Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

A Study of Clustering Algorithm in Data Science

* **R. Anto Priyadharshini, B.Nithya**
Gonzaga College of Arts and Science for Women, Kathampallam, Elathagiri, Krishnagiri, Tamil Nadu, India.
*Corresponding author Email: priyamic25@gmail.com

Abstract. Data clustering is a process of partitioning data points into meaningful clusters such that a cluster holds similar data and different cluster hold dissimilar data. It is an unsupervised approach to classify into the following two categories: Firstly, hard clustering, where a data object can belong to a single and distinct cluster and secondly, soft clustering, where a data object can belong to different cluster. In this report we have made a comparative study of three major data clustering algorithm highlighting their merits and demerits. These algorithms are: k-means, fuzzy c-means and K-NN clustering algorithm. Choosing an appropriate clustering algorithm for grouping the data takes various factors into account for illustration one is the size of data to be partitioned.

1. Introduction

Many companies have recognized the strategic importance of the knowledge hidden in their database and, therefore, have built data warehouse. A data warehouse is a collection of data from multiple source, integrated into a common repository and extended by summary information for a purpose of analysis. There are two categories of data warehousing. Derived Information is present for the purpose of analysis. The environment is dynamic. In such an environment, either manual analysis supported by appropriated visualization tools or (semi) automatic data mining may be performed. Data mining has been defined as the application of data analysis and discovery algorithm that – under acceptable computational efficiency limitations-produce a particular enumeration of patterns over the data. Several data mining tasks have been identified, e.g., clustering, classification and summarization. Typical results of data mining are as follows: Clustering of item which is typically bought together by some set of customers (clustering in a data warehouse storing sales transaction). Symptoms distinguishing disease A form B (classification in a medical data warehouse). Description of the typical WWW access patterns (summarization in the data warehouse of an internet provider).

2. Clustering Algorithm In Data Mining

Cluster analysis or clustering is the task of grouping a set of objects in such a way that object in the same group(called as cluster) are more similar to each other than to those in other group (cluster). It is main task of exploratory data mining, and

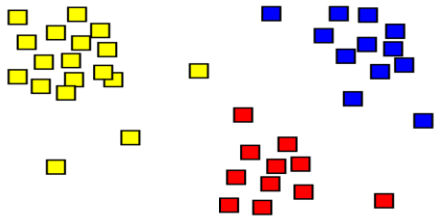


Figure 1 the coloring of the squares into three cluster.

a common technique for statistical data analysis, used in many field, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. The result of a cluster analysis shown as the coloring of a square into three cluster. A clustering is essentially a set of such cluster, usually containing all objects in the data set. Additionally, it may specify the relationship of the cluster to each other, for example, a hierarchy of Cluster embedded in each other. Clustering's can be roughly distinguished as: Hard clustering: Each object belongs to a cluster or not. Soft clustering (fuzzy clustering): Each object belongs to each cluster to a certain degree.

3. Algorithms

Clustering algorithms can be categorized based on their cluster model, as list above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms.

Connectivity based clustering (Hierarchical clustering): Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of object being more related to nearby object further away. These algorithms connect "object" to form "cluster" based on their distance. In a dendrogram the y-axis marks the distance at which the cluster merge, while the object are placed along the x-axis such that the clusters don't mix.

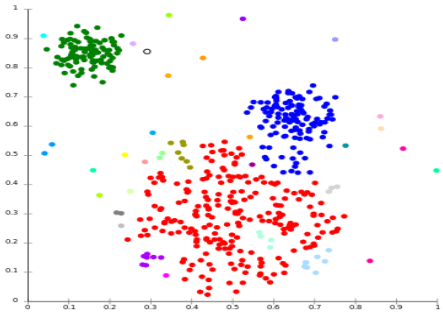


Figure 2 Linkage Clustering

Single-linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into small parts, while before it was still connected to the second largest due to the single-link effect.

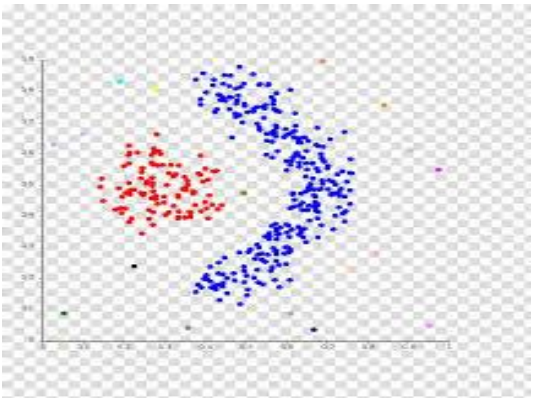


Figure 3 Single –linkage on density-based cluster.

20 clusters extracted, most of which contain single elements, and since linkage clustering does not have a notion of "noise".

Centroid-based clustering: Centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of a data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the cluster center and assign the objects to the nearest cluster, such that the squared distancefrom the cluster are minimized. K-means has a number of interesting theoretical properties. First, it partition the data spaceinto a Structure known as a voronai diagram. Second its conceptually close to a nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as variation of model based classification, and Lloyd's algorithm as a variation of the expectation-maximization algorithm for this model discussed below.

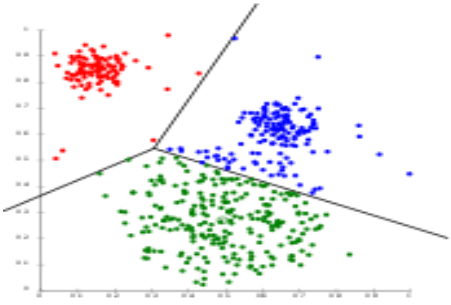


Figure 4 K-means clustering

K-means separate data into Voronoi-cell, which assumes equal-sized cluster(not adequate here)

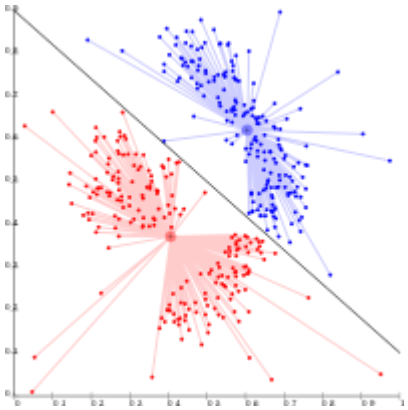


Figure 5 K-means cannot represent density-based clusters

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriority. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step (3).

Merits

1) Fast, robust and easier to understand.
2) Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.
3) Gives best result when data set are distinct or well separated from each other.

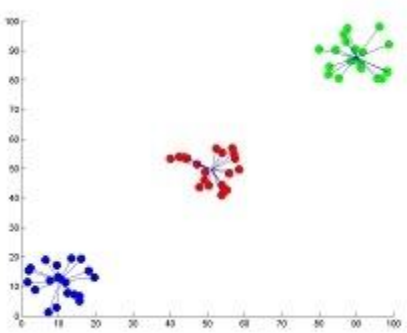


Figure 6 Showing the result of k-means for *'N'* = 60 and *'c'* = 3

Note: For more detailed figure for k-means algorithm please refer to k-means figure sub page.

Demerits

1) The learning algorithm requires apriori specification of the number of cluster centers.

2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.

3) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).

4) Euclidean distance measures can unequally weight underlying factors.

5) The learning algorithm provides the local optima of the squared error function.

6) Randomly choosing of the cluster center cannot lead us to the fruitful result. Pl. refers Fig.

7) Applicable only when mean is defined i.e. fails for categorical data.

8) Unable to handle noisy data and outliers.

9) Algorithm fails for non-linear data set.

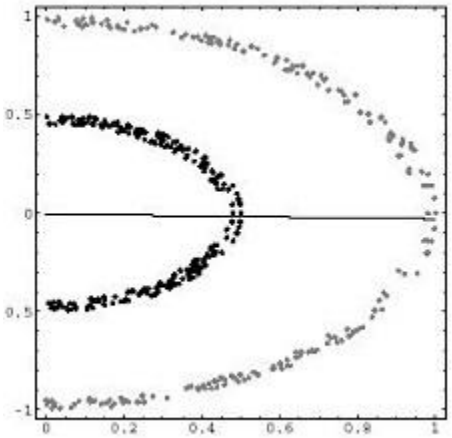


Figure 7 Showing the non-linear data set where k-means algorithm fails

4. Clustering Algorithm In Data Warehosing

Data Warehous The business analyst get the information from the data warehouse to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages: Since a data warehouse can gather information quickly and efficiently, it can enhance business productivity. A data warehouse provides us a consistent view of customers and items; hence, it helps us manage customer relationship. A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period in a consistent and reliable manner.

Three-Tier data Warehouse Architecture Generally a data warehouse adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture. Bottom Tier- The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the extract, Clean, Load, and refresh functions. Middle Tier – In the middle tier, we have the OLAP Server that can be implemented either of the following ways. By Relational OLAP (ROLAP), which is an extended relational database management system? The ROLAP maps the operations on multidimensional data to standard

relational operations. By Multidimensional OLAP (MOLAP) model, this directly implements the multidimensional data and operations. Top-Tier – This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

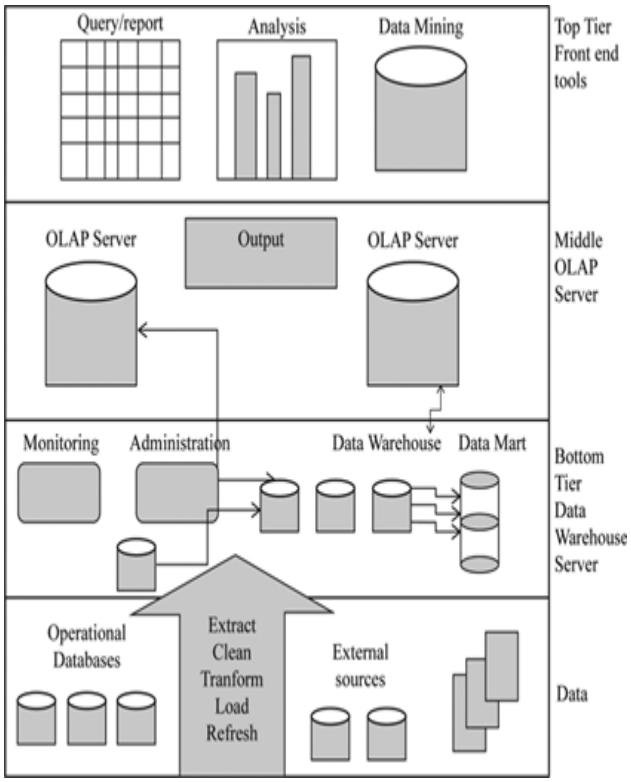


Figure 8 Three tier Architecture of data warehouse

Fuzzy C-Means Clustering

The AlgorithmFuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. The algorithm is composed of the following steps:

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

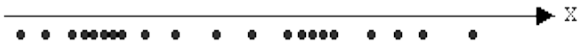
$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $|| U^{(k+1)} - U^{(k)}|| < \varepsilon$ *then STOP; otherwise return to step 2.*

Remarks: As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behavior of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centers of clusters. For a better understanding, we may consider this simple mono-dimensional example. Given a certain data set, supposed to represent it as distributed on an axis. The figure below shows this:



Looking at the picture, we may identify two clusters in proximity of the two data concentrations. We will refer to them using 'A' and 'B'. In the first approach shown in this tutorial - the k-means algorithm - we associated each datum to a specific centroid; therefore, this membership function looked like this:

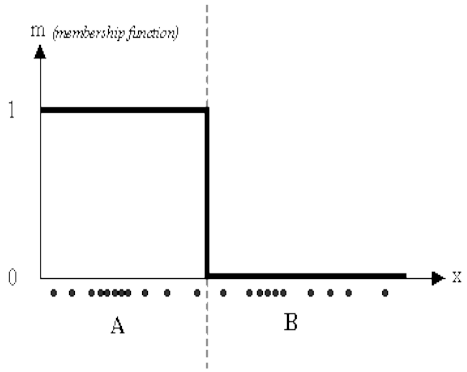


Figure 9

In the FCM approach, instead, the same given datum does not belong exclusively to a well-defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.

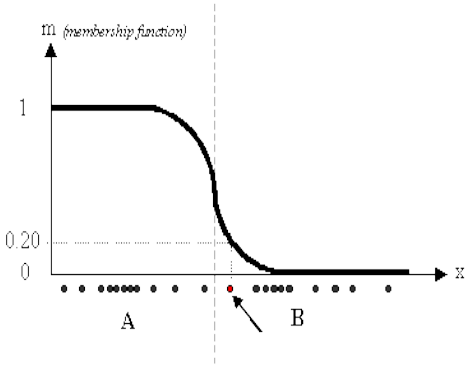


Figure 10

In the figure above, the datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of 'm' indicates the degree of membership to A for such datum. Now, instead of using a graphical representation, we introduce a matrix U whose factors are the ones taken from the membership functions:

$$U_{MC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix} \quad U_{MC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

(a) (b)

The number of rows and columns depends on how many data and clusters we are considering. More exactly we have C = 2 columns (C = 2 clusters) and N rows, where C is the total number of clusters and N is the total number of data. The generic element is so indicated: u_{ij}. In the examples above we have considered the k-means (a) and FCM (b) cases. We can notice that in the first case (a) the coefficients are always unitary. It is so to indicate the fact that each datum can belong only to one cluster. Other properties are shown below:

- $u_{ij} \in [0,1] \qquad \forall i, j$
- $\sum_{j=1}^{C} u_{ik} = 1 \qquad \forall i$
- $0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$

An example Here, we consider the simple case of a mono-dimensional application of the FCM. Twenty data and three clusters are used to initialize the algorithm and to compute the U matrix. Figures below (taken from our interactive demo) show the membership value for each datum and for each cluster. The color of the data is that of the nearest cluster according to the membership function.

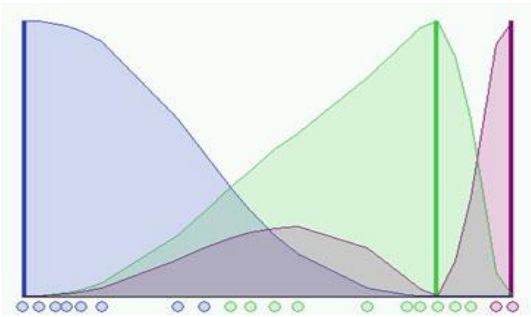


Figure 11

In the simulation shown in the figure above we have used a fuzzyness coefficient m = 2 and we have also imposed to terminate the algorithm when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < 0.3$. The picture shows the initial condition where the fuzzy distribution depends on the particular position of the clusters. No step is performed yet so that clusters are not identified very well. Now we can run the algorithm until the stop condition is verified. The figure below shows the final condition reached at the 8th step with m=2 and ε=0.3:

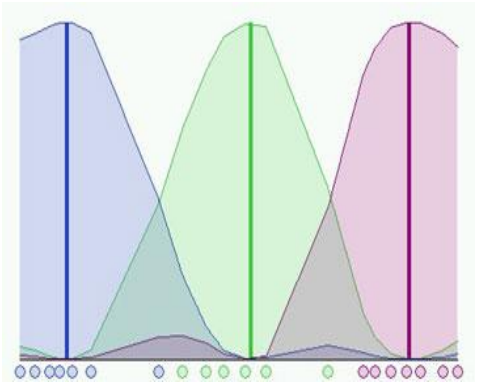


Figure 12

Is it possible to do better? Certainly, we could use an higher accuracy but we would have also to pay for a bigger computational effort. In the next figure we can see a better result having used the same initial conditions and ε=0.01, but we needed 37 steps!

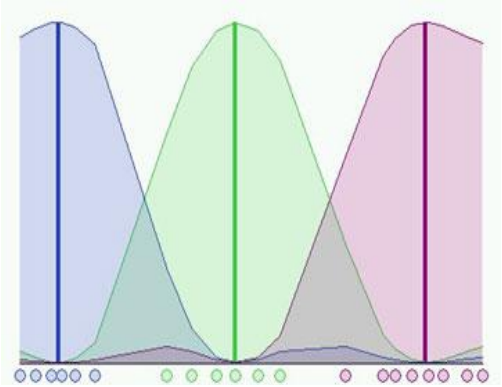


Figure 13

It is also important to notice that different initializations cause different evolutions of the algorithm. In fact it could converge to the same result but probably with a different number of iteration steps.

5. Conclusion

In this paper the authors have reviewed the three major clustering algorithms, namely, the k-means, the fuzzy c-means and the KNN clustering algorithm and have made a comparative study taking account their merits and demerits and all other factors which may influence the criterion in choosing an appropriate clustering algorithm for partitioning a given dataset. In k-means algorithm, the main principle is to minimize the sum of squares of distances between data and corresponding clustering centroids. Hence, the distance measure is the quintessential criterion in classification and formation of clusters. The KNN algorithm uses neighborhood classification as the predication value of the new query instant. The fuzzy c means algorithm, on the other hand, is a soft clustering algorithm. It uses fuzzy sets to cluster data, such that each data point may belong to two or more clusters (overlapping clustering) with different degrees of membership. The data objects on the boundaries between several cluster or not forced to fully belong to one of the clusters, but are rather assigned membership degrees between 0 and 1 indicating their partial membership. Based on the various constraints and conventions, it was observed that the fuzzy c-means algorithm is the most widely used algorithm of all the three. Future Enhancement: Advanced Clustering Algorithm and analyses the shortcoming of the standard k-means, SOM and HAC clustering algorithm. Because the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data point a number of times during every iteration, which makes the effiency of standard clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in O (nk) time without sacrificing the accuracy of clusters. Experimental results show the Advanced Clustering Algorithm can improve the execution time, quality of SOM algorithm and works well on high dimensional data set. So the proposed method is feasible.

Reference

[1]. An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount,Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.

[2]. Research issues on K-means Algorithm: An Experimental Trial Using Matlab by Joaquin Perez Ortega, Ma. DelRocio Boone Rojas and Maria J. Somodevilla Garcia.

[3]. The k-means algorithm - Notes by Tan, Steinbach, Kumar Ghosh.

[4]. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[5]. k-means clustering by ke chen.

[6]. J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57

[7]. J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York.

[8]. Tariq Rashid: "Clustering" http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html

[9]. Hans-Joachim Mucha and Hizir Sofyan: "Nonhierarchical Clustering" http://www.quantlet.com/mdstat/scripts/xag/html/xaghtmlframe149.html

[10]. Hans-Joachim Mucha and Hizir Sofyan: "Nonhierarchical Clustering".

[11]. Rob shot, Rod Gamache, John Vert and Milk Mass 'Window NT Cluster for Availability and Scalability' Microsoft Online Research Paper ,Micro soft Corporation.

[12]. Jim Gray 'QqJim Gray's NT Cluster Research Agenda 'Microsoft Online Research Papers, Micro soft Corporation.