Chapter 7. Cluster Analysis

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

Contents

What is Cluster Analysis?

- **Cluster: a collection of data objects**
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- **Cluster analysis**
  - Finding similarities between data according to the characteristics found in the data
    - Here, the **similarity measure** must be defined first
  - Grouping similar data objects into clusters
- **Unsupervised learning: no predefined classes**

What is Cluster Analysis?

- **Dissimilarity/Similarity metric:** Similarity is expressed in terms of a distance function, typically metric: *d* (*i, j*)

i j

- **The definitions of distance functions**
  - Usually very different for interval-scaled, Boolean, categorical, ordinal ratio, and numerical features

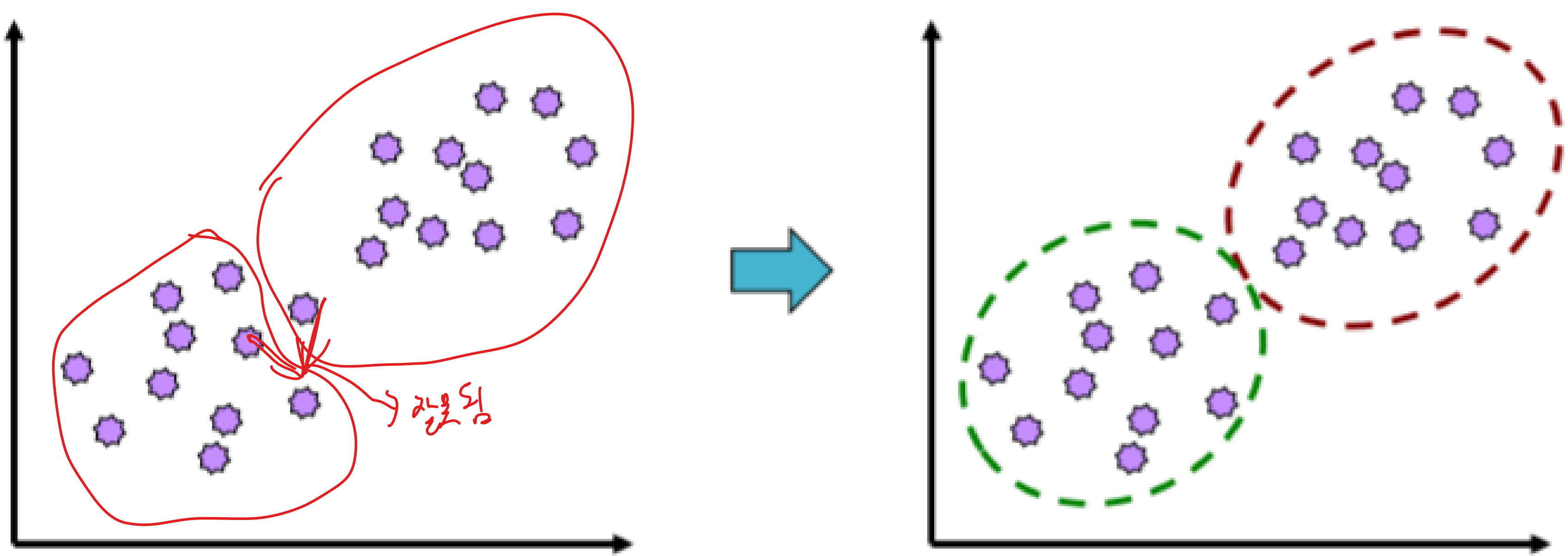
- **Hard to define "similar enough" or "good enough" of a cluster analysis**
  - The answer is typically highly subjective

주관적

Quality: What Is Good Clustering?

- **A good clustering method** will produce high quality clusters with
  - high intra-cluster similarity
  - low inter-cluster similarity

같은 cluster에 있는 점은 high similarity
다른 〃 low 〃



Applications

- **Data visualization and distribution analysis**
  - To know the data itself

- **Spatial Data Analysis** location 카페 마케팅 카페 어디에 차릴지
  - Detect spatial clusters or for other spatial mining tasks

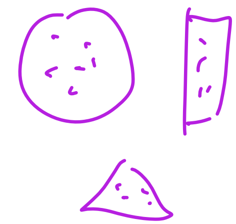
- **Economic Science (especially market research)**
  - Identify customers whose behaviors are similar

- **WWW**

  web log data
  - Cluster Weblog data to discover groups of similar access patterns

Requirements of Clustering / Distance Functions

- Ability to deal with different types of features
- Ability to handle dynamic data dynamic 더 들어오면
- Discovery of clusters with an arbitrary shape
  - Some algorithm may look for circle shapes, while others may not



- Minimal requirements for domain knowledge to determine hyper-parameters
- Able to deal with noises and outliers
- Insensitive to the order of input data 어떤 알고리즘은 input order가 바뀌면 결과가 바뀐다
- High dimensionality
- Scalability
- Incorporation of user-specified constraints

나이가 50~70 사이인 것
female인 것

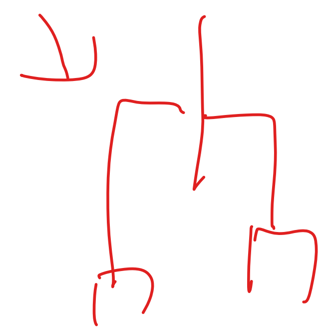
Contents

Major Clustering Approaches

- **Partitioning approach:**

clusters

  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of pairwise-distances within a cluster
  - Examples: k-means, k-medoids, CLARANS
- **Hierarchical approach:**
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Examples: Diana, Agnes, BIRCH, ROCK, CHAMELEON



- **Density-based approach:**
  - Based on some density functions
  - Examples: DBSACN, OPTICS

Centroid, Radius, and Diameter of a Cluster

- **Centroid:** the "middle" of a cluster
  - t_{ik}: a data point associated with the k-th cluster

$$c_k = \frac{\sum_{i=1}^{N}(t_{ik})}{N_k}$$

(1, 2, 3)
(2, 1, 4)
number of data inside cluster k
k th

- **Radius:** square root of an average squared distance from any point of the cluster to its centroid

작을수록 클러스터 ↑
cluster

$$R_k = \sqrt{\frac{\sum_{i=1}^{N}(t_{ik} - c_k)^2}{N}}$$

centroid

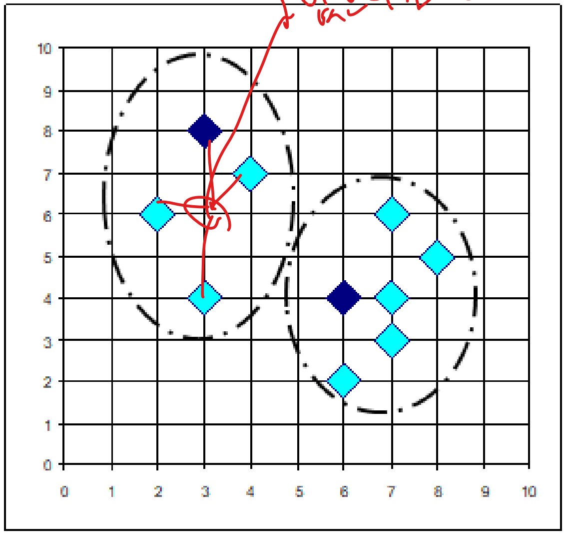
- **Diameter:** square root of an average squared distance between all possible pairs of points in the cluster

$$D_k = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}(t_{ik} - t_{jk})^2}{N(N-1)/2}}$$

nC_2

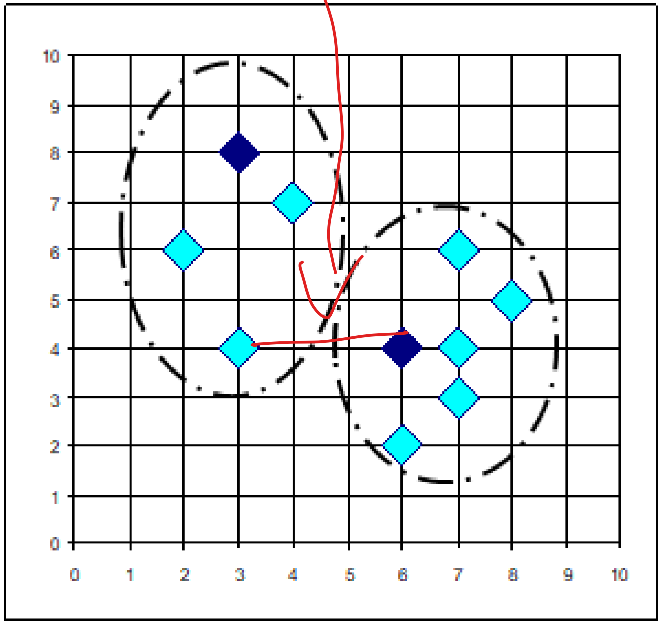
cluster 내의 모든 pair

없으니까 centroid



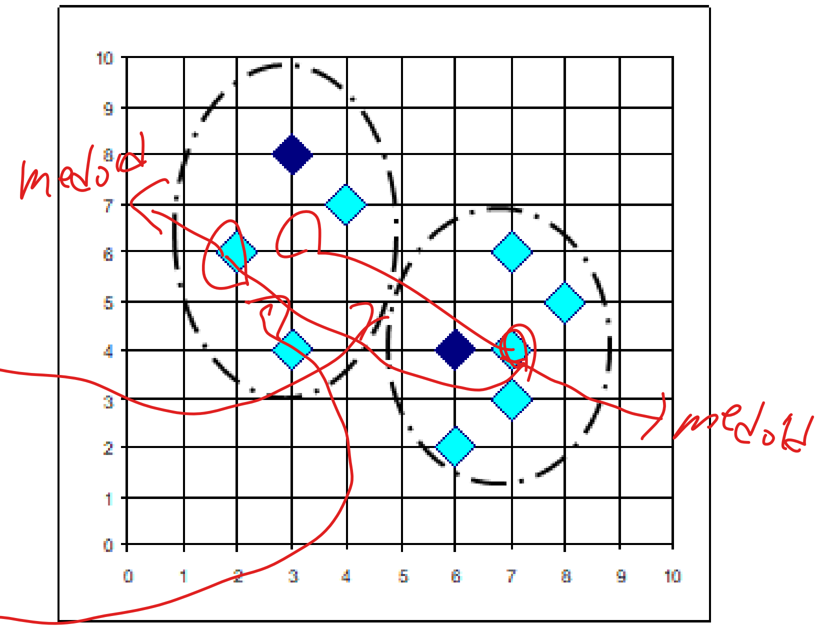
Distance between Clusters

- **Single link:** smallest distance between an element in one cluster and an element in the other
  - $\mathbf{dis(K_i, K_j) = min(t_{ip}, t_{jq})}$
- **Complete link:** largest distance between an element in one cluster and an element in the other
  - $\mathbf{dis(K_i, K_j) = max(t_{ip}, t_{jq})}$



Distance between Clusters

- **Average:** average distance between an element in one cluster and an element in the other
  - **dis(K_i, K_j) = avg(t_{ip}, t_{jq})** 모든 possible
- **Centroid:** distance between the centroids of two clusters
  not real object
  - **dis(K_i, K_j) = dis(c_i, c_j)**
- **Medoid:** distance between the medoids of two clusters
  real object
  - **dis(K_i, K_j) = dis(M_i, M_j)**
  - Medoid: one chosen, centrally located (real) object in the cluster



Contents

// clustering

Partitioning Methods: Basic Concept

- **Partitioning methods:** Construct a partition of *N* data points into a set of *K* clusters, having the minimum sum of squared distances of objects to their representative (e.g., centroid, medoid) of a cluster

the number of data in the m^th cluster

cluster의 data

m^th cluster

k cluster

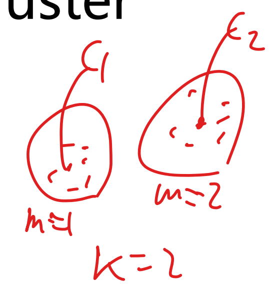
each cluster

$$\sum_{m=1}^{K}\sum_{i=1}^{N_m}(\boldsymbol{c}_m - \boldsymbol{t}_i^m)^2$$

representative
ex) centroid

i^th data point

the number of (pre-defined) clusters



minimize 할수록 좋다

- **Given *K*, find a partition of *K* clusters that minimizes the above partitioning criterion**
  - Global optimal: evaluate "all" possible partitions (impossible!)
  - Heuristic methods: *k-means* and *k-medoids* algorithms
    - **k-means**: Each cluster is represented by the centroid of the cluster
    - **k-medoids** : Each cluster is represented by one of the objects (medoid) in the cluster

The *K-Means* Clustering Method

- ***K-means* clustering algorithm is implemented in four steps:**

1. Randomly select *k* seed points, and assign objects to one of those seeds.

real
not real

2. **Compute** seed points as the centroids of the clusters of the current partition

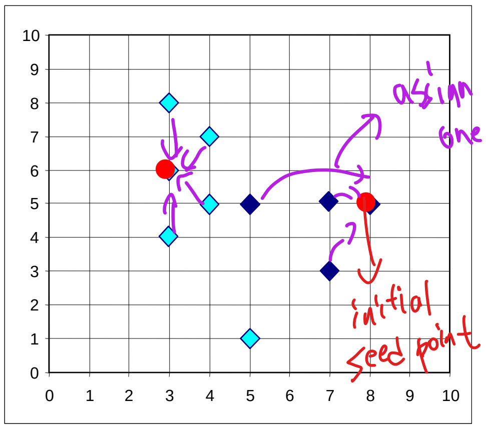
repeat

3. **Assign** each object to the nearest centroid, forming new clusters

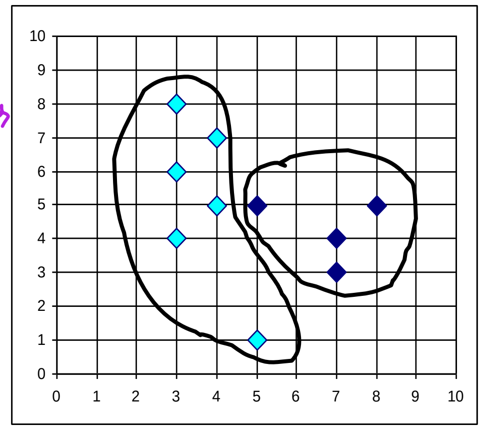
4. Go back to Step 2, stop when no more new assignment

The *K-Means* Clustering Method

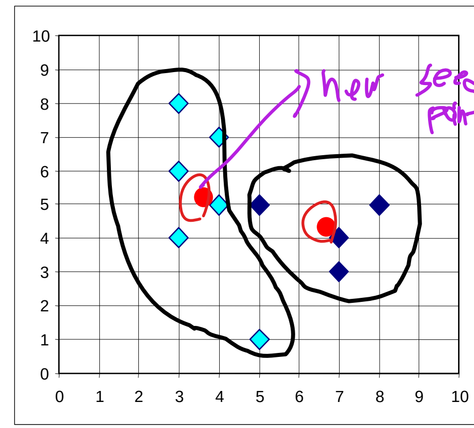
- **Iterative solution:** compute centroid, then compute assignments, then compute centroid, ...



Assign each objects to most similar center



Update the cluster means

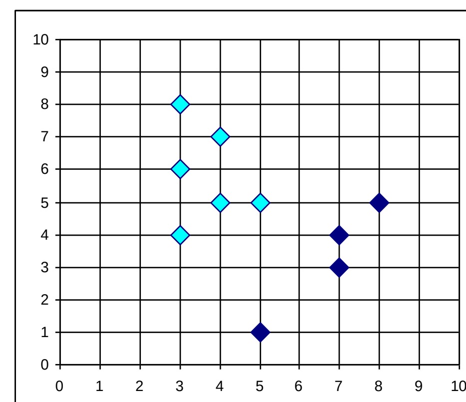


reassign

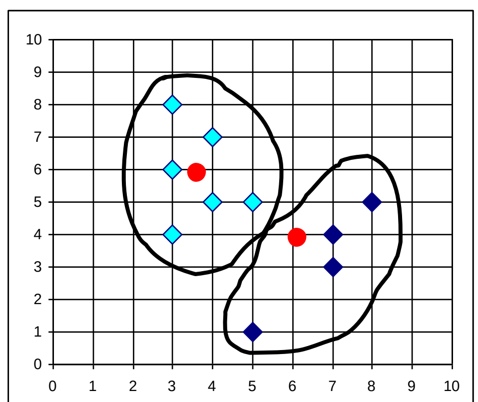
K=2

Randomly choose 2 objects as the initial 2 cluster centers

reassign



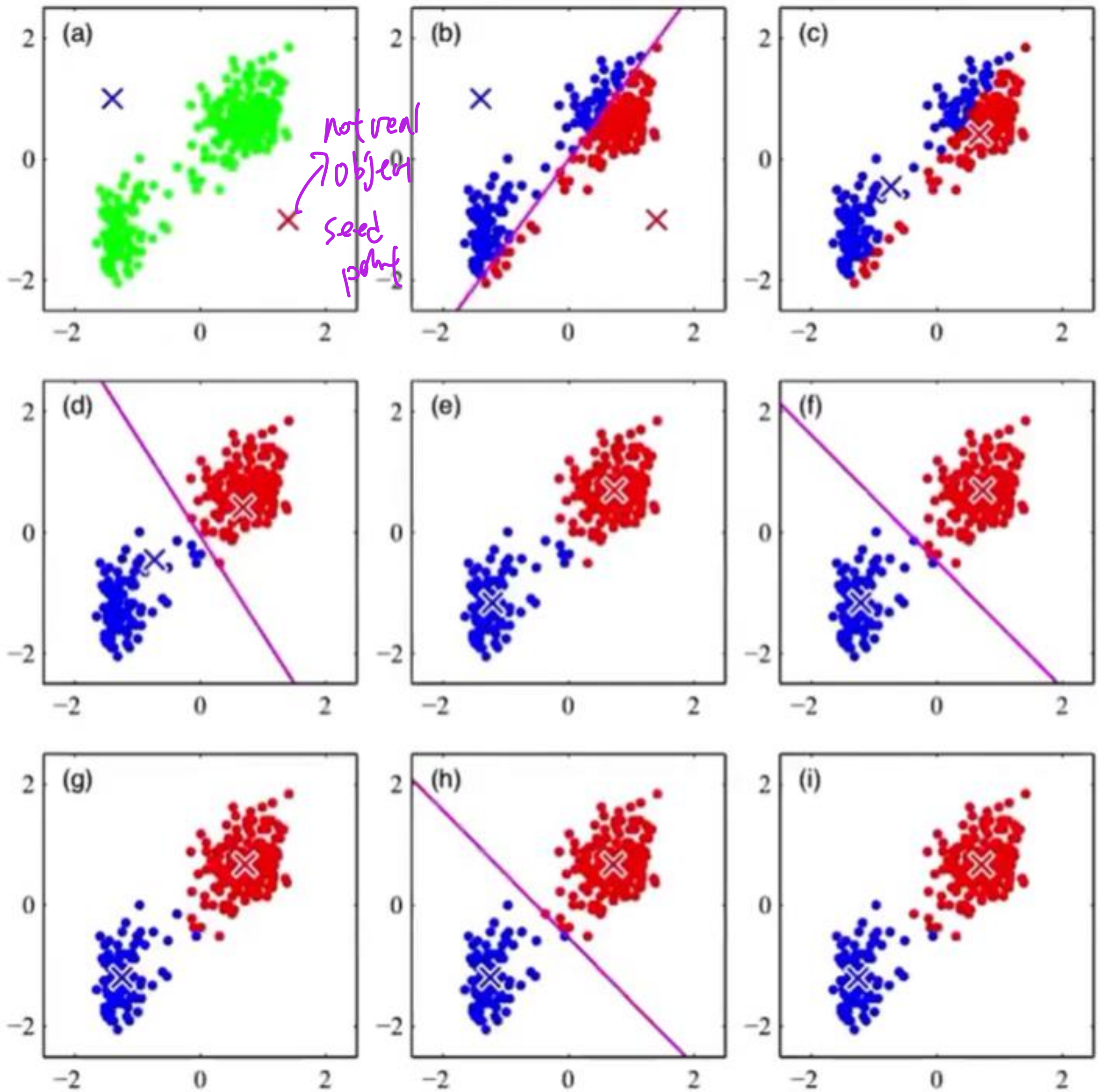
Update the cluster means



chang
없으면
stop

The *K-Means* Clustering Method

❑ Example



Thank You