VOL. 10, NO. 4, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

CLUSTERING OF DATA SETS BY USING FUZZY ALGORITHM

M. Saravanan[1] and V. L. Jyothi[2]
[1]Sathyabama University, Chennai, India
[2]Jeppiaar Engineering College, Chennai, Tamilnadu, India
E-Mail: mail2saravananme@yahoo.co.in

ABSTRACT

In this Technological era Clustering is inevitable. For any function arrangement of Data is a primary task. The collected Data has to be grouped based on their features, Clustering is a method of arranging same or similar attributes and that attributes which are closer to each other are also grouped together. Clustering is formed of three major process initializing Data is the principle process, Data sets are selected randomly and distance metrics are used. Iteration reduction is a great challenge as for clustering is concerned. Fuzzy c-means is applied with the intention of reducing iteration. This Fuzzy c-means permits one data to function in two sets. When iteration is reduced clustering will be more effective. This paper deals with intervention of Fuzzy c-means algorithm in a specified Data set which thereby is to reduce iteration to make the function flaw less and reliable.

Keywords: clustering, C –means, iteration, data sets, metrics.

1. INTRODUCTION

Data mining is used to remove data from huge datasets and review into helpful information. The superior stage meta information that may be palpable when looking at raw data. Data mining can be considered to be an inter-disciplinary field involving concepts from machine learning, database technology, clustering and visualization among others. Data mining goes beyond the scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques of data analysis. A data analysis system that does not handle large amounts of data should be categorized as machine learning system, a statistical data analysis tool or an experimental result prototype. The huge amounts of stored data contains knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. Database Management Systems used to manage these data sets at present only allow the user to access information explicitly present in the databases

Clusters is a collection of same elements occurring strictly together. The types of clusters are like elite, overlie etc. K-harmonic means (KHM) is popular clustering algorithm technique.

With the help of these you can find sum of all squared distance by using K-harmonic.

And also used to solve the small area from the input data point to the cluster centre. It is better for election of initial cluster center. If any input data is nearby any one center based on that it gives

Result. The K-means algorithm try to find the cluster centers, such that result of all distances of each data point Xi to its adjacent cluster center is reduced. K-Harmonic Means clustering algorithm is based on soft membership a data point belongs to all clusters, dynamic weighting data not close to any center are boosted by a higher weight in the next iteration and the cluster center is updated using all data points weighted by both soft membership and dynamic weighting this will easily moves into local optima.

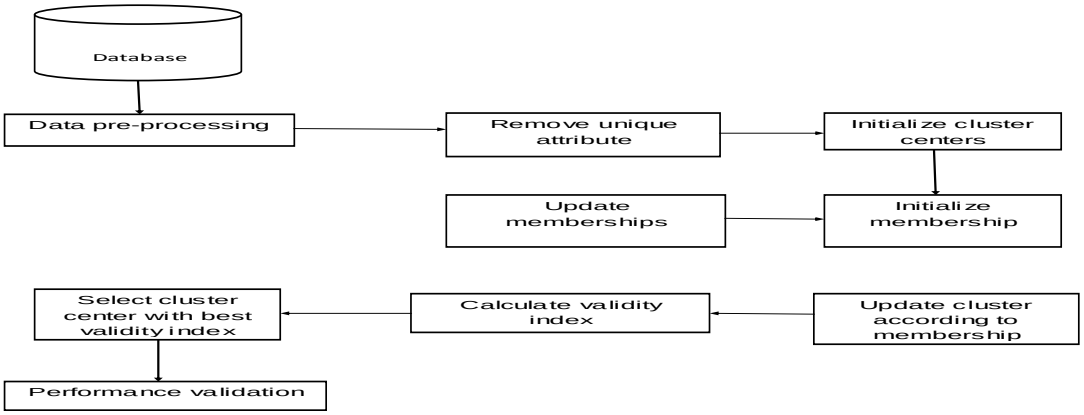
Improved gravitational search algorithm was used. This algorithm is based on the hybrid data clustering. Optimization algorithm based on the law of gravity, namely Gravitational Search Algorithm (GSA) is proposed. This algorithm is based on the Newtonian gravity. Every particle in the universe attracts every other particle with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them.

In gravitational search algorithm, all the individuals can be viewed as objects with masses. The objects attract each other by the gravity force, and the force makes all of them move towards the ones with heavier masses. The objects transform information by the gravitational force, and the objects with heavier masses become heavier. When any object jumps out of its range, the original GSA just pulls it back to the fringe. The object will be assigned a boundary value. Some seven data sets are correlated. It overwhelmed the convergence speed and it need more run time.

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Fuzzy c-means is used to group the data points in definite number of clusters. The purpose is to recognize the groupings of data from large dataset and it produce a short representation. Datasets will be grouped into clusters. Collective fuzzy c-means algorithm is used to form the cluster group. It is done by using partitional clustering algorithm. It avoid selecting the random variable. The cluster group will be formed with less number of iterations.

2. RELATED WORK

The definition of the problem can be defined as a detailed and operational description of the differences between the existing situation and the desired situation.



System architecture

Materials implemented

- Preprocessing
- Grouping the object using FCM
- Calculate the center vectors
- Modified membership

2.1 Preprocessing

Preprocessing is the main step in data mining process. It is used to remove the unwanted attributes. Data gathered will be loosely controlled so there occurred impossible combinations. Sometimes missing values may also be occurred. It leads to misleading results. Analyze the data before processing and find the missing values and unwanted data.

```
Min da=min (data (:,sc_num+1:dacol));
Max da=max (data (:,sc_num+1:dacol));
max_min da=max da-min da;
max_min da(max_min da==0)=1;
```

Those data's should be removed. Select the following parameters. It preprocess the data to remove the unwanted attributes for example if some values is missed then the whole data's will get deleted. The required number of clusters N, 2<N<k . it measures the distance as Euclidean distance, a fixed parameter and initial (at zero iteration) matrix is equal to the object ownership with the initial cluster centers.

2.2 Grouping the objects using FCM

Objects from the different sets will be grouped by using the fuzzy means. It can make the entire group visible or invisible. Objects can be grouped by selecting a parenting axes by children. A number of similar individuals that occur together as a two or more consecutive consonants or vowels in a segment of speech b: a group of houses c: an aggregation of stars or galaxies that appear close together in the sky and are gravitationally associated.

2.3 Calculate the canter vectors

Centroid value will be find by calculating the center vectors by using the formula. The center value will be calculated. The graph is drawn between the objective function value and iteration count. Depending upon the selection of datas graph will differ slightly according the center values that occurred for the particular data.
In the t-th iteration step in the known matrix is computed in accordance with the above solution of differential equations.

$$Cj= \sum_{i=1}^{N} (u_{ij}^{m} * xi / u_{ij}^{m})$$

2.4 Modified membership

It means grouping of formation. Group is a collection of data. Each data sets will have the distinct centroid value. By selecting the data it will have the values that is calculated for the corresponding data. Data will be modified based on the centroid value. Modified data produced according to the following equation,

$$Uij=1/(\sum_{k=1}^{c} (||x_i-c_j||/||x_i-c_k||))^{2/(m-1)}$$

Here, yi represents as no of datas: C represents as Centroid value.

3. METHODS USED

The method used is Fuzzy C-means. It is to identify natural groupings of data from a large dataset to produce a concise presentation. Dataset will be grouped to n-clusters with every data point in the dataset belonging to every cluster. The set of data points are taken. Assume some points to be data centers. The cluster will be formed by selecting the cluster center. Using the centroid formula the center value is found so that the clusters can be formed by selecting the data sets. Membership function μij is calculated by the formula. Objective function value is done by using the data's. Each time the graph is different by selecting the data's. K is the iteration step. Iteration

will depend upon the data given in the dataset. When the values getting constant, then the iterations will be stopped and it is the result of efficiency and performance.

4. RESULTS

Figure-1 shows that collecting the data's from the data set that is stored in a system. It merged each data's. The preprocessing is done by finding the missing values and remove the unwanted attribute. Then all the data are collected to one point.

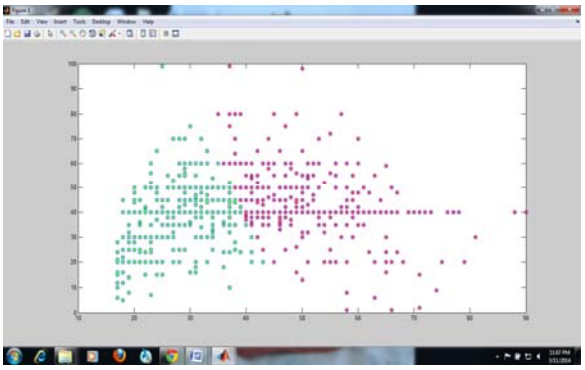


Figure-1. Collecting the Data's.

Figure-2 shows that the data will be varied each and every time on selecting the different data sets.

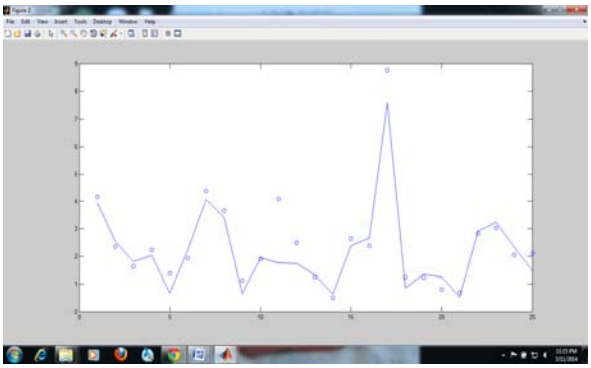


Figure-2. Data variation.

Figure-3 shows that the iteration value for each objective function value. It depends upon the data and it increases the performance.

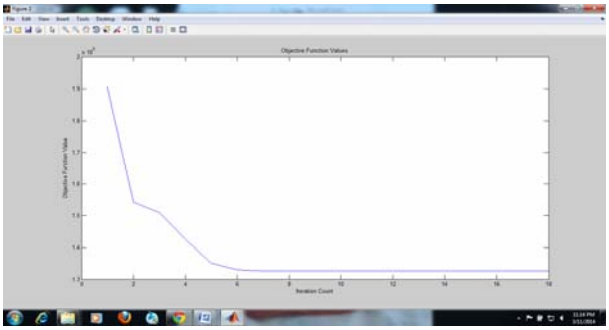


Figure-3. Finding iteration count.

Figure-4 shows that data will be modified according to the data's.

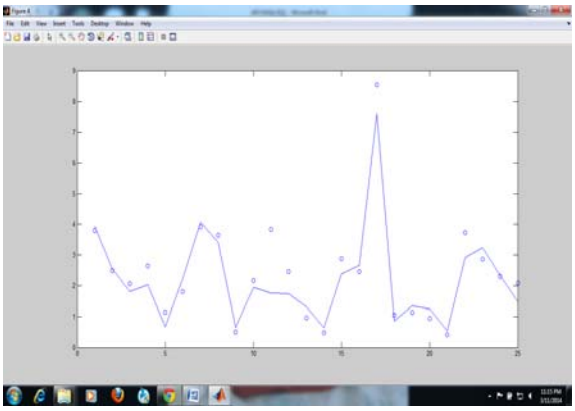


Figure-4. Modified data.

Figure-5 shows that the fitness function for each identification differs based on the iteration value.

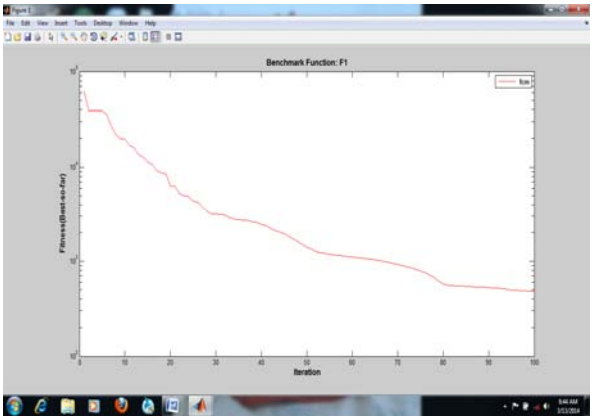


Figure-5. Benchmark function.

Figure-6 shows that the cluster will be formed for the random values.

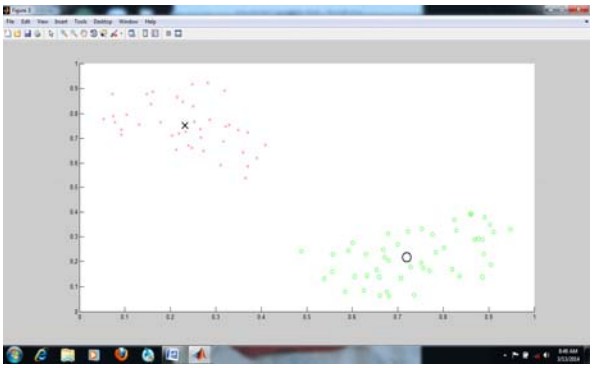


Figure-6. Cluster for random values.

Figure-7 shows that cluster is formed by calculating the center values for the data sets.

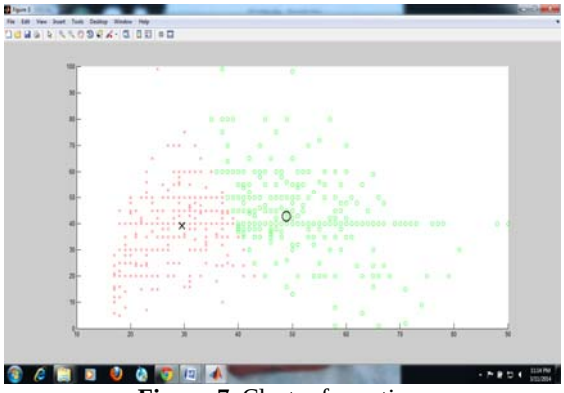


Figure-7. Cluster formation.

Figure-8 shows that formation is done by calculating the detection rate and the threshold values. The cluster is formed in step by step process. It is done by calculating centroid formula using fuzzy c means algorithm. Each and every graph explains the slight variation when changing the data and it will change every time running the program. The cluster will be formed based on the given input and also based on the number of clusters. The threshold and detection rate is calculated and the graph is drawn according to the threshold values.

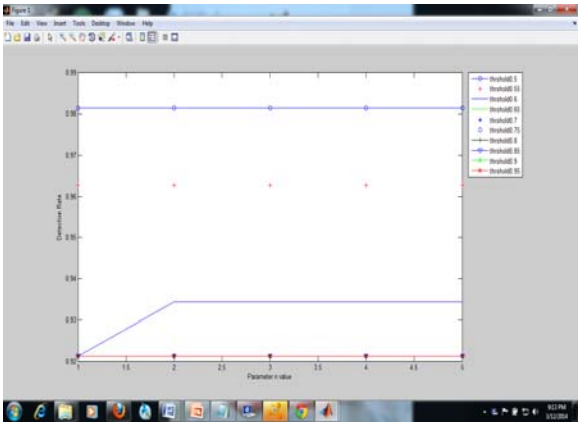


Figure-8. Thresold and detection rate.

5. CONCLUSIONS

Cluster formation is done by using collective fuzzy c-means algorithm. The steps to form the cluster is to load and plot the data and then start clustering and finally cluster center is saved. First it preprocess the data to remove the unwanted attributes and then after removing it finds the centroid value in order to group the objects by using the fuzzy c-means formula. Fuzzy algorithm allows the gradual membership of data points to cluster measured as degrees in 0 and 1. This gives the flexibility to express the data points can belong to more than one cluster. For starting the clustering data, choose the clustering function fuzzy c means and save the cluster center. The objects will be grouped together by calculating the vectors and cluster is formed effectively.

5. FUTURE WORK

Clustering methods employing have to be necessarily robust and fuzzy, to be able to handle large percentage of outliers and overlap. They also need to be of low complexity to deal with extremely large data sets. In this paper it requires some time for execution and it is done with some data sets. In future, we may integrate other fuzzy algorithm to get the more efficient cluster by using the real time data's and can try to reduce the iterations than getting in proposed work and also fuzzy algorithm can help in selecting superior quality clusters.

REFERENCES

[1] Mohammad Doraghinejad., Mailhe maghfoori., Hossein Nezambadi-pour. 2012. A Hybrid Algorithm Based on Gravitational Search Algorithm for Unimodal Optimization. Shahid Bahonor University of Kerman, Iran.

[2] Jyothi Bankapali *et al.* 2011. International Journal on Computer Science and Engineering (IJCSE). Combining K-harmonic mean and hierarchical algorithm for robust and efficient data clustering with cohesion self-merging. ISSN: 0975-3397, Vol. 3. No. 6.

[3] Huiqin Chen., Sheng Li. and Zheng Tang. 2011. Hybrid Gravitational Search Algorithm with Random-key Encoding Scheme Combined with Simulated Annealing. IJCSN International Journel of Computer Science and Network Security. Vol. No. 6.

[4] K. Thangavel. and N. Karthikeyani Visalakshi. 2009. Ensemble based Distributed k-Harmonic Means Clustering. International Journal of Recent Trends in Engineering. Vol. 2 , No. 1.

[5] Yang F.Q., Sn T.L. and Zhang C.H. 2009. An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. Expert System with Applications. 36(6): 984-9852.

[6] Samarjee Boroh., Mrinal Kanti Ghose. 2009. Performance Analysis of AIM-Kmeans and K-means in quality cluster generation. Journal of computing. Volume1, Issue 1.

[7] Esmat Rashedi., Hossein Nezambadi-pour., Saeid Saryazdi. 2009. A Gravitational Search Algorithm". Information Sciences. p. 179.

[8] Zhou H. and Lin Y. H. Accurate integration of multi-view range images using k-means clustering. Pattern Recognition. 41(1): 152-175.

[9] Unler A. and Gungor Z. 2008. Applying K-harmonic means clustering to the part-machine classification

problem. Expert Systems with Applications. doi:10.1016/j.eswa.2007.11.048.

[10] Halberstadt W. and Douglas T. S. 2008. Fuzzy clustering for the detection of Tuberculous Meningitis – associated hyperdensity in CT images. Computers in Biology and Medicine. 38(2): 165-170.

[12] Cui X. and Potok T. k. 2005. Document Clustering using Particle Swarm Optimization. In IEEE swarm intelligence symposium Pasadena. California.

[12] Hammerly G. and Elkan C. 2002. Alternatives to the k-means algorithm that find better clusterings. In: Proceedings of the 11th international conference on information and knowledge, management. pp. 600-607.