnykov, V., 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. Journal of Computational andStatistics 19, 354–376
6. Lozano, J.A., Pena, J.M., Larranaga, P., 1999. An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recognition Letters20, 1027–1040.
7. Gnanadesikan, R., Harvey, J.W., Kettenring, J.R., 1993. Mahalanobis metrics for cluster analysis.Sankhya, Series a 55, 494–505.
8. Kaufman, L., Rousseuw, P.J., 1990. Finding Groups in Data. John Wiley & Sons, New York.
9. Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.
10. MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium 1, 281–297.
11. Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.
12 . Mohana Prasad. K, Sabitha. R, "Meta Physical Algorithmic Representation for Flawless Clustering" Journal of Theoretical and Applied Information Technology , 2015, Volume 76, NO 1,PP 82-87.
13 Mohana Prasad. K, Sabitha. R,"Evolution Of An Algorithm For Formulating Efficient Clusters To Eliminate Limitations" International ournal of Applied Engineering J Research , 2015, Volume 9, Issue 23, pp. 20111-20118.
14. .Mohana Prasad. K, Sabitha. R," Yoking of Algorithms for Effective Clustering",Indian Journal of Science and Technology,2015, Vol 8(22),pp 1-4.