Survey on Comparison of Cluster Algorithm in Data Mining

R. Jayasri[1], Dr. E. George Dharma Prakash Raj[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] Bharathidasan University

Abstract- *A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters. The most distinct characteristic of data mining is that it deals with very large data sets (gigabytes or even terabytes). This requires the algorithms used in data mining to be scalable. However, most algorithms currently used in data mining do not scale well when applied to very large data sets because they were initially developed for other applications than data mining which involve small data sets. We present a fast clustering algorithm used to cluster categorical data. The main advantages of this method is that, can easily categorize the objects and the dissimilar objects can be removed easily. This paper is a survey on Clustering Algorithm for best group of data selection.*

Keywords- Data mining, clustering methods, Fast Clustering Algorithm, k- mean Algorithm.

I. INTRODUCTION

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. We could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

II. CLASSIFICATION OF CLUSTERING ALGORITHMS IN DATA MINING

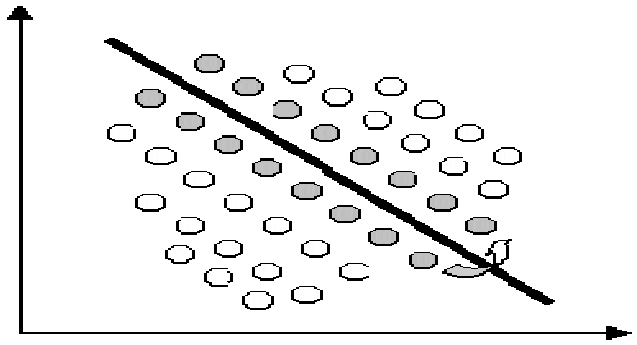
Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

(i) Each of these algorithms belongs to one of the clustering types listed above.
(ii) K-means is an exclusive clustering algorithm
(iii) Fuzzy C-means is an overlapping clustering algorithm
(iv) Hierarchical clustering is obvious and lastly.
(v) Mixture of Gaussian is a probabilistic clustering algorithm.

* In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure below, where the separation of points is achieved by a straight line on a bi-dimensional plane.

*On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.



*Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.

*Finally, the last kind of clustering uses a completely probabilistic approach.

We propose four of the most used clustering algorithms are,

- K-means clustering Algorithm
- Fuzzy C-means clustering Algorithm
- Hierarchical clustering Algorithm
- Mixture of Gaussians clustering Algorithm

Another most using Clustering Algorithm is,

- Density based clustering algorithm
- kernel k-means clustering algorithm
- Quality Threshold (QT) clustering algorithm
- MST based clustering algorithm
- Fast Clustering and Subset Selection Algorithm

k- means clustering algorithm:

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 apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because different location causes different result.

Fuzzy C-means clustering Algorithm:

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one.

Hierarchical clustering Algorithm:

A hierarchical clustering algorithm is based on the union between the two nearest clusters.The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.

Hierarchical clustering algorithm is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and
ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Mixture of Gaussians clustering Algorithm:

This algorithm assumes apriori that there are 'n' Gaussian and then algorithm try to fits the data into the 'n' Gaussian by expecting the classes of all data point and then maximizing the maximum likelihood of Gaussian centers.

Density based clustering algorithm:

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity.

Kernel k-means clustering algorithm:

This algorithm applies the same trick as k-means but with one difference that here in the calculation of distance, kernel method is used instead of the Euclidean distance.

Quality Threshold (QT) clustering algorithm:

This algorithm requires the apriori specification of the threshold distance within the cluster and the minimum number of elements in each cluster. Now from each data point we find all its candidate data points. Candidate data points are those which are within the range of the threshold distance from the given data point. This way we find the candidate data points for all data point and choose the one with large number of candidate data points to form cluster.

MST based clustering algorithm:

First construct MST (minimum spanning tree) using Kruskal algorithm and then set a threshold value and step size. We then remove those edges from the MST, whose lengths are greater than the threshold value. We next calculate the ratio between the intra-cluster distance and inter-cluster distance and record the ratio as well as the threshold. We update the threshold value by incrementing the step size. Every time we obtain the new (updated) threshold value, we repeat the above procedure.

Fast Clustering and Subset Selection Algorithm:

A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining.

III. K- MEANS CLUSTERING ALGORITHM

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 apriori. 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 knows 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.

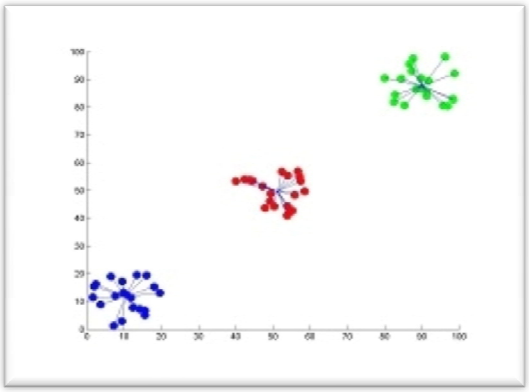


Figure: Showing the result of k-means for 'N' = 60 and 'c' = 3

Advantages

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, t, d << n.
3) Gives best result when data set are distinct or well separated from each other.

Disadvantages

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.
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.

IV. FUZZY C-MEANS CLUSTERING ALGORITHM

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly,

summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

Algorithmic steps for Fuzzy c-means clustering

Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3) Compute the fuzzy centers $'v_j'$ using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...$$

4) Repeat step 2) and 3) until the minimum *'J'* value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.
   where, *'k'* is the iteration step. *'β'* is the termination criterion between [0, 1].
   'U = $(\mu_{ij})_{n*c}$*'* is the fuzzy membership matrix. *'J'* is the objective function.

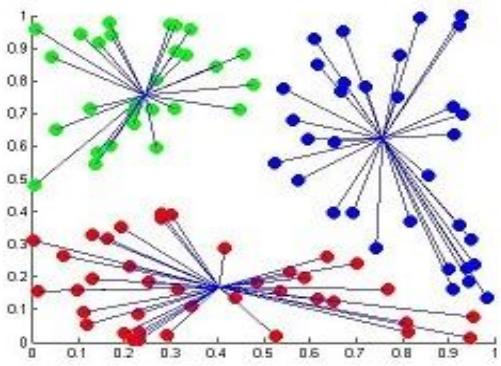


Fig I: Result of Fuzzy c-means clustering

Advantages

1) Gives best result for overlapped data set and comparatively better then k-means algorithm.
2) Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

Disadvantages

1) Apriori specification of the number of clusters.
2) With lower value of β we get the better result but at the expense of more number of iteration.
3) Euclidean distance measures can unequally weight underlying factors.

V. HIERARCHICAL CLUSTERING ALGORITHM

A hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Hierarchical clustering algorithm is of two types there are,

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm is exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.

Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point, there are many available methods. Some of them are:

1) Single-nearest distance or single linkage.
2) Complete-farthest distance or complete linkage.
3) Average-average distance or average linkage.
4) Centroids distance.
5) Ward's method - sum of squared Euclidean distance is minimized.

This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many numbers of clusters should be actually present.

Algorithmic steps for Agglomerative Hierarchical clustering

Let $X = \{x_1, x_2, x_{3,..}... x_n\}$ be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering

m. Set the level of this clustering to L(m) = d[(r),(s)].
4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).
5) If all the data points are in one cluster then stop, else repeat from step 2).

Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.

Advantages

1) No apriori information about the number of clusters required.
2) Easy to implement and gives best result in some cases.

Disadvantages

1) Algorithm can never undo what was done
2) Time complexity of at least O($n^2 log\ n$) is required, where *'n'* is the number of data points.
3) Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:
    i) Sensitivity to noise and outliers.
    ii) Breaking large clusters.
    iii) Difficulty handling different sized clusters and convex shapes
4) No objective function is directly minimized.
5) Sometimes it is difficult to identify the correct number of clusters by the dendogram.

VI. MIXTURE OF GAUSSIANS CLUSTERING ALGORITHM

This algorithm assumes apriori that there are 'n' Gaussian and then algorithm try to fits the data into the 'n'Gaussian by expecting the classes of all data point and then maximizing the maximum likelihood of Gaussian centers.

Algorithmic steps for Expectation Maximization (EM) clustering

Let X={x_1, x_2, x_3, ..., x_n} be the set of data points V = {μ_1, μ_2, μ_3, ..., μ_c} be the set of means of Gaussian P = {p_1, p_2, p_3, …, p_c} be the set of probability of occurrence of each Gaussian

1) On the i^{th} iteration initialize.
2) Compute the "expected" classes of all data points for each class using:
3) Compute "maximum likelihood μ" given our data class membership distribution using:

$$p_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t)}{R}$$

Where, 'R' is the number of data points.

Advantage

1) Gives extremely useful result for the real world data set.

Disadvantage

1) Algorithm is highly complex in nature.

VII. DENSITY BASED CLUSTERING ALGORITHM

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity.

Density Reachability- A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbors who are within distance ε.

Density Connectivity- A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".

Algorithmic steps for DBSCAN clustering

Let X = {x_1, x_2, x_3, ..., x_n} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

1) Start with an arbitrary starting point that has not been visited
2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
3) If there is sufficient neighborhood around this point then clustering process starts and point is marked as visited

else this point is labeled as noise (Later this point can become the part of the cluster).

4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.
5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
6) This process continues until all points are marked as visited.

Advantage

1) Does not require a-priori specification of number of clusters.
2) Able to identify noise data while clustering.
3) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

Disadvantages

1) DBSCAN algorithm fails in case of varying density clusters.
2) Fails in case of neck type of dataset.

VIII. FAST CLUSTERING AND SUBSET SELECTION ALGORITHM

A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The most distinct characteristic of data mining is that it deals with very large data sets (gigabytes or even terabytes). This requires the algorithms used in data mining to be scalable. However, most algorithms currently used in data mining do not scale well when applied to very large data sets because they were initially developed for other applications than data mining which involve small data sets. A fast clustering algorithm used to cluster categorical data. The main advantages of this method is that, can easily categorize the objects and the dissimilar objects can be removed easily.

The Fast Clustering algorithm follows the steps are,

(i) Irrelevant feature removal and redundant feature elimination,
(ii) Then Minimum Spanning Tree construction and the data set specify the Complete Graph (G),
(iii) The features are divided in Tree partition and pair wise find the related features selected from complete Graph (G)..
(iv) Best subset of feature is selected from large amount of data.

Advantages

(i) Categorize the objects and the dissimilar objects can be removed easily.
(ii) Effectively remove the irrelevant features and redundant feature.
(iii) Easy to retrieve the data.
(iv) Easy to classify new instance rapidly.
(v) Performance good better than traditional feature subset selection algorithm.
(vi) It produces the accurate result.

IX. CHALLENGES IN CLUSTERING

Requirements

The main requirements that a clustering algorithm should satisfy are:

- Scalability;
- Dealing with different types of attributes;
- Discovering clusters with arbitrary shape;
- Minimal requirements for domain knowledge to determine input parameters;
- Ability to deal with noise and outliers;
- Insensitivity to order of input records;
- High dimensionality;
- Interpretability and usability.

Possible Applications

Clustering Algorithms can be applied in many fields, for instance:

- **Marketing:** finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- **Biology:** classification of plants and animals given their features;
- **Libraries:** book ordering;
- **Insurance:** identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- **City-planning:** identifying groups of houses according to their house type, value and geographical location;
- **Earthquake studies:** clustering observed earthquake epicenters to identify dangerous zones;

X. COMPARISON OF CLUSTER ALGORITHM

ADVANTAGE AND DISADVANTAGE OF VARIOUS CLUSTER ALGORITHMS

S.no	Cluster Algorithm	ADVANTAGES	DISADVANTAGES
1	K-means clustering Algorithm	(i) Fast, robust and easier to understand. (ii)Relatively efficient. (iii)Gives best result when data set are distinct or well separated from each other.	(i)The learning algorithm requires apriori specification of the number of cluster centers. (ii) Applicable only when mean is defined i.e. fails for categorical data. (iii) Unable to handle noisy data and outliers. (iv) Algorithm fails for non-linear data set.
2	Fuzzy C-means clustering Algorithm	(i) Gives best result for overlapped data set and comparatively better then k-means algorithm. (ii) Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.	(i) Apriori specification of the number of clusters. (ii) With lower value of β we get the better result but at the expense of more number of iteration. (iii) Euclidean distance measures can unequally weight underlying factors.
3	Hierarchical clustering algorithm	(i) No apriori information about the number of clusters required. (ii) Easy to implement and gives best result in some cases.	(i) Algorithm can never undo what was done previously. (ii) Time complexity. (iii)Sensitivity to noise and outliers. (vi) Breaking large clusters.

			(v) Difficulty handling different sized clusters and convex shapes (vi) No objective function is directly minimized.
4	Mixture of Gaussians clustering algorithm	(i) Gives extremely useful result for the real world data set.	(i)Algorithm is highly complex in nature.
5	Density based clustering algorithm	(i) Does not require a-priori specification of number of clusters. (ii) Able to identify noise data while clustering. (iii) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.	(i)DBSCAN algorithm fails in case of varying density clusters. (ii) Fails in case of neck type of dataset.
6	Kernel k-means clustering algorithm	(i)Algorithm is able to identify the non-linear structures. (ii)Algorithm is best suited for real life data set is large.	(i) Number of cluster centers need to be predefined. (ii) Algorithm is complex in nature and time complexity.
7	Quality Threshold (QT) clustering algorithm	(i) Quality Guaranteed - Only clusters that pass a user-defined quality threshold will be returned. (ii) Number of clusters is not specified apriori. (iii) All possible clusters are considered - Candidate cluster is generated with respect to every data points and tested in order of size against quality criteria.	(i) Computationally Intensive and Time Consuming - Increasing the minimum cluster size or increasing the number of data points can greatly increase the computational time. (ii) Threshold distance and minimum number of element in the cluster has to be defined apriori.

8	MST based clustering algorithm	(i) Comparatively better performance then k-means algorithm.	(i) Threshold value and step size needs to be defined apriori
9	Fast Clustering and Subset Selection Algorithm	(i) Categorize the objects and the dissimilar objects can be removed easily. (ii) Effectively remove the irrelevant features and redundant feature. (iii) Easy to retrieve the data. (iv) Easy to classify new instance rapidly. (v) Performance good better than traditional feature subset selection algorithm. (vi) It produces the accurate result	

XI. CONCLUSION

In this survey on Clustering algorithms are used in many applications. Clustering has long been used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid,Which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. The most distinct characteristic of data mining is that it deals with very large data sets (gigabytes or even terabytes). This requires the algorithms used in data mining to be scalable. However, most algorithms currently used in data mining do scale well when applied to very large data sets because they were initially developed for other applications than data mining. Present Clustering algorithm used to cluster categorical data. The main advantages of this method is that, can easily categorize the objects and the dissimilar objects can be removed easily.

REFERENCE

[1] k-means and Hierarchical Clustering by Andrew W. Moore.

[2] Hierarchical Document Clustering by Benjamin C. M. Fung, Ke Wang and Martin Ester.

[3] How to explain Hierarchical Clustering by S. P. Borgatti.

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

[5] BIRCH: An efficient data clustering method for very large databases by T. Zhang, R. Ramakrishnan and M. Livny.

[6] Fuzzy c-means by Balaji K and Juby N Zacharias.

[7] Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation by Weiling Cai, Songcan Chen and Daoqiang Zhang.

[8] Clustering with Gaussian Mixtures by Andrew W. Moore.

[9] http://en.wikipedia.org/wiki/Cluster_analysis.

[10] A Comparison of Fuzzy and Non-Fuzzy clustering Techniques in Cancer Diagnosis by X.Y. Wang and J.M. Garibaldi.

[11] Probability Density Estimation from Optimally Condensed Data Samples by Mark Girolami and Chao He.