INTRODUCTION TO DATA ANALYTICS

Class # 26

Clustering techniques

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology
IIIT Sri City

TOPICS TO BE COVERED…

- Introduction to clustering
- Similarity and dissimilarity measures
- Clustering techniques
- Partitioning algorithms
- Hierarchical algorithms
- Density-based algorithm

CLUSTERING TECHNIQUES

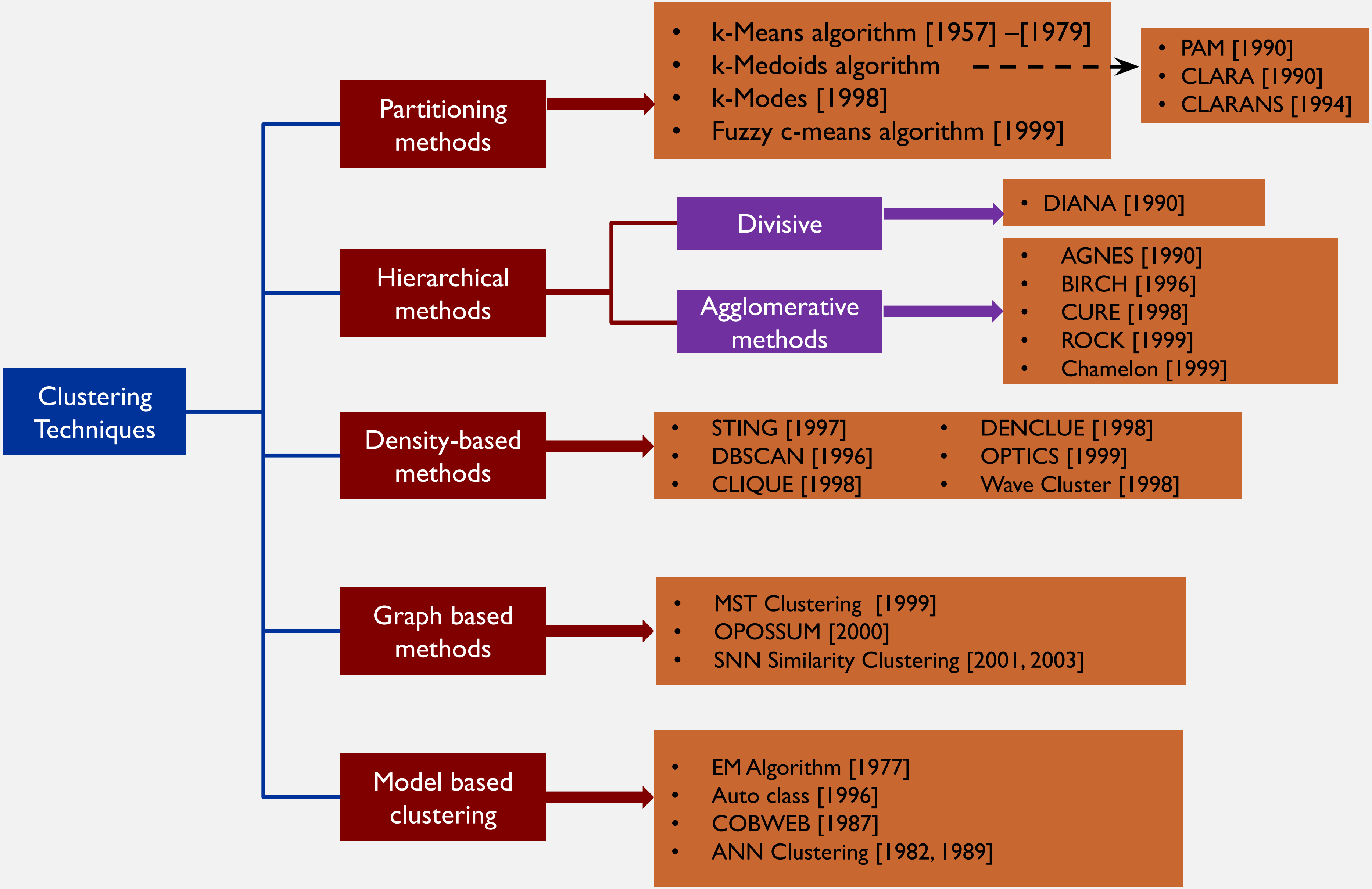
- Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications.

- As a result, many clustering techniques have been reported in the literature.

- Let us categorize the clustering methods. In fact, it is difficult to provide a crisp categorization because many techniques overlap to each other in terms of clustering paradigms or features.

- A broad taxonomy of existing clustering methods is shown in the next slide.

- It is not possible to cover all the techniques in this lecture series. We emphasize on major techniques belong to partitioning and hierarchical algorithms.



Clustering Techniques

Partitioning methods

- k-Means algorithm [1957] –[1979]
- k-Medoids algorithm
- k-Modes [1998]
- Fuzzy c-means algorithm [1999]

- PAM [1990]
- CLARA [1990]
- CLARANS [1994]

Hierarchical methods

Divisive

- DIANA [1990]

Agglomerative methods

- AGNES [1990]
- BIRCH [1996]
- CURE [1998]
- ROCK [1999]
- Chamelon [1999]

Density-based methods

- STING [1997]
- DBSCAN [1996]
- CLIQUE [1998]
- DENCLUE [1998]
- OPTICS [1999]
- Wave Cluster [1998]

Graph based methods

- MST Clustering [1999]
- OPOSSUM [2000]
- SNN Similarity Clustering [2001, 2003]

Model based clustering

- EM Algorithm [1977]
- Auto class [1996]
- COBWEB [1987]
- ANN Clustering [1982, 1989]

CLUSTERING TECHNIQUES

- In this lecture, we shall cover the following clustering techniques only.
  - Partitioning
    - k-Means algorithm
    - PAM (k-Medoids algorithm)
  - Hierarchical
    - DIANA (divisive algorithm)
    - AGNES
    - ROCK

    } (Agglomerative algorithm)
  - Density – Based
    - DBSCAN

K-MEANS ALGORITHM

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].

- Given a set of *n* distinct objects, the k-Means clustering algorithm partitions the objects into *k* number of clusters such that intracluster similarity is high but the intercluster similarity is low.

- In this algorithm, user has to specify *k*, the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

K-MEANS ALGORITHM

The algorithm can be stated as follows.

- First it selects *k* number of objects at random from the set of n objects. These *k* objects are treated as the centroids or center of gravities of *k* clusters.

- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.

- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).

- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

K-MEANS ALGORITHM

Algorithm 24.1: k-Means clustering

Input: D is a dataset containing *n* objects, *k* is the number of cluster

Output: A set of *k* clusters

Steps:

1. Randomly choose *k* objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
   - Compute distance between the current objects and *k* cluster centroids
   - Assign the current object to that cluster to which it is closest.
3. Compute the "cluster centers" of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

K-MEANS ALGORITHM

Note:

1) Objects are defined in terms of set of attributes. $A = \{A_1, A_2, \ldots\ldots, A_m\}$ where each A_i is continuous data type.
2) Distance computation: Any distance such as L_1, L_2, L_3 or cosine similarity.
3) Minimum distance is the measure of closeness between an object and centroid.
4) Mean Calculation: It is the mean value of each attribute values of all objects.
5) Convergence criteria: Any one of the following are termination condition of the algorithm.
   - Number of maximum iteration permissible.
   - No change of centroid values in any cluster.
   - Zero (or no significant) movement(s) of object from one cluster to another.
   - Cluster quality reaches to a certain level of acceptance.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- **Table 24.1: 16 objects with two attributes A_1 and A_2.**

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Fig 24.1: Plotting data of Table 24.1

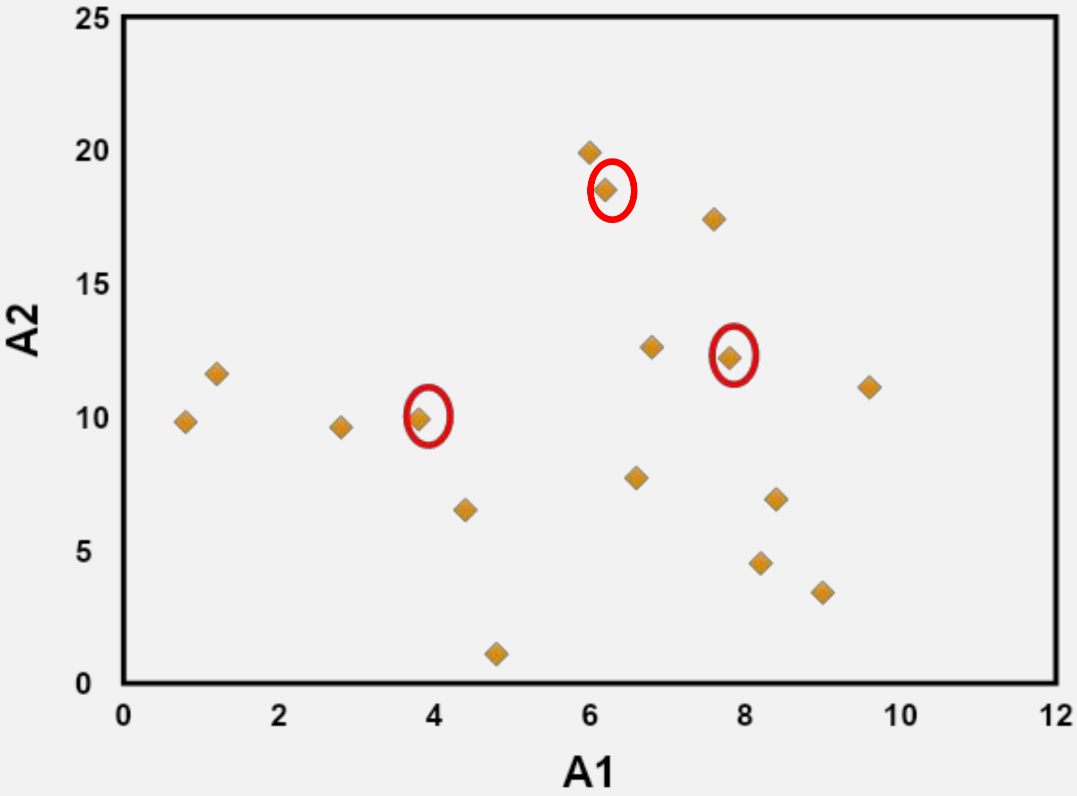


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- Suppose, k=3. Three objects are chosen at random shown as circled (see Fig 24.1). These three centroids are shown below.

Initial Centroids chosen randomly

Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1, d_2 and d_3 denote the distance from an object to c_1, c_2 and c_3 respectively. The distance calculations are shown in Table 24.2.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 24.2.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

Table 24.2: Distance calculation

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Fig 24.2: Initial cluster with respect to Table 24.2

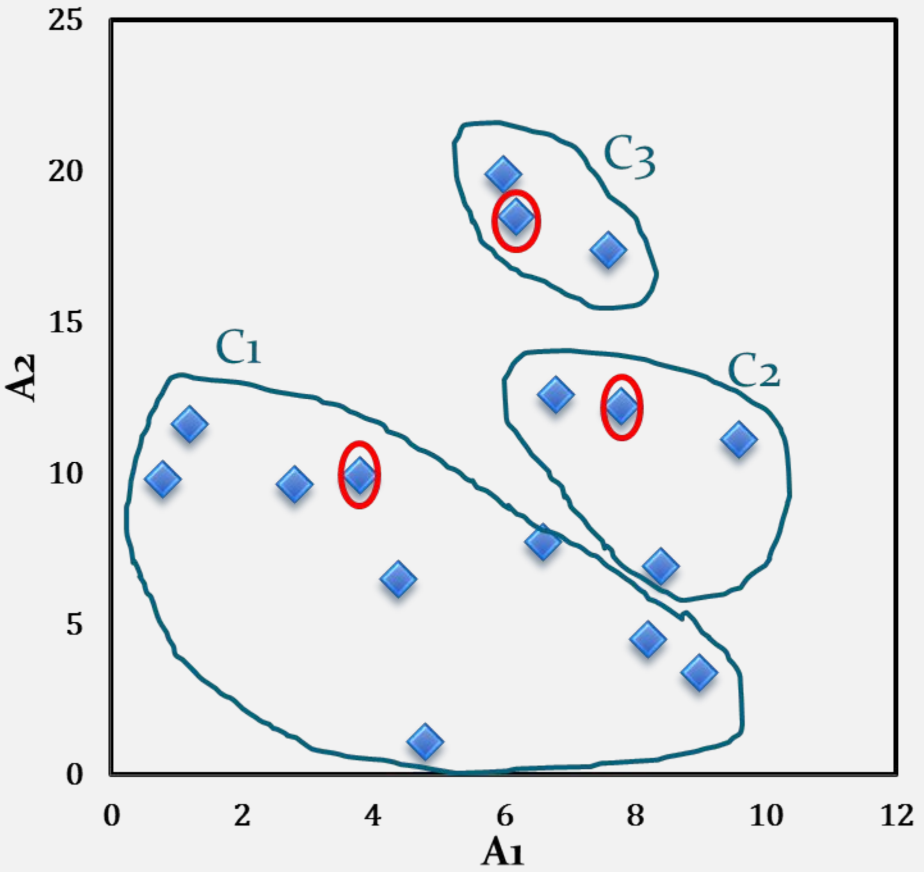


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 24.3.

Calculation of new centroids

New Centroid	Objects	
	A1	A2
c_1	4.6	7.1
c_2	8.2	10.7
c_3	6.6	18.6

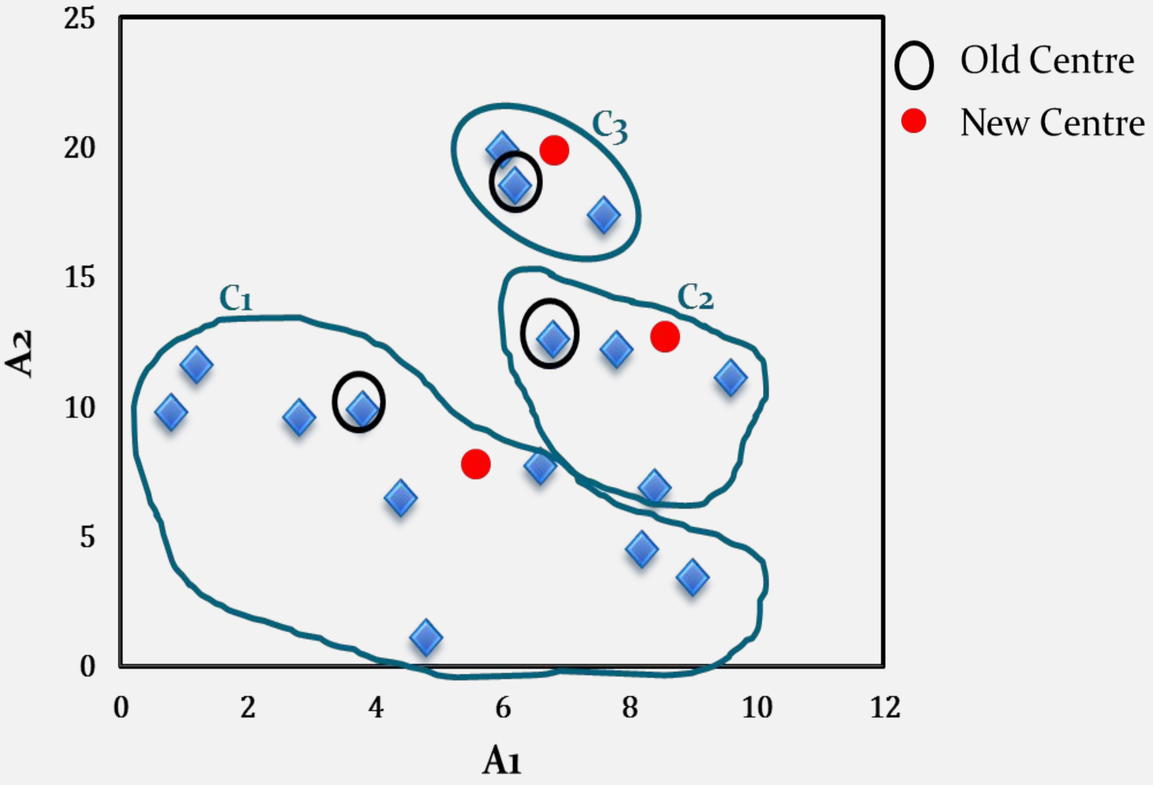


Fig 24.3: Initial cluster with new centroids

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 24.4.

Note that point p moves from cluster C_2 to cluster C_1.

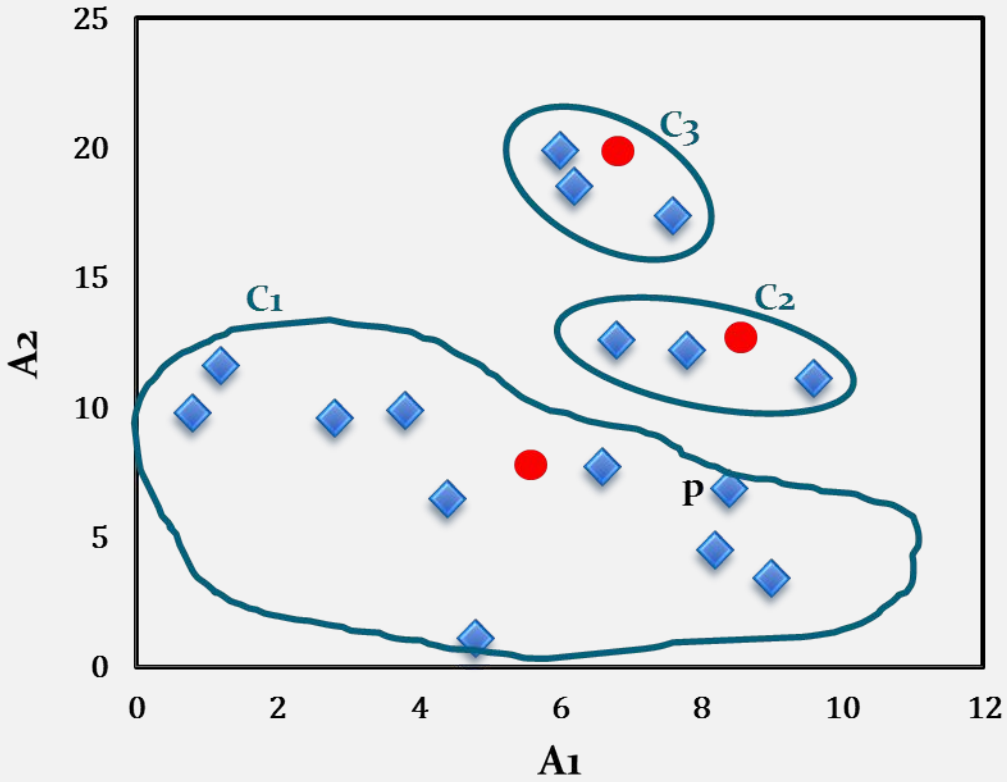


Fig 24.4: Cluster after first iteration

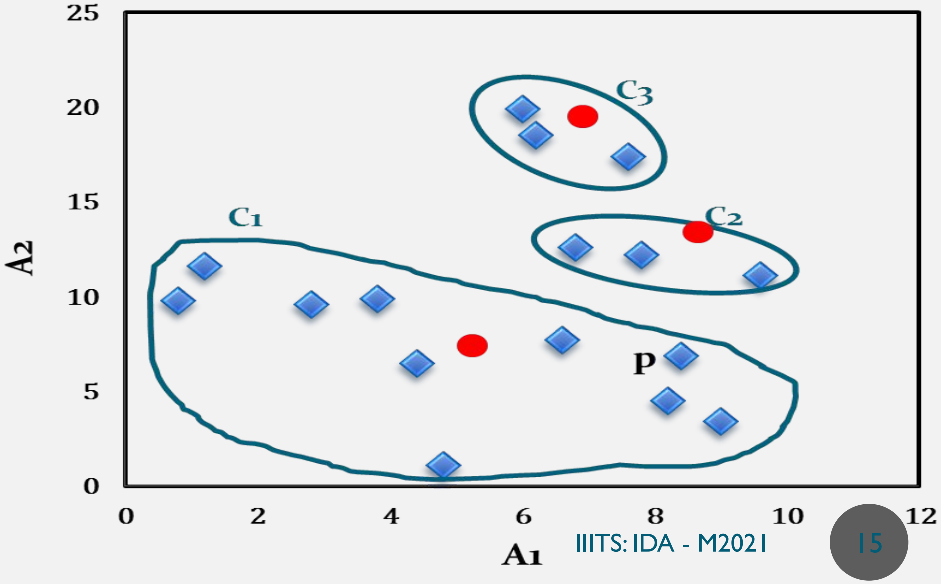
ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 24.5 is same as Fig 24.4.

Cluster centres after second iteration

Centroid	Revised Centroids	
	A1	A2
c_1	5.0	7.1
c_2	8.1	12.0
c_3	6.6	18.6

Fig 24.5: Cluster after Second iteration



COMMENTS ON K-MEANS ALGORITHM

Let us analyse the k-Means algorithm and discuss the pros and cons of the algorithm. We shall refer to the following notations in our discussion.

- Notations:
  - x : an object under clustering
  - n : number of objects under clustering
  - $\boldsymbol{C}_i$: the *i-th* cluster
  - c_i : the centroid of cluster $\boldsymbol{C}_i$
  - n_i : number of objects in the cluster $\boldsymbol{C}_i$
  - c : denotes the centroid of all objects
  - k : number of clusters

COMMENTS ON K-MEANS ALGORITHM

1. Value of k:

- The k-means algorithm produces only one set of clusters, for which, user must specify the desired number, *k* of clusters.
- In fact, *k* should be the best guess on the number of clusters present in the given data. Choosing the best value of *k* for a given dataset is, therefore, an issue.
- We may not have an idea about the possible number of clusters for high dimensional data, and for data that are not scatter-plotted.
- Further, possible number of clusters is hidden or ambiguous in image, audio, video and multimedia clustering applications etc.
- There is no principled way to know what the value of *k* ought to be. We may try with successive value of k starting with 2.
- The process is stopped when two consecutive *k* values produce more-or-less identical results (with respect to some cluster quality estimation).

COMMENTS ON K-MEANS ALGORITHM

Example 24.1: k versus cluster quality

- Usually, there is some objective function to be met as a goal of clustering. One such objective function is sum-square-error denoted by SSE and defined as

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2$$

- Here, $x - c_i$ denotes the error, if x is in cluster $\boldsymbol{C_i}$ with cluster centroid c_i.

- Usually, this error is measured as distance norms like L_1, L_2, L_3 or Cosine similarity, etc.

COMMENTS ON K-MEANS ALGORITHM

Example 24.1: k versus cluster quality

- With reference to an arbitrary experiment, suppose the following results are obtained.

k	SSE
1	62.8
2	12.3
3	9.4
4	9.3
5	9.2
6	9.1
7	9.05
8	9.0

- With respect to this observation, we can choose the value of $k \approx 3$, as with this smallest value of k it gives reasonably good result.

- Note: If $k = n$, then SSE=0; However, the cluster is useless!

COMMENTS ON K-MEANS ALGORITHM

2. Choosing initial centroids:

- Another requirement in the k-Means algorithm to choose initial cluster centroid for each *k* would be clusters.
- It is observed that the k-Means algorithm terminate whatever be the initial choice of the cluster centroids.
- It is also observed that initial choice influences the ultimate cluster quality. In other words, the result may be trapped into local optima, if initial centroids are chosen properly.
- One technique that is usually followed to avoid the above problem is to choose initial centroids in multiple runs, each with a different set of randomly chosen initial centroids, and then select the best cluster (with respect to some quality measurement criterion, e.g. SSE).
- However, this strategy suffers from the combinational explosion problem due to the number of all possible solutions.

COMMENTS ON K-MEANS ALGORITHM

2. Choosing initial centroids:

- A detail calculation reveals that there are $c(n,k)$ possible combinations to examine the search of global optima.

$$c(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}(i)^n$$

- For example, there are $o(10^{10})$ different ways to cluster 20 items into 4 clusters!

- Thus, the strategy having its own limitation is practical only if

1) The sample is negatively small (~100-1000), and
2) k is relatively small compared to n (i.e.. $k \ll n$).

COMMENTS ON K-MEANS ALGORITHM

3. Distance Measurement:

- To assign a point to the closest centroid, we need a proximity measure that should quantify the notion of "closest" for the objects under clustering.

- Usually Euclidean distance (L_2 norm) is the best measure when object points are defined in n-dimensional Euclidean space.

- Other measure namely cosine similarity is more appropriate when objects are of document type.

- Further, there may be other type of proximity measures that appropriate in the context of applications.

- For example, Manhattan distance (L_1 norm), Jaccard measure, etc.

COMMENTS ON K-MEANS ALGORITHM

3. Distance Measurement:

Thus, in the context of different measures, the sum-of-squared error (i.e., objective function/convergence criteria) of a clustering can be stated as under.

Data in Euclidean space (L_2 norm):

$$SSE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} (c_i - x)^2$$

Data in Euclidean space (L_1 norm):

The Manhattan distance (L_1 norm) is used as a proximity measure, where the objective is to minimize the sum-of-absolute error denoted as SAE and defined as

$$SAE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} |c_i - x|$$

COMMENTS ON K-MEANS ALGORITHM

Distance with document objects

Suppose a set of *n* document objects is defined as *d* document term matrix (DTM) (a typical look is shown in the below form).

Document	**Term**			
	t_1	t_2		t_n
D_1				
D_2				
D_n				

Here, the objective function, which is called Total cohesion denoted as TC and defined as

$$TC = \sum_{i=1}^{k} \sum_{x \in C_i} \cos(x, c_i)$$

where $\cos(x, c_i) = \frac{x \cdot c_i}{\|x\| \|c_i\|}$

$$x \cdot c_i = \Sigma_j \, x_j \, c_{ij} \quad \text{and} \quad \|x\| = \sqrt{\Sigma_j^p \, {x_j}^2}$$

$$\hat{x} = \Sigma_{j=1}^p \, \hat{x}_j \qquad \hat{c}_i = \Sigma_{j=1}^p \, \hat{c}_{ij} \qquad \|c_{ij}\| = \sqrt{\Sigma_j^p \, {c_{ij}}^2}$$

COMMENTS ON K-MEANS ALGORITHM

Note: The criteria of objective function with different proximity measures

1. SSE (using L_2 norm) : To minimize the SSE.
2. SAE (using L_1 norm) : To minimize the SAE.
3. TC(using cosine similarity) : To maximize the TC.

COMMENTS ON K-MEANS ALGORITHM

4. Type of objects under clustering:

- The k-Means algorithm can be applied only when the mean of the cluster is defined (hence it named k-Means). The cluster mean (also called centroid) of a cluster C_i is defined as

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

- In other words, the mean calculation assumed that each object is defined with numerical attribute(s). Thus, we cannot apply the k-Means to objects which are defined with categorical attributes.
- More precisely, the k-means algorithm require some definition of cluster mean exists, but not necessarily it does have as defined in the above equation.
- In fact, the k-Means is a very general clustering algorithm and can be used with a wide variety of data types, such as documents, time series, etc.

COMMENTS ON K-MEANS ALGORITHM

Note:

1) When SSE (L_2 norm) is used as objective function and the objective is to minimize, then the cluster centroid (i.e. mean) is the mean value of the objects in the cluster.

2) When the objective function is defined as SAE (L_1 norm), minimizing the objective function implies the cluster centroid as the median of the cluster.

The above two interpretations can be readily verified as given in the next slide.

COMMENTS ON K-MEANS ALGORITHM

Case 1: SSE

We know,

$$SSE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} (c_i - x)^2$$

To minimize SSE means, $\frac{\partial(SSE)}{\partial c_i} = 0$

Thus,

$$\frac{\partial}{\partial c_i}\left(\sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} (c_i - x)^2\right) = 0$$

Or,

$$\sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} \frac{\partial}{\partial c_i} (c_i - x)^2 = 0$$

COMMENTS ON K-MEANS ALGORITHM

Or,

$$\sum_{x \in C_i} 2(c_i - x) = 0$$

Or,

$$n_i \cdot c_i = \sum_{x \in C_i} x$$

Or,

$$\boxed{c_i = \frac{1}{n_i} \sum_{x \in C_i} x}$$

- Thus, **the best centroid for minimizing SSE of a cluster is the mean of the objects in the cluster**.

COMMENTS ON K-MEANS ALGORITHM

Case 2: SAE

We know,

$$SAE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} |c_i - x|$$

To minimize SAE means, $\frac{\partial(SAE)}{\partial c_i} = 0$

Thus,

$$\frac{\partial}{\partial c_i}\left(\sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} |c_i - x| \quad \right) = 0$$

Or,

$$\sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} \frac{\partial}{\partial c_i} |c_i - x| = 0$$

COMMENTS ON K-MEANS ALGORITHM

Or,

$$\sum_{x \in \boldsymbol{C}_i} \left\{ (x - c_i)\Big|_{if\ x > \boldsymbol{C}_i} + (c_i - x)\Big|_{if\ \boldsymbol{C}_i > x} \right\} = 0$$

Solving the above equation, we get

$$\boxed{c_i = median\,\{x | x \in \boldsymbol{C}_i\}}$$

- Thus, the best centroid for minimizing SAE of a cluster is the median of the objects in the cluster.

COMMENTS ON K-MEANS ALGORITHM

5. Complexity analysis of k-Means algorithm

Let us analyse the time and space complexities of k-Means algorithm.

Time complexity:

The time complexity of the k-Means algorithm can be expressed as

$$T(n) = O(n \times m \times k \times l)$$

where n = number of objects

m = number of attributes in the object definition

k = number of clusters

l = number of iterations.

Thus, time requirement is a linear order of number of objects and the algorithm runs in a modest time if $k \ll n$ and $l \ll n$ (the iteration can be moderately controlled to check the value of l).

COMMENTS ON K-MEANS ALGORITHM

5. Complexity analysis of k-Means algorithm

Space complexity: The storage complexity can be expressed as follows.

It requires $n \times m$ space to store the objects and $n \times k$ space to store the proximity measure from n objects to the centroids of k clusters.

Thus the total storage complexity is

$$S(n) = O(n \times (m + k))$$

That is, space requirement is in the linear order of n if $k \ll n$.

COMMENTS ON K-MEANS ALGORITHM

6. Final comments:

Advantages:

- k-Means is simple and can be used for a wide variety of object types.

- It is also efficient both from storage requirement and execution time point of views. By saving distance information from one iteration to the next, the actual number of distance calculations, that must be made can be reduced (specially, as it reaches towards the termination).

Limitations:

- The k-Means is not suitable for all types of data. For example, k-Means does not work on categorical data because mean cannot be defined.

- k-means finds a local optima and may actually minimize the global optimum.

COMMENTS ON K-MEANS ALGORITHM

6. Final comments:

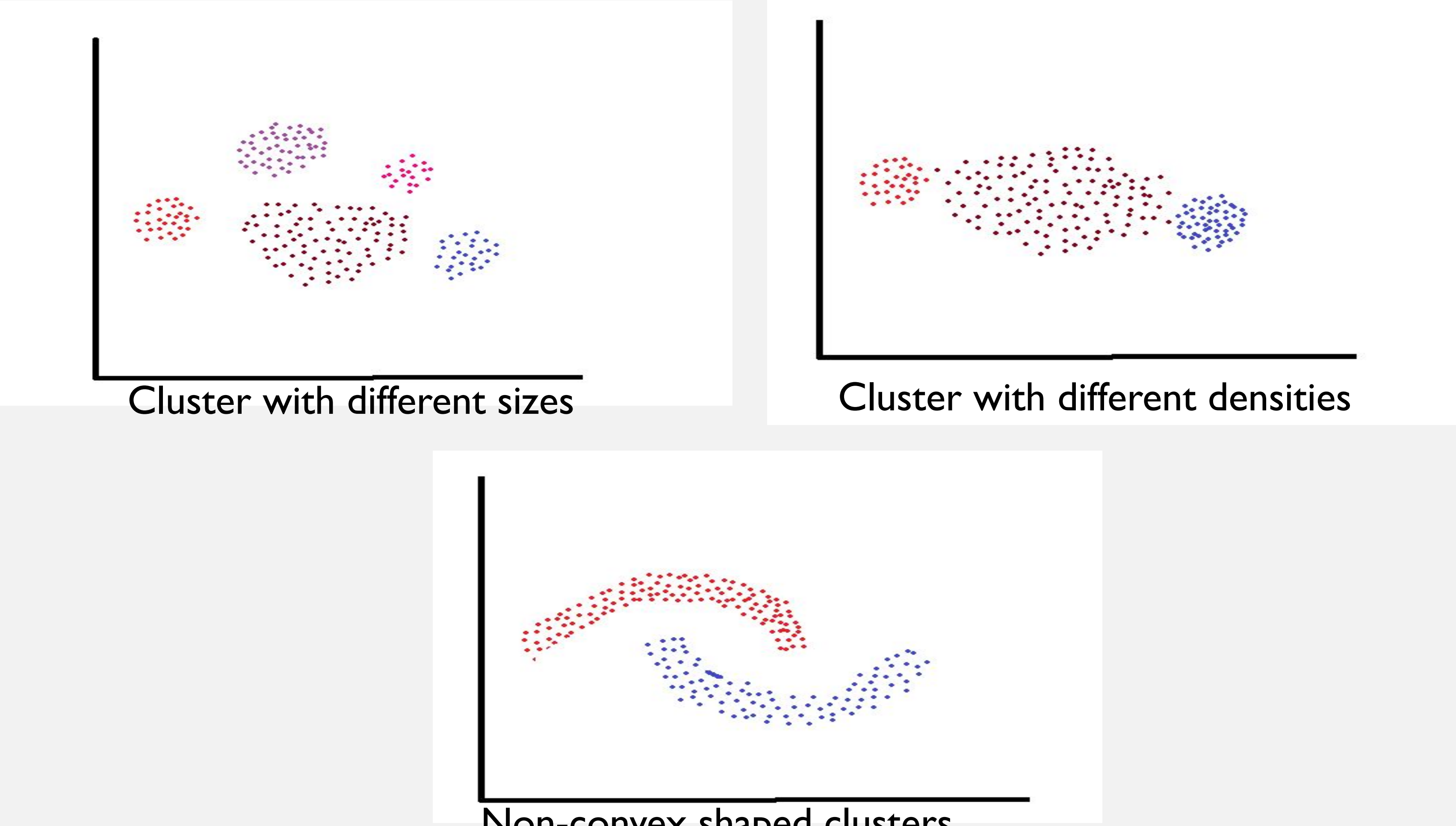
Limitations :

- k-means has trouble clustering data that contains outliers. When the SSE is used as objective function, outliers can unduly influence the cluster that are produced. More precisely, in the presence of outliers, the cluster centroids, in fact, not truly as representative as they would be otherwise. It also influence the SSE measure as well.

- k-Means algorithm cannot handle non-globular clusters, clusters of different sizes and densities (see Fig 24.6 in the next slide).

- k-Means algorithm not really beyond the scalability issue (and not so practical for large databases).

COMMENTS ON K-MEANS ALGORITHM



Cluster with different sizes

Cluster with different densities

Non-convex shaped clusters

Fig 24.6: Some failure instance of k-Means algorithm

DIFFERENT VARIANTS OF K-MEANS ALGORITHM

There are a quite few variants of the k-Means algorithm. These can differ in the procedure of selecting the initial *k* means, the calculation of proximity and strategy for calculating cluster means. Another variants of k-means to cluster categorical data.

Few variant of k-Means algorithm includes

- Bisecting k-Means (addressing the issue of initial choice of cluster means).

    I. M. Steinbach, G. Karypis and V. Kumar "A comparison of document clustering techniques", *Proceedings of KDD workshop on Text mining,* 2000.

- Mean of clusters (Proposing various strategies to define means and variants of means).
    - B. zhan "Generalised k-Harmonic means – Dynamic weighting of data in unsupervised learning", *Technical report, HP Labs*, 2000.
    - A. D. Chaturvedi, P. E. Green, J. D. Carroll, "k-Modes clustering", *Journal of classification*, Vol. 18, PP. 35-36, 2001.
    - D. Pelleg, A. Moore, "x-Means: Extending k-Means with efficient estimation of the number of clusters", *17th International conference on Machine Learning,* 2000.

DIFFERENT VARIANTS OF K-MEANS ALGORITHM

- N. B. Karayiannis, M. M. Randolph, “Non-Euclidean c-Means clustering algorithm”, *Intelligent data analysis journal*, Vol 7(5), PP 405-425, 2003.

- V. J. Olivera, W. Pedrycy, “Advances in Fuzzy clustering and its applications”, Edited book. John Wiley [2007]. (Fuzzy c-Means algorithm).

- A. K. Jain and R. C. Bubes, “Algorithms for clustering Data”, Prentice Hall, 1988.

  Online book at http://www.cse.msu.edu/~jain/clustering_Jain_Dubes.pdf

- A. K. Jain, M. N. Munty and P. J. Flynn, “Data clustering: A Review”, *ACM computing surveys*, 31(3), 264-323 [1999]. Also available online.

THE K-MEDOIDS ALGORITHM

Now, we shall study a variant of partitioning algorithm called k-Medoids algorithm.

Motivation: We have learnt that the k-Means algorithm is sensitive to outliers because an object with an "extremely large value" may substantially distort the distribution. The effect is particularly exacerbated due to the use of the SSE (sum-of-squared error) objective function. The k-Medoids algorithm aims to diminish the effect of outliers.

Basic concepts:

- The basic concepts of this algorithm is to select an object as a cluster center (one representative object per cluster) instead of taking the mean value of the objects in a cluster (as in k-Means algorithm).
- We call this cluster representative as a cluster medoid or simply medoid.

1. Initially, it selects a random set of *k* objects as the set of medoids.
2. Then at each step, all objects from the set of objects, which are not currently medoids are examined one by one to see if they should be medoids.

THE K-MEDOIDS ALGORITHM

- That is, the k-Medoids algorithm determines whether there is an object that should replace one of the current medoids.
- This is accomplished by looking all pair of medoid, non-medoid objects, and then choosing a pair that improves the objective function of clustering the best and exchange them.
- The sum-of-absolute error (SAE) function is used as the objective function.

$$SAE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i, x \notin M \text{ and } c_m \in M} |x - c_m|$$

Where c_m denotes a medoid

M is the set of all medoids at any instant

x is an object belongs to set of non-medoid object, that is, x belongs to some cluster and is not a medoid. i.e. $x \in \boldsymbol{C}_i, x \notin M$

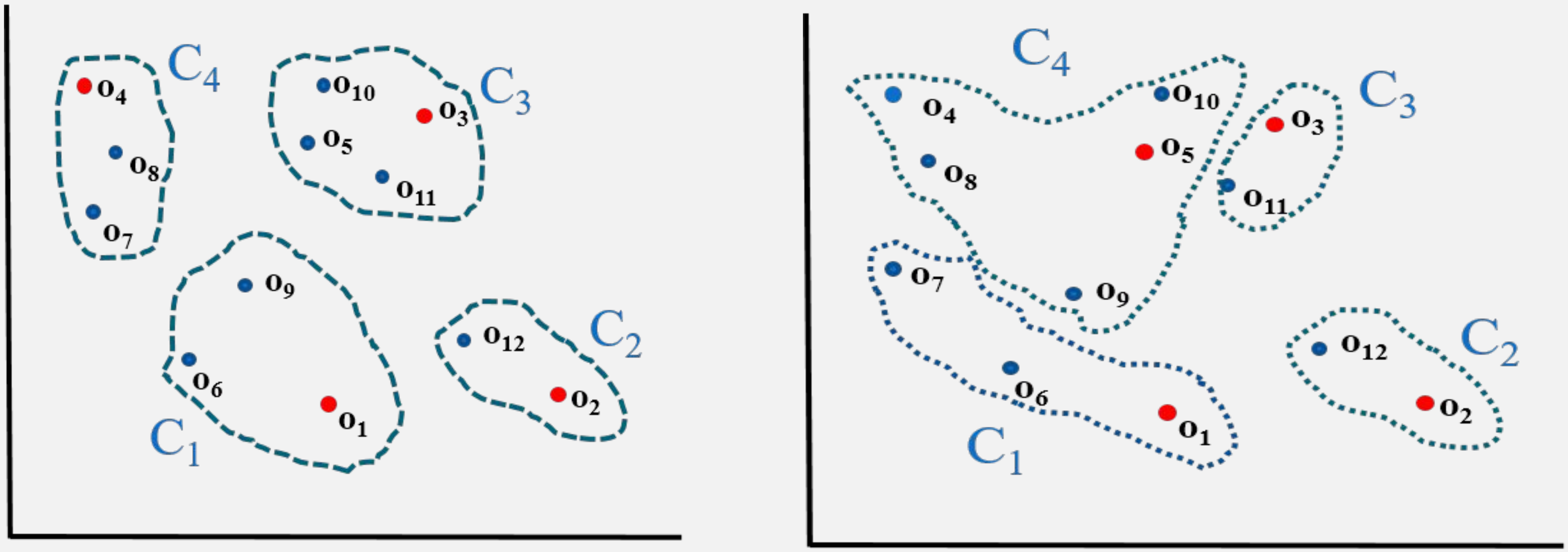
PAM (PARTITIONING AROUND MEDOIDS)

- For a given set of medoids, at any iteration, it select and exchange which has minimum SAE.
- The procedure terminates, if there is no change in SAE in successive iteration (i.e. there is no change in medoid).
- This k-Medoids algorithm is also known as PAM (Partitioning around Medoids).

Illustration of PAM

- Suppose, there are set of 12 objects $O(o_1, o_2, \ldots, o_{12})$ and we are to cluster them into four clusters. At any instant, the four cluster $C_1, C_2, C_3\ and\ C_4$ are shown in Fig. 24.7 (a). Also assume that $o_1, o_2,\ o_3, and\ o_4$ are the medoids in the clusters $C_1, C_2, C_3\ and\ C_4$, respectively. For this clustering we can calculate SAE.
- There are many ways to choose a non-medoid object to be replaced by any one medoid object. Out of these, suppose, if o_5 is considered as candidate medoid instead of $o_4,$ then it gives the lowest SAE. Thus, the new set of medoids would be $o_1, o_2, o_3, and\ o_5$. The new cluster is shown in Fig 24.7 (b).

PAM (PARTITIONING AROUND MEDOIDS)

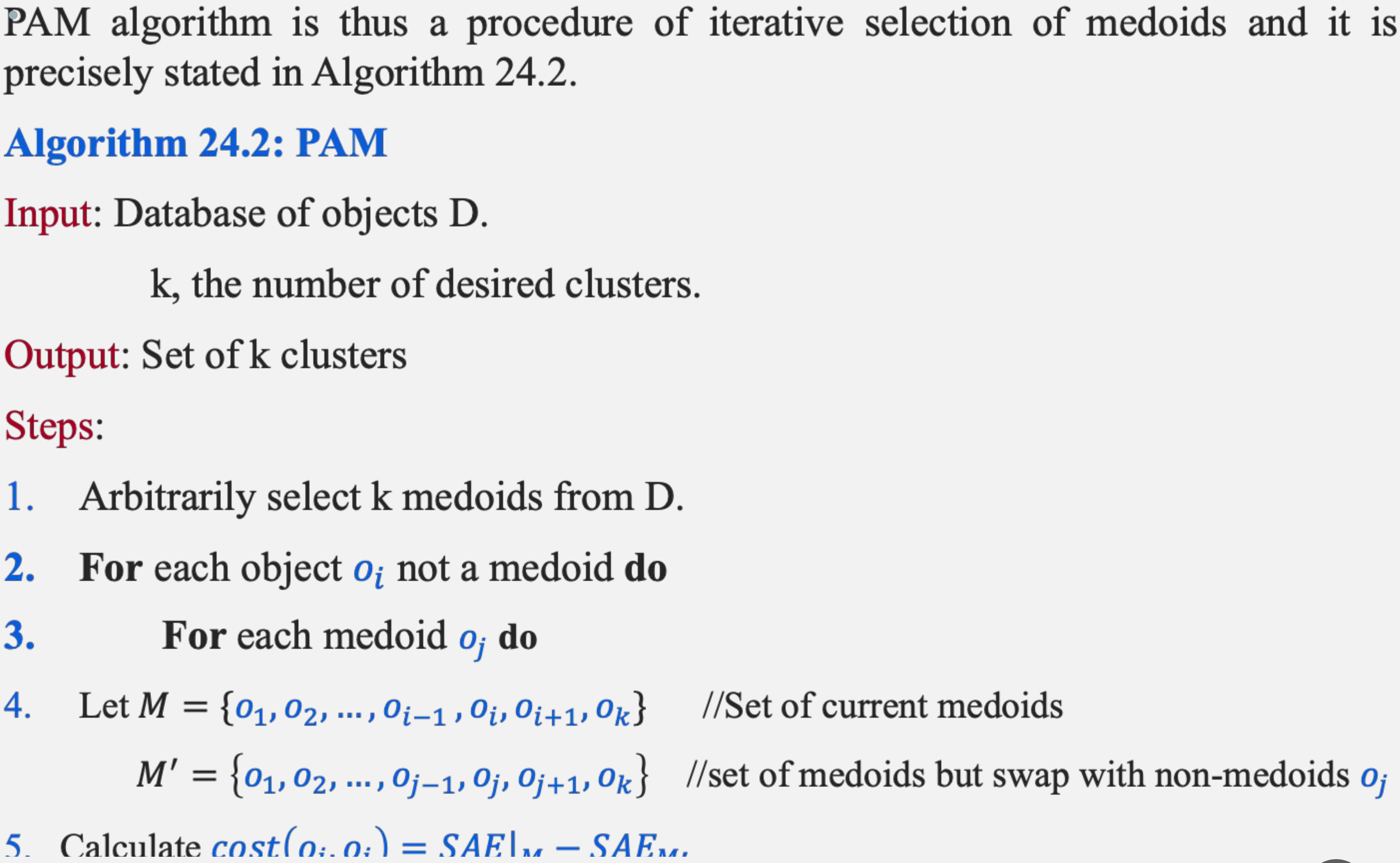


(a) Cluster with $o_1, o_2, o_3, and\ o_4$ as medoids

(b) Cluster after swapping $o_4\ and\ o_5$ (o_5 becomes the new medoid).

Fig 24.7: Illustration of PAM

PAM (PARTITIONING AROUND MEDOIDS)



PAM algorithm is thus a procedure of iterative selection of medoids and it is precisely stated in Algorithm 24.2.

Algorithm 24.2: PAM

Input: Database of objects D.

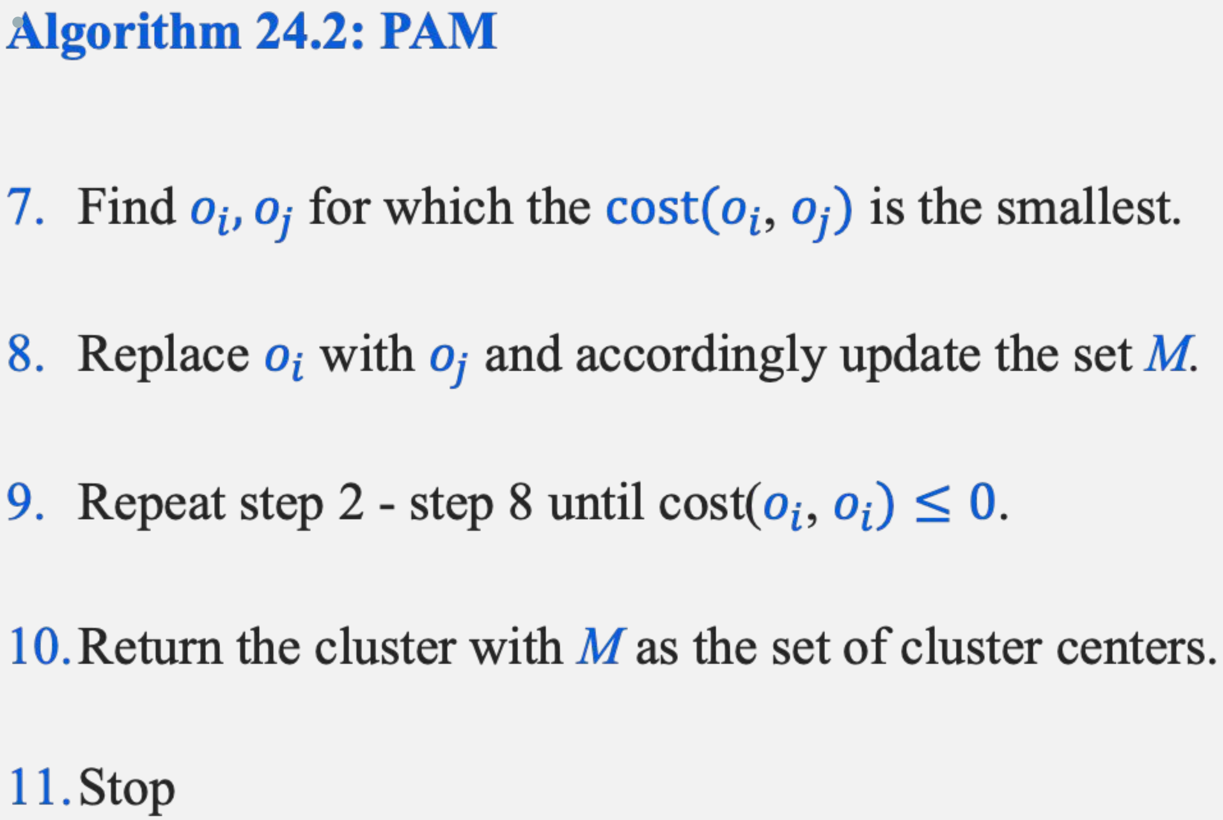
k, the number of desired clusters.

Output: Set of k clusters

Steps:

```
1.  Arbitrarily select k medoids from D.
2.  For each object $o_i$ not a medoid do
3.      For each medoid $o_j$ do
4.  Let $M = \{o_1, o_2, \dots, o_{i-1}, o_i, o_{i+1}, o_k\}$   //Set of current medoids
        $M' = \{o_1, o_2, \dots, o_{j-1}, o_j, o_{j+1}, o_k\}$   //set of medoids but swap with non-medoids $o_j$
5.  Calculate $cost(o_i, o_j) = SAE|_M - SAE_{M'}$
```

PAM (PARTITIONING AROUND MEDOIDS)



Algorithm 24.2: PAM

```
7. Find $o_i, o_j$ for which the cost($o_i$, $o_j$) is the smallest.
8. Replace $o_i$ with $o_j$ and accordingly update the set $M$.
9. Repeat step 2 - step 8 until cost($o_i$, $o_i$) $\leq 0$.
10. Return the cluster with $M$ as the set of cluster centers.
11. Stop
```

COMMENTS ON PAM

1. **Comparing k-Means with k-Medoids:**

- Both algorithms needs to fix k, the number of cluster prior to the algorithms. Also, both algorithm arbitrarily choose the initial cluster centroids.
- The k-Medoid method is more robust than k-Means in the presence of outliers, because a medoid is less influenced by outliers than a mean.

2. **Time complexity of PAM:**

- For each iteration, PAM consider $k(n-k)$ pairs of object o_i, o_j for which a cost $cost(o_i, o_j)$ determines. Calculating the cost during each iteration requires that the cost be calculated for all other non-medoids o_j. There are $n-k$ of these. Thus, the total time complexity per iteration is $n(n-k)^2$. The total number of iterations may be quite large.

3. **Applicability of PAM:**

- PAM does not scale well to large database because of its computation

OTHER VARIANTS OF K-MEDOIDS ALGORITHMS

- There are some variants of PAM that are targeted mainly large datasets are CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based upon RANdomized Search), it is an improvement of CLARA.

References:

For PAM and CLARA:

- L. kaufman and P. J. Rousseew, “Finding Groups in Data: An introduction to cluster analysis”, John and Wiley, 1990.

For CLARANS:

- R. Ng and J. Han, “Efficient and effective clustering method for spatial Data mining”, Proceeding very large databases [VLDB-94], 1994.

Any question?