International Journal of Computer Science
Oddity...Probe...Reviste...

http://www.ijcsjournal.com Volume 1, Issue 2, No 2, 2013. ISSN: 2348-6600
Reference ID: IJCS-012 PAGE NO: 060-063

Clustering Algorithms and Techniques: A Survey

S.Neelamegam #1, Dr.E.Ramaraj *2

#1 Department of Computer science and Engineering,
Alagappa University, Karaikudi.
India.

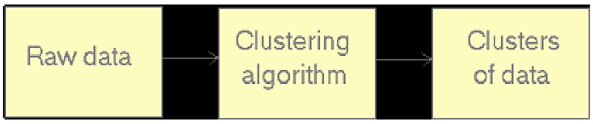
*2 Professor ,Department of Computer science and Engineering,
Alagappa University, Karaikudi.
India.

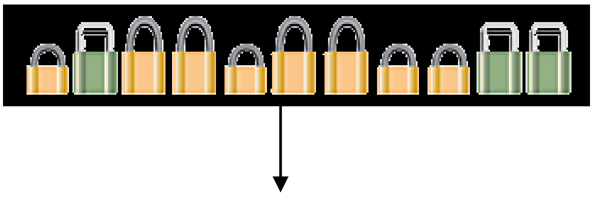
***Abstract*— The goal of this survey provide the best knowledge about the clustering and their algorithms. Nowadays lot of algorithms is available in clustering. This survey very useful for initial level clustering practitioners. Different clustering algorithms are discussed in this paper.**

***Index Terms*— Clustering, Machine learning, Data mining.**

I. INTRODUCTION

Clustering is an unsupervised learning technique. Clustering is the process for grouping the similar object. Each group called cluster. Clustering is one of the most important process (step) in Data mining. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. Finding a 'structure' in a collection of unlabeled data. A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. A good clustering methods produce a high quality clusters. The quality of the clustering method measured by ability to discover some or all of the hidden patterns.





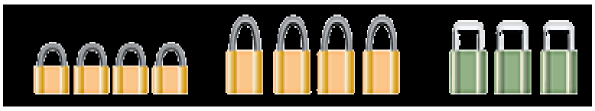


Fig 1 CLUSTERING PROCESS.

Figure 1 shown how to cluster raw data. Here three type of locks are raw data. Clustering algorithm is used to cluster the raw data. Finally three clusters are present. The three type of locks are grouped into the similar lock.

II. TYPE OF CLUSTERING

Generally Clustering methods are four types, Partitioning, Hierarchy, Distance based, Probabilistic.

2.1. PARTITIONING METHOD

The partitioning method is the famous method in clustering the data set. The partitioning method is divide data into proper subset. Partitioning methods reallocate instance moving one cluster to another, starting from an initial partition. Usually starts with random partitioning. Construct a partition of n documents into a set of K clusters. Input data are given set of documents and number of K. To find a partition of k clusters that optimizes the chosen partitioning criterion. Each cluster represented by a centroid or a cluster representative. The portioning method algorithms are k-means, k-medoids, fuzzy c-means. These three techniques are deal with large data sets.

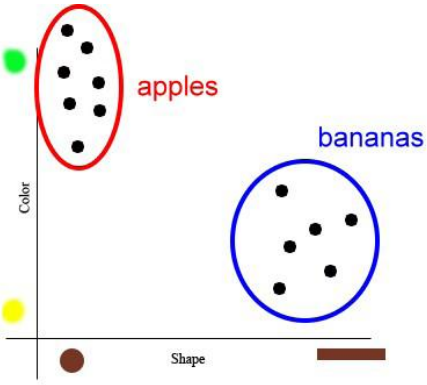


FIG 2. PARTITIONING BASED CLUSTERING

Figure 2 shown sample of partition based clustering, This example contains two clusters one is apple another one is banana

K-Mean Algorithm:

The K-Means method is the simple and easiest algorithm for clustering the real world large data. K-means cluster finds the pure subsets in the large number of clusters. K-means cluster initially create k cluster i.e. no of class in the data. Suppose cluster are not available in data randomly create k clusters.

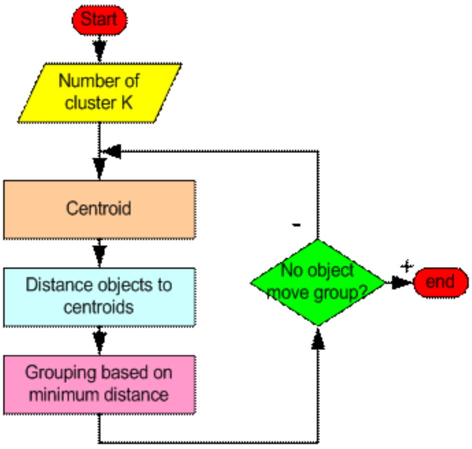

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Fig 3. K-MEANS CLUSTERING PROCESS

To find centroid for each cluster. Then measure the distance between the clusters. Form K clusters by assigning each point to its closest centroid. Assign the centroid points to nearest points in each cluster. Recompute the new unit positions as cluster centers.

```
Algorithm 1. K-Means Algorithm

1: Select k points as initial Centroids.
2: repeat
3:     from k clusters by assigning each point to its closest
       Centroids
4:     recompute the centroid of each cluster
5: until Centroids not change.
```

K- Medoids Algorithm:

The K-Medoids algorithm is similar to k-means algorithm. Minimize the sensitivity of k-means to outliers. The actual object is representing by cluster mean value. Each reaming objects are represented with object medoids. First chosen representative randomly. The basic strategy of K-medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoids) for each cluster.

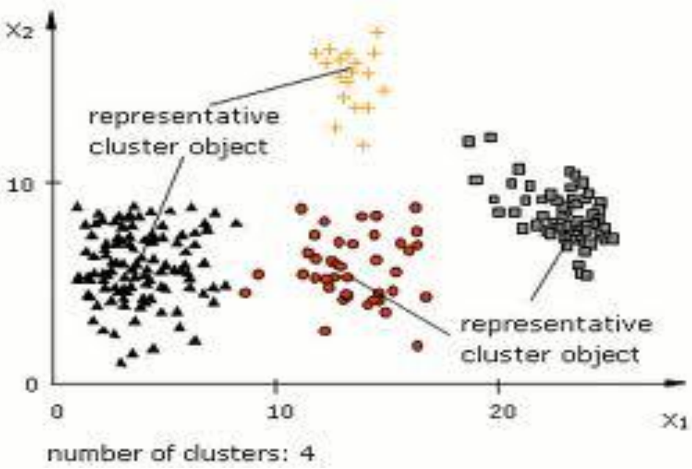


Fig 4. K- MEDOIDS REPRESENTATIVE CLUSTERING

K-Medoids algorithm randomly creates the representative object. The representative object is called medoids.

```
Algorithm 2. K- Medoids Algorithm
```

```
1: Arbitrarily choose k objects in D as the initial
      representative objects.
2: Assign each remaining object to the cluster with the nearest
      medoids.
3: Randomly select a non medoids object $O_{random}$.
4. Compute the total points S of swapping object $O_i$ with
      $O_{random}$ .
5: If S<0 then swap $O_i$ with $O_{random}$ to form the new set of k
      medoids.
6: Until no change.
```

Fuzzy c-means Algorithm:

Fuzzy c-means clustering (FCM) works alike fuzzy logic. FCM method allows cluster singe piece of data into two or more cluster. FCM mostly used in pattern reorganization. Fuzzy c-means is closely related to Soft k-means algorithm. FCM is actually used improve the clustering accuracy under the noise

```
Algorithm3. Fuzzy C-means:

1: Initialize U= [$u_{ij}$] matrix, U(0).
2: At k-step: Calculate the centers vector, C(K)=[Cj] with(k)
3: Update U(K), U(K+1).
4: If || U(K+1) – U(K)|| threshold then STOP; otherwise return
      step 2.
5: The computation of the updated membership function is the
condition for minimization of the objective function
```

Fuzzy C-means clustering is better than the hard c-means clustering. Fuzzy c-means clustering also known as fuzzy ISODATA. FCM performs the fuzzy partition in the data.

2.2 HIERARCHICAL METHOD

Hierarchical clustering method is second important clustering in Data mining. Hierarchical clustering is a method of cluster analysis which Seeks to build a hierarchy of clusters. Hierarchical clustering algorithms are grouped object in Hierarchical (tree) structure. This method of algorithms divided into two types hierarchical clustering and divisive hierarchical clustering. Given a set of N items to be clustered, and an NxN distance (or similarity) matrix.

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Steps in process:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3. Compute distances (similarities) between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

5. can be done in different ways, which is what distinguishes *single-link* from *complete-link* and *average-link* clustering. In *single-link* clustering

Nearest-Neighbour, Farthest-Neighbour, Minimum spanning tree algorithms are works under the hierarchical clustering techniques. The basics of hierarchical clustering include Lance-Williams formula, idea of conceptual clustering, now classic algorithms SLINK, COBWEB, as well as newer algorithms CURE and CHAMELEON.

Hierarchical clustering is often displayed graphically using a tree –like diagram called dendrogram, which displays both the cluster-sub cluster relationships and the order in which the clusters were merged or split.

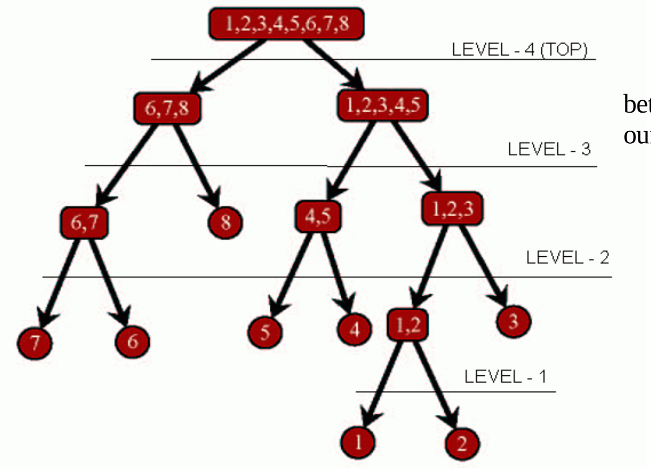


Fig 5: STRUCTURE OF HIERARCHICAL CLUSTERING.

Figure 5. Explains how to clustered the data in level by level. The Hierarchical clustering is clustered by N*N distance matrix.

2.3 DISTANCE BASED METHOD

Distance generally, the distance between two points is taken as a common metric to assess the similarity among the components of a population. The commonly used distance measure is the Euclidean metric which defines the distance between two points p= (p1, p2, ...) and q = (q1, q2, ...) is given by :

$$d = \sqrt{\sum_{i=1}^{k} (p_i - q_j)^2}$$

It's easy to handle. Distance Based method is assign the distance measure between each data. Find the distance between same cluster objects and minimize the distance. The different cluster have maximized distance in the clusters.

Issues :

1. Requires defining a distance (similarity) measure in situation where it is unclear how to assign it
2. What relative weighting to give to one attribute vs another?
3. Number of possible partition us super exponential

2.4 PROBABILISTIC

There are several ways to model the relationship between distances and probabilities. The simplest model, and our working principle (or axiom), is the following:

1. Fo r each x∈D, and each cluster C k,
2. pk(x) dk(x)=constant, depending on x.
3. Cluster membership is thus more probable the closer the data point is to the cluster center. Note that the constant in is independent of the cluster k.

III. Conclusion

In this paper describe all available methods in clustering techniques. This paper is very useful for clustering practitioners and it provides the best knowledge about the clustering technologies.

IV. References

1. Data Clustering and Its Applications ://members.tripod.com/asim_saeed/paper.htm
2. A Survey of Clustering Techniques *International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010*
3. Introduction to partitioning based clustering methods with a robust example ISSN 14564378
4. http://www.scialert.net/fulltext/?doi=itj.2011.478.484 &org=11
5. A survey of cluster analysis International journal of Computer science.Vol-7 No-12, October 2010.
6. Probabilistic D-Clustering Journal of Classification 25 (2007) DOI: 10.1007/s00357-007-0021-y.
7. Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software, Inc
8. A Survey: Clustering Ensembles Techniques World Academy of Science, Engineering and Technology 26 2009
9. A Possibilistic Fuzzy c-Means Clustering Algorithm IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 13, NO. 4, AUGUST 2005.