New approach for clustering – based on quick sort

Sabeen Govind P V [#1] , Archana K N [#2], Reenu M [#3] , Philomina Simon [#4]
[1,2,3,4] *Department of computer science*
University of Kerala, India

Abstract

Clustering refers to the process of grouping samples so that samples in a group look similar. In this paper we propose a new clustering algorithm based on quick sort. Experimental result shows that our algorithm gives comparable results with existing algorithms. Complexity of this algorithm is also less.

Keywords *Clustering, Quick sort, K-Means.*

1. INTRODUCTION

Clustering can be considered as an unsupervised learning problem [1]. A cluster is a collection of object which are 'similar' between them and 'dissimilar' to the objects belonging to other clusters [4]. Cluster analysis (CA) is an exploratory data analysis tool for organizing observed data (e.g. people, things, events, brands, companies) into meaningful taxonomies, groups, or clusters, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown. Clustering can be applied in various fields like marketing, biology, libraries, etc. Cluster analysis is the obverse of factor analysis. Whereas factor analysis reduces the number of variables by grouping them into a smaller set of factors, cluster analysis reduces the number of observations or cases by grouping them into a smaller set of clusters. Based on the properties of cluster generated clustering algorithm can be broadly classified in to hierarchical clustering and partitional clustering. In hierarchical clustering we start with every data point in a separate cluster and we keep merging the most similar pairs of data points/clusters until we have one big cluster left. This is called a bottom-up or agglomerative method. Partitional clustering, attempts to directly decompose the data set into a set of disjoint clusters.

2. RELATED WORKS

This section introduces the related clustering techniques found in the literature.

2.1 Single linkage, complete linkage and average linkage algorithms

The single linkage algorithm is obtained by defining the distance between two clusters to be the smallest distance between two points such that one point is in each cluster. This method is referred to as minimum method. But in complete linkage algorithm we are taking the largest distance between two points. This method is referred to as maximum method. The Average linkage algorithm is obtained by defining the distance between two clusters to be the average distance between two points such that one point is in each cluster.

2.2 .Ward's Algorithm

Another commonly used approach in hierarchical clustering is Ward's method [8]. This approach does not combine the two most similar objects successively. Instead, those objects whose merger increases the overall within-cluster variance to the smallest possible degrees are combined

2.3 .K-Means Algorithm

K-means is a simple algorithm proposed by MacQueen(1967) that has been adapted to many problem domains. In K-means algorithm each cluster is associated with a centroid and each point is assigned to the cluster with the closest centroid [5].

2.4. Fuzzy C-Means clustering

Fuzzy c-means clustering is based on optimization of the basic c-means objective function. We can optimize this function by using a variety of methods, including iterative minimization or genetic algorithm [10].

2.5. Grid based clustering

Grid-based clustering methods start by forming a grid structure of cells from the objects of the input dataset. Each object is classified in a cell of the grid. The clustering is performed on the resulting grid structure. STING [6] is a grid based clustering.

2.6. Model based clustering

Model-based clustering methods typically assume that the objects in the input dataset match a model which is often a statistical distribution. COBWEB [1] is a model based conceptual clustering.

2.7. Density based clustering

It discovers cluster based on the density of points in regions. They are capable to produce arbitrary shapes clusters and filter out noise. DBSCAN [7] is a density based clustering proposed by Ester.

3. PROPOSED ALGORITHM

The proposed algorithm is based on quick sort [9]. Quick sort uses divide- and- conquer strategy. This algorithm works on data sets which consist of numerical data values. The data value is considered to be on a 2D plane which has both 'x' co-ordinate and 'y' co-ordinate. We are taking 'x' value and sorting is performed on 'x' coordinate values. We get sorted 1-dimensional array of 'x', from which the median value is calculated. It can be done on the 'y' axis values also which also gives a similar result but the cluster values will be different. The algorithm then proceeds further by dividing the whole data set into two halves, first one to the left of Mid _Value and second to the right. These halves cab be further divided by taking Median of newly formed clusters and so on. The algorithm terminates when the desired number of cluster is reached or when further clustering cannot be performed.

3.1. Partitioning Process

The partitioning is applied by calculating the distance of the 'x' values from each other. Since sorting is already performed the data values at each cluster will be closer. The Mid value will go to the cluster which has minimum distance when it is compared to the right 'x', Data(Mid_Value -1), and left 'x' , Data(Mid_Value +1), is calculated. This process can be applied to two or three nearest values of Median. If both clusters have equal distance Mid_value can go to either group.

3.2. Pseudocode

Input : 'n' number of data points (x,y).
Output : Required number of Clusters.

Step 1: Sort the given 'x' values in the data points using Quick sort algorithm.

If x value is repeated sort 'y' values of only identical x values.

Sorted_Array= Quick_Sort(Datapoints(1,x));

Step 2 : Compute the mid _value of the sorted array.
Mid_Value=(n+1)/2 for even 'n' values and
Mid_Value=n/2 for odd 'n' values.

Step 3: Compute the distance between successor and Mid_value and predecessor and Mid_value.

If (Mid_Value + Mid_Value – 1) > (Mid_Value + Mid_Value+1) divide array as

Arr1=Sorted_Array(1…Mid_Value-1)
Arr2=Sorted_Array(Mid_Value …n)

else

Arr1=Sorted_Array(1…Mid_Value)
Arr2=Sorted_Array(Mid_Value+1 …n)

Step 4 : Display the clusters.

4. EXPERIMENTAL RESULTS

Simulations are done in MATLAB. We are randomly selects N data points and the output is shown in the below figure. In this example we take N=50.

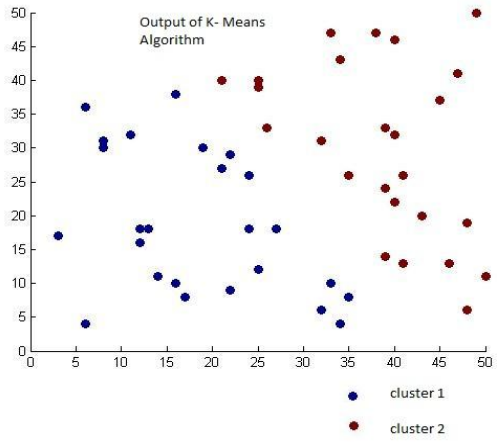

Fig (a) Output of K-means algorithm

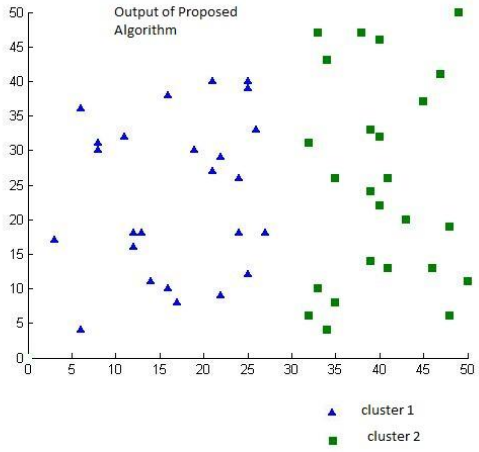

Fig (b) Output of Proposed method

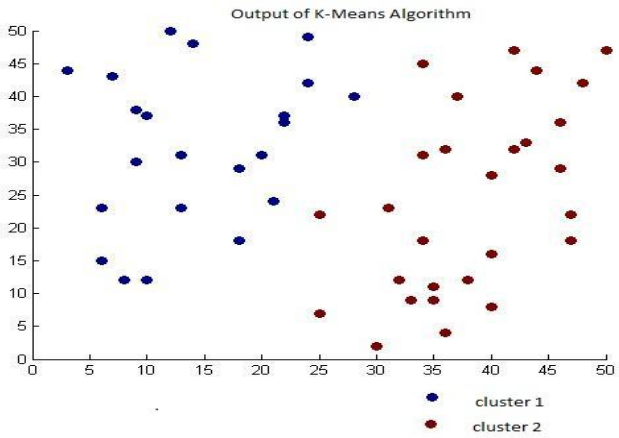

Fig (c) Output of K-means algorithm

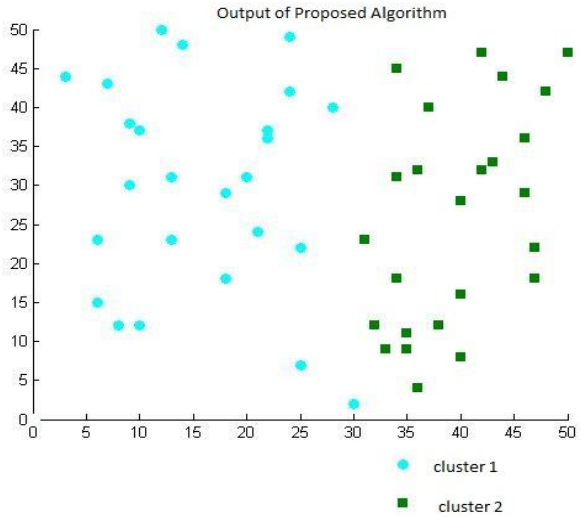

Fig (d) Output of proposed method

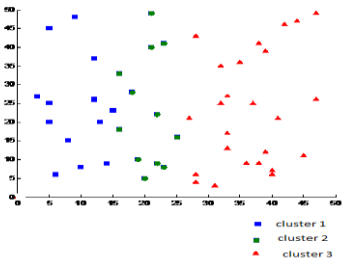

Fig (e) Output of proposed method (K=3)

Figure (a) shows the output of K-means algorithm (Number of clusters K=2) ,figure (b) shows the output of proposed method for the same data points (x,y)

Figure (c) and (d) represents the output of K-means and proposed algorithm simultaneously for some another randomly generated data points.

Figure (e) shows 50 randomly generated data points are put in to three clusters. (K=3)

Following table shows various clustering algorithms and their complexities [2].

Algorithm	Time complexity
K-means	O(Nkd) (time) O(N+K) (space)
Fuzzy c-means	Near O(N)
Hiearchical clustering	$O(N^2)$ (time) $O(N^2)$ (space)
Proposed method	$O(N^2)$

5. CONCLUSIONS

Clustering can be used to reduce the amount of data and to induce a categorization. In this paper a new method of clustering based on quick sort is presented. The method is simple to implement and experimental results shows that the proposed method is giving comparable results with K-means algorithm. We can extend the method for any number of clusters by just using a recursive function.

REFERENCES

[1] Rui Xu, Donald Wunsch "Survey of Clustering Algorithms",IEEE Transactions on Neural Networks Vol 16, No. 3, May 2005.

[2] Madhuri A. Taya, M.M.Raghuwanshi " Review on Various Clustering Methods for the Image Data" in Journal of Emerging Trends in Computing and Information Sciences vol 2,special issue 2010.p.p 35-39.

[3] Naz, Majeed, Irshad " Image segmentation using fuzzy clustering: A survey" International Conference on Emerging Technologies (ICET),p.p 181 – 186,2010.

[4] Song Yu-chen , Jia Xiao-liang , Meng Hai-dong " Comparative study of clustering methods based on linear data distribution" International Conference on Management Science and Engineering (ICMSE), p.p 377 - 384 ,2012.

[5] Hui Xiong, Junjie Wu, and Jian Chen,"K-Means Clustering Versus Validation Measures:A Data-Distribution Perspective". IEEE Transaction on Man,and cybernetics-Part B:Cybernetics, Vol. 39, No. 2, April 2009.

[6] W. Wang, J. Yang and R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining". Proc. Int. Conf. VLDB, pp. 186-195, 1997.

[7] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD"96), Portland, AAAI Press pp. 291-316, 1996.

[8] Earl Gose, Richard Johnsonbaugh, Steve Jost, Book on "Pattern Recognition and Image Analysis"

[9] Alfred Aho, D.Ullman, Book on "Data structures and algorithms".

[10] Liu Su-hua , Hou Hui-fang " A combination of mixture genetic algorithm and Fuzzy C-means clustering algorithm" IEEE international symposium on Medicine and Education, Vol 1,p.p 254-258,2009