Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have highly similarity, but are very dissimilar to objects in other clusters. Dissimilarties are similarities are accessed based on the attribute values describing the objects and often involve distance measures. Clustering as a data mining tool has its roots in many application areas such as biology, security, business intelligence, and web search.

You will learn several basic clustering techniques organizing into the following categories:

1. Partitioning Methods
2. Hierarchical Methods
3. Density based Methods
4. Grid based Methods
5. Model based clustering Methods
6. Clustering of high Dimensional data

Cluster Analysis:— cluster analysis or simply clustering is the Process of Partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to object in other clusters. The set of clusters resulting from a cluster analysis can be referred to as clustering. In this context, different clustering Methods may generate different clustering on the same data set. The Partitioning is not performed by humans but by the clustering Algorithm.

Hence, clustering is useful in that can lead to the discovery of previously unknown groups within the data.

Requirements for cluster Analysis :→ The following are the typical requirements of clustering in data mining

* Scalability
* Ability to deal with different types of attribute
* Requirements for domain knowledge to determine I/P parameter
* Ability to deal with noisy data
* Incremental clustering and Insensitivity to I/P order
* Capability of clustering high dimensional data
* Constraint based clustering
* Interpretability and usability

The following are orthogonal Aspects with which clustering Methods can be compared:

* The Partitioning criteria
* separation of clusters
* Similarity Measure
* Clustering space :-

Overview of Basic clustering Methods :-

Method	General characteristics
(1) Partitioning Methods	- find mutually exclusive clusters of spherical shape - Distance based - May use mean or Medoid to represent - Effective for small to medium data set

Hierarchical Method	- clustering is a hierarchical decomposition (i.e multiple levels) - cannot correct erroneous merges or splits - May incorporate other technique like Microclustering or consider object "linkages"
Density based Method	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low density regions. - cluster density, each point must have a minimum number of points within its neighbourhood. - May filter out outliers
Grid based Method	- use a multiresolution grid based structure - fast Processing time (typically independent of the number of data objects, yet dependent on grid size)

Partitioning Methods:- Given a set of n objects, a Partitioning Method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$, that is, it divides the data into k groups such that each group must contain at least one object.

The cluster are formed to optimize one objective partitioning criterion such as a dissimilarity function based on distance, so that the objects within a cluster are "similar" to one another and "dissimilar" to object in other clusters in terms of data set attributes

The most well known and commonly used Partitioning Methods are (i) K-Means (ii) K-Medoids

K-Means: A Centroid based technique :- Suppose a dataset D, contains n objects in D into K- clusters, $C_1, C_2 \dots C_k$ that is $C_i \subset D$ and $C_i \cap C_j = \phi$ for $(i \leq k, j \leq k)$. An objective function is used to assess the Partitioning quality so that objects with in a cluster are similar to one another but dissimilar to objects in other clusters. This is, the objective function aims for high intracluster similarity and low Interchuster similarity.

A Centroid based Partitioning Method uses the Centroid of a Cluster, C_i to represent that cluster. Conceptually, the Centroid of a cluster is a central Point. The Centroid can be defined in various ways such as by the Mean or the Medoid of the objects (or Points) assigned to the cluster.

How does the K-Means algorithm work :- The K-Means algorithm defines the centroid of a cluster as the Mean value of the Points with the cluster. It Proceeds as follows. First, it randomly selects K of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance bw the object and the cluster Mean. The K- Means algorithm then iteratively improves the within- Cluster variation. For each cluster, it computes the new Mean using the objects assigned to the cluster in the Previous iteration. All the objects are then reassigned using the updated Means as the new cluster centers. The Iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the Previous Round.

10.2

figure: The K-Means Partitioning algorithm :-

Algorithm: K-Means. The K Means algorithm for Partitioning, where each cluster's center is represented by the Mean value of the objects in the cluster.

Input:

* K: The number of clusters,
* D: a data set containing n objects

Output: A set of k clusters

Method:

(1) Arbitrarily choose K objects from D as the initial cluster centers;
(2) repeat
(3) (Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster Means, that is, calculate the Mean value of the objects for each cluster;
(5) until no change;

Example:— Consider a set of objects located in 2-D space, as depicted in figure 10.3 (a). Let K=3, i.e, the User wanted like the objects to be partitioned into three clusters.

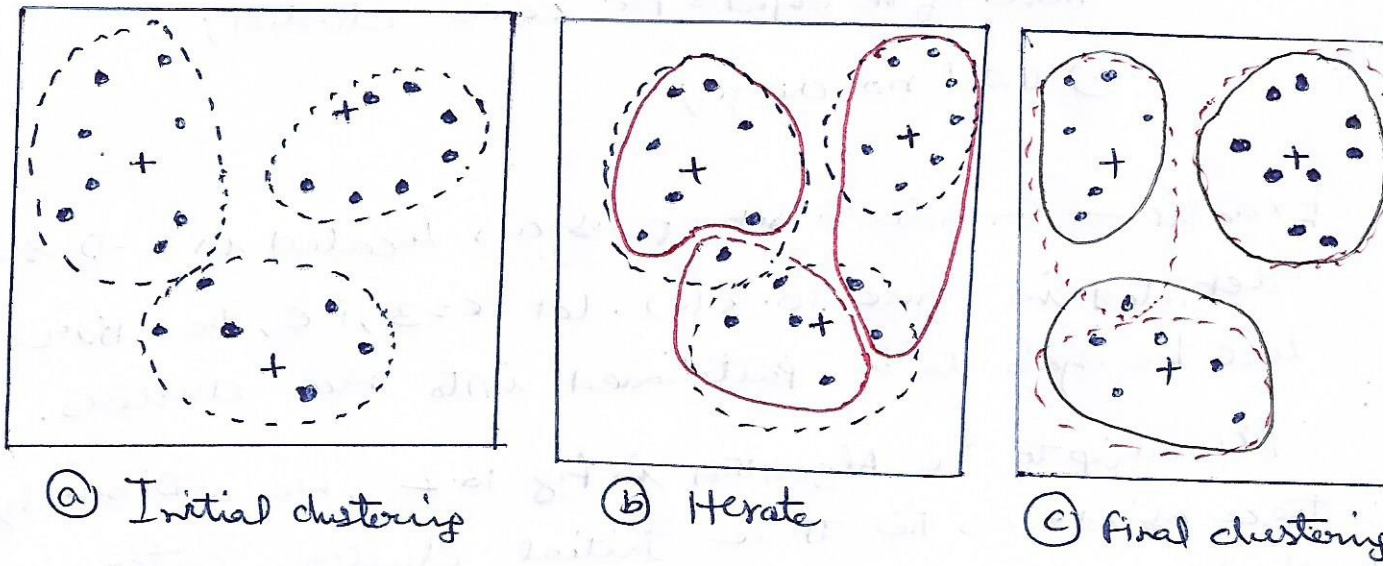
According to the Algorithm in fig 10.2, we arbitrarily choose three objects as the three Initial cluster centers, where cluster centers are marked by a +. Each object is assigned to a cluster based on the cluster center to which it is nearest. Such a distribution forms silhouettes en

dotted curves, as shown in figure 10.3 (a).

Next the cluster centers are updated. i.e, The Mean Value of each cluster is recalculated based on the current objects in the cluster. Using the new cluster centers, the objects are re-distributed to the clusters based on which cluster center is nearest. Such a redistribution forms new Silhouettes encircled by dashed curves, as shown in figure 10.2 (b)

This process iterates, leading to figure 10.3(c). The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as iterative relocation. Eventually, no reassignment of the objects in any cluster occurs and so the process terminates.

The resulting clusters are returned by the clustering process. The time complexity of the K-means algo is $O(nkt)$, where n is the total number of objects, K is the number of clusters, and t is the number of iterations. Normally $K \leq n$ and $t \leq n$.



(a) Initial clustering (b) Iterate (c) Final clustering

fig:- Clustering of a set of objects using the K-means Method for Update cluster Centers and reassign objects accordingly.

K-Mean clustering Algorithm : →

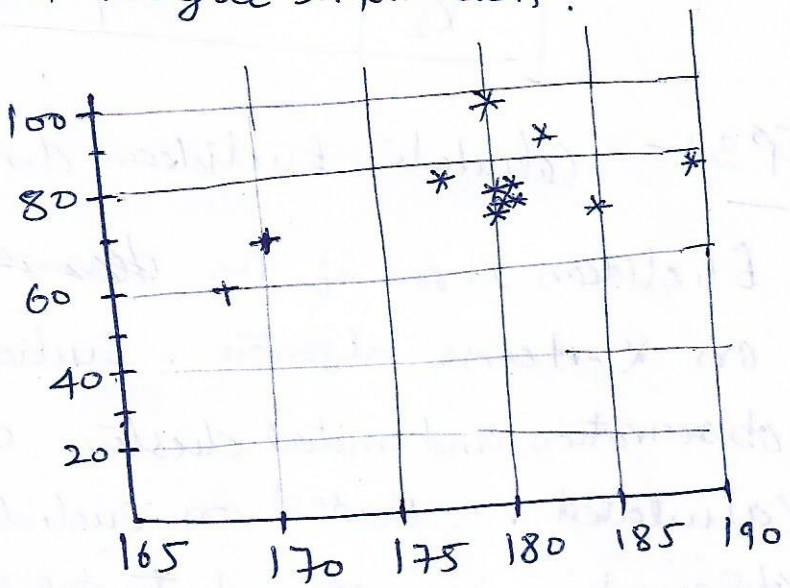
one of the most frequently used unsupervised algorithms is K-Means. K-means clustering is exploratory data analysis technique - This is non-hierarchical method of grouping objects together.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

For Ex:- we have height and weight information, using these two variables, we need to group the objects based on height and weight information.

Data Sample

Height	weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76



if you look at the above diagram, you are able to see two visible clusters/segments and we want those to be identified using K-Means algorithm.

Step 1:- Input Dataset, Clustering variables and maximum No. of clusters (K in Means clustering) in this dataset, only Two Variable - height and weight - are considered for Clustering. K=2

Step 2:- initialize cluster centroid; in this example, value of K is considered as 2. Cluster Centroids are initized with first 2 observations

Cluster	Initial Centroid	
	Height	weight
K_1	185	72
K_2	170	56

step 3:- Calculate Euclidean distance

Euclidean is one of the ~~Means~~ distance Measures used on K-Means algorithm. Euclidean distance b/w of a observation and initial cluster centroid 1 and 2 is Calculated. Based on euclidean distance each observation is assigned to one of the clusters - based on Minimum distance

$$\text{Euclidean distance} = \sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$$

where :— X_H: observation value of variable height

H_1: Centroid value of cluster 1 for variable height

X_W: observation value of variable weight

W_1: Centroid value of cluster 1 for variable weight

first Two observations

Height	weight
185	72
170	56

~~Step 3:~~ ~~calculate~~

Now Initial cluster centroid are:-

Cluster	Updated centroid	
	Height	weight
K_1	185	72
K_2	170	56

Euclidean distance calculation from each of the clusters is calculated

Euclidian distance from Cluster 1	Euclidean distance from Cluster 2	Assignment
$(185-185)^2+(72-72)^2$ $=0$	$(185-170)^2+(72-56)^2$ $=21.93$	1
$(170-185)^2+(56-72)^2$ $=21.93$	$(170-170)^2+(56-56)^2$ $=0$	2

We have considered two observations for assignment only because we knew the assignment, and there is no change in centroids as these two observations were only considered as initial centroids.

Step 4:- Move on to next observation and calculate Euclidean Distance.

Height	weight
168	60

Euclidean distance from cluster 1	Euclidean distance from cluster 2	Assignment
$(168-185)^2+(60-72)^2$ $\Rightarrow 20.808$	$(168-185)^2+(56-60)^2$ $\Rightarrow 4.472$	2

Since distance is Minimum from cluster 2, so the observation is assigned to cluster 2. Now revise Cluster Centroid - Mean value Height and weight as cluster centroids. Addition is only to cluster 2, so Centroid of cluster 2 will be updated

Updated cluster centroids

	Updated centroid	
Cluster	Height	weight
K=1	185	72
K=2	$(170+168)/2$ $\Rightarrow 169$	$(56+60)/2$ $\Rightarrow 58$

Step 5:- Calculate Euclidean distance for the next observation, assign next observation based on minimum Euclidean distance and update the cluster centroids

Next observation

Height	weight
179	68

Euclidean Distance Calculation and assignment

Euclidean distance from cluster 1	Euclidean distance from cluster 2	Assignment
~~7.211~~	14.142	1

Update cluster centroid

Cluster	update centroid	
	Height	weight
K=1	182	70.667
K=2	169	59

Continue the ~~steps~~ until all observations are assigned

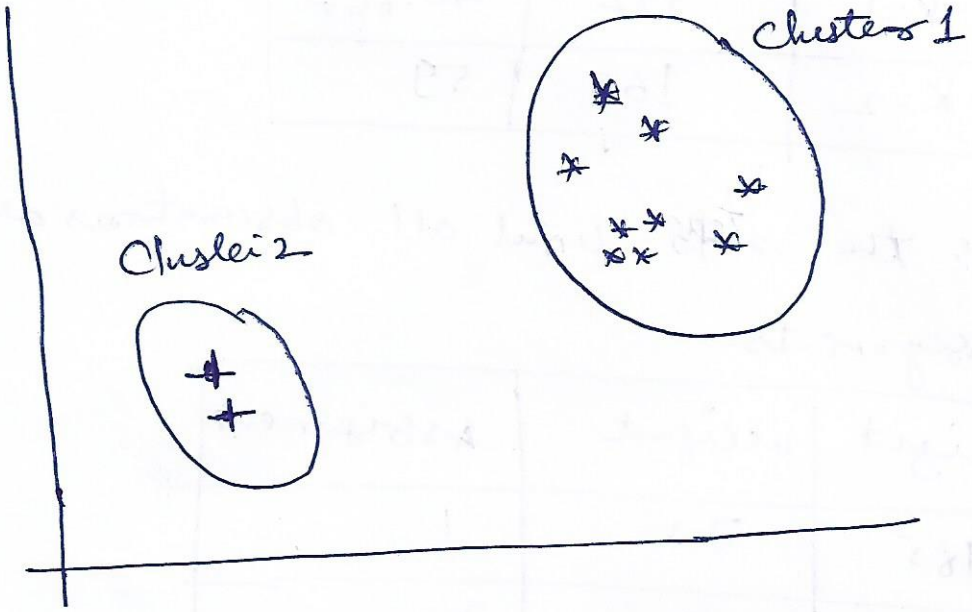
final assignments

Height	weight	Assignment
185	72	1
170	56	2
168	60	2
179	68	1
182	72	1
188	77	1
180	71	1
180	70	1
183	84	1
180	88	1
180	67	1
170	76	1

Final cluster centroids

Cluster	Updated centroid	
	Height	Weight
K=1	182.8	72
K=2	169	58

that is what was expected based on Two dimensional plot.



Important Considerations in K-Means:-

(1) Scale of Measurements influences Euclidean Distance, so variable Standardisation becomes necessary

(2) Depending on expectations - you may requires Outlier treatement

(3) K-Means Clustering may be biased on initial centroids - called Cluster seeds.

(4) Maximum clusters is typically inputs and may also impact the clusters getting created.