DA-Final-5) Clustering

1. Clustering Introduction

1-1. Clustering 정의

- Cluster: A collection of (similar) data objects,
  - Similar (or related) to one another within the same group
  - Dissimilar (or unrelated) to the objects in other groups
- Clustering (or cluster analysis, segmentation, …)
  - Grouping similar data objects into a cluster
  - Unsupervised learning (no predefined classes or labels)

Unsupervised learning & Data grouping

- Why clustering?
  1. To get insight into data distribution with applications to
     - Biology, information retrieval, marketing, climate analysis, land use analysis, city planning, etc.
  2. To preprocess data before running other algorithms, e.g.
     - Micro-clustering (summarization) for supervised learning or data compression
     - Clustering for outlier detection or finding nearest neighbors

     ex. streaming data → micro-clustering → Data mining

1-2. Clustering quality

- A good clustering method will produce high quality clusters
  - High intra-class similarity: cohesive within clusters
    (내부) (응집력)
  - Low inter-class similarity: distinctive between clusters
    (외부)
- The quality of a clustering method depends on
  - The similarity measure used by the method
  - Its ability to discover some or all of the hidden patterns
- Similarity or distance metric:
  - The definitions of distance functions are usually different depending on the type of attributes such as interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
  - Weights should be associated with different variables considering the applications and data semantics
- Quality of clustering:
  - There is usually a separate "quality" function that measures the "goodness" of a cluster.
  - It is hard to define "similar enough" or "good enough"
  - The answer is typically highly subjective
    (주관적인)

1-3. Clustering에서 고려해야 할 점들

- Partitioning criteria
  - Single level vs. (multi-level) hierarchical clustering
- Separation of clusters
  - Exclusive vs. non-exclusive (i.e. each object may belong to multiple clusters)
- Similarity measure
  - Distance-based (e.g. Euclidian, road network, vector) vs. connectivity-based (e.g. density or contiguity)
    (연속성)
  - Multiple types of attributes: Numerical, binary, categorical, ordinal, linked, etc.
- Clustering space
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

    → distance가 의미X, clustering 어렵다.

    → low dim으로 mapping해 clustering 가능

    > Dimensionality reduction
- Constraint-based clustering
  - Constraints may be given by users or domain knowledge
- Clustering shape and size
  - Spherical shape and equivalent size (e.g. k-means), arbitrary shape and inequivalent size (e.g. DBSCAN)
- Incremental clustering and insensitivity to input order
  (무감각)

  streaming data → 순서와 상관없이 Data mining해야 하는데 어렵다.

1-4. Clustering methods

많이 사용

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g. minimizing the sum of square errors
  - k-means, k-medoids
- Hierarchical approach:
  - Create a hierarchical (multi-level) clusters
  - Agglomerative clustering (AGNES), Diana, BIRCH, CAMELEON

    (병합)
- Density-based approach:
  - Based on connectivity or density
  - DBSCAN, OPTICS, DenClue
- Grid-based approach:
  - based on the grid of feature space
  - STING, WaveCluster, CLIQUE
- Model-based:
  - A model is hypothesized for each cluster and tries to find the best fit of that model
  - EM, SOM, COBWEB
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - COD (obstacles), constrained clustering
- Link-based clustering:
  - Objects are often linked together in various ways
  - SimRank, LinkClus

2.Clustering methods

2-1. Partitioning methods

Ex. K-means, K-medoids, k-modes

- Partitioning method: Partitioning m objects into K clusters, such that the sum of distances is minimized.
- Input: $D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$
- Output: assignment of each object to a cluster $[z_1, \dots, z_m]$ (zi에 대해) where $z_i \in \{1, \dots, K\}$
- Objective:

각 cluster의 centroid

$$L(\mathbf{z}, \mathbf{c}) = \sum_{i=1}^{m} dist\,(\mathbf{x}_i - \mathbf{c}_{z_i})$$

- Given K, find a partition of K clusters that minimizes the objective (combinatorial problem). (조합)
- Heuristic methods:
  - mean ← k-means (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster centroid
  - median ← k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster
  - k-modes: for categorical data
    ↳ 가장 빈번하게 등장하는 값

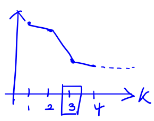
2-1-(1). K-means clustering

- Algorithms
  1. Randomly set K centroids
  2. Assign $\mathbf{z}$ to their nearest centroids.
  3. Compute centroids as the mean points of each cluster
  4. Repeat Step 2 until $\mathbf{z}$ do not change
- Example: one-dimensional data
  - Input: $D = \{0,2,10,12\}$
  - Output: $[z_1, z_2, z_3, z_4]$ where $z_i \in \{1,2\}$ $(K = 2)$ 2 clusters
  - Initialization (random): $c_1 = 0, c_2 = 2$:
  - Iteration 1:
    - Step 1: $z_1 = 1, z_2 = 2, z_3 = 2, z_4 = 2$
    - Step 2: $c_1 = 0, c_2 = 8$
  - Iteration 2: $\frac{0}{1}=0$ $\frac{2+10+12}{3}=8$
    - Step 1: $z_1 = 1, z_2 = 1, z_3 = 2, z_4 = 2$:
    - Step 2: $c_1 = 1, c_2 = 11$

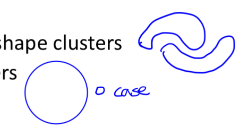
zi가 same!

$\frac{0+2}{2}=1$ $\frac{10+12}{2}=11$

- Need distance function
  - Distance function must be designed carefully reflecting domain knowledge.
  - For categorical data, means can be replaced by modes (i.e. k-modes).
- Find local optima
  - Results are different depending on the initial centroids.
  - Advanced heuristics are proposed (e.g. k-means++).
- Need to specify K
  - How to adjust K?
  - Compute the loss for every K and find where the loss decreases rapidly.



- Cluster shape
  - Spherical shape: not proper for non-convex shape clusters
  - Equivalent size: not proper for skewed clusters

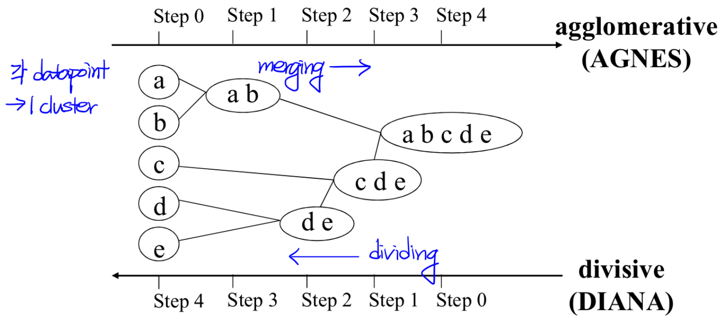


2-1-(2). K-medoids clustering (PAM)

- Motivation: Means is more sensitive to outliers than median.
- PAM uses the most centrally located object as the reference points and the rest is the same as k-means.
- Computing the central points requires additional scan of data. → 시간↑ > K-means
- Improvements to reduce the computational overhead are proposed (e.g. CLARA, CLARANS)

2-2. Hierarchical methods

- Do not require the number of clusters as input

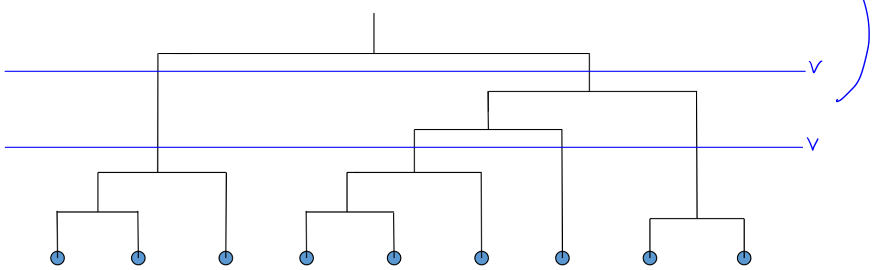


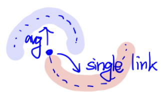
Agglomerative(상향식) clustering (AGNES)

- Repeatedly merge nodes that are similar the most until all nodes are merged to a single cluster.

Divisive clustering(DIANA)

- Start a single cluster of all nodes and repeatedly divide them until every node forms a cluster on its own.
  cluster 내부 노드들 간 거리를 계산 → 시간↑
- Dendrogram: a tree of clusters
  (계통수)
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level.





Distance measure between clusters

- **Single link**: smallest distance between an element in one cluster and an element in the other → non-convex shape
- **Complete link**: largest distance between an element in one cluster and an element in the other
- **Average**: avg distance between an element in one cluster and an element in the other → convex shape (ex. sphere)
- **Centroid**: distance between the centroids of two clusters
- **Medoid**: distance between the medoids of two clusters

빠르게 계산가능 (update와 추가) ↔ 다른 measure는 특정 data point를 찾아야한다.

Distance measure affects on the shape of the final clusters; especially, using single link would produce clusters of non-convex shape.

Major weakness of hierarchical clustering methods:

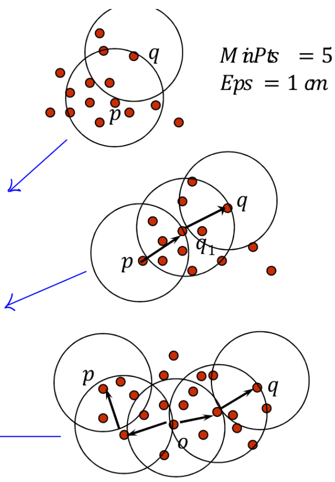
- Can never undo what was done previously
- Do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects

Advanced hierarchical clustering methods:

- BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
- CHAMELEON (1999): hierarchical clustering using dynamic modeling

2-3. Density-based methods

- Density-based clustering features:
  - Discover clusters of arbitrary shape.
  - Handle noise well.
  - Need one scan of data at most.
  - Need density parameters as termination condition.

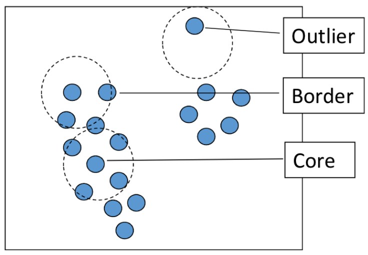


- Two parameters:
  - epsilon (ε) Eps: Maximum radius of the neighborhood from p
  - min points $MinPts$: Minimum number of points within Eps
- $N_{Eps}(p)$: $\{q \mid \mathrm{dist}(p,q) \le Eps\}$, the points within Eps
- p is a core point if $|N_{Eps}(p)| \ge MinPts$.
- *Directly density-reachable*: : core point p의 eps 반경 내에 존재
  - A point q is directly density-reachable from a point p w.r.t Eps and $MinPts$, if $q \in N_{Eps}(p)$, and p is a core point.
- *Density-reachable*: eps 반경 chain으로 도달가능
  - A point q is density-reachable from a point p w.r.t. Eps and $MinPts$, if there is a chain of points $q_1, \ldots, q_n$, $q_1 = p$, $q_n = q$ such that q_{i+1} is directly density-reachable from q_i.
- *Density-connected*:
  - A point q is density-connected to a point p w.r.t. Eps and $MinPts$, if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$.

2-3-(1). DBSCAN

DBSCAN (Density-based spatial clustering of applications with noise)

- A cluster is defined as a maximal set of density-connected points.
- Discovers clusters of arbitrary shape.
- Outliers are those not density-connected to any clusters.



$Eps = 1cm$
$MinPts = 5$

Algorithm

- Arbitrary select a point p.
- Retrieve all points density-reachable from p w.r.t. Eps and $MinPts$.
- If p is a core point, a cluster is formed. (directly density-reachable)
- If p is not, nothing is density-reachable, so visit the next point of the database.
- Continue the process until all of the points have been processed.

2-3-(2). OPTICS

(DBSCAN의 확장)
- OPTICS: Ordering Points To Identify the Clustering Structure [SIGMOD'99]
  - Detect meaningful clusters in data of varying density
  - Slower than DBSCAN

2-3-(3). DENCLUE

- DENCLUE:
  - DENsity-based CLUstering by Hinneburg & Keim [KDD'98]
  - A cluster is modeled by kernel density estimation (KDE)
- Influence of y on x:

$$f_{Gauss}(x, y) = e^{-\frac{||x-y||^2}{2\sigma}}$$

- Total influence on x:

$$f(x) = \sum_{i=1}^{n} f_{Gauss}(x, x_i)$$

- Overall density of the data space can be calculated as the sum of the influence function of all data points
- Clusters can be determined mathematically by identifying density attractors, which are local maxima of the overall density function. overall 밀도가 최대한 높은 곳!
- Data points are assigned to density attractors by hill climbing, i.e., points going to the same local maximum are put into the same cluster.
- Merge density attractors that are connected through paths of high density (> threshold)
  granularity 결정

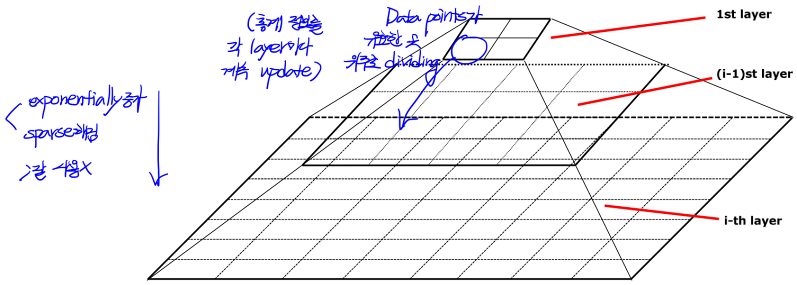
σ: 분산 척도, ε: 경계척도

2-4. Grid-based methods

- Using multi-resolution grid data structure
- Several interesting methods
  - **STING** [VLDB'97]
    - STatistical INformation Grid approach
  - **CLIQUE** [SIGMOD'98]
    - Both grid-based and subspace clustering

2-4-(1). STING

- Wang, Yang and Muntz (VLDB'97):
  - The spatial area is divided into rectangular cells.
  - There are several levels of cells corresponding to different levels of resolution.
    (layers)



- Each cell at a high level is partitioned into a number of smaller cells in the next lower level.
- Statistical information of each cell is calculated and stored beforehand and is used to answer queries.
  - Count, mean, std, min, max.
  - Type of distribution—normal, uniform, etc.
  - Parameters of higher level cells can be easily calculated from parameters of lower level cell.
- Use a top-down approach to answer spatial data queries:
  - Remove the irrelevant cells from further consideration
  - When finish examining the current layer, proceed to the next lower level
  - Repeat this process until the bottom layer is reached
- Advantages:
  - Query-independent, easy to parallelize, incremental update
- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

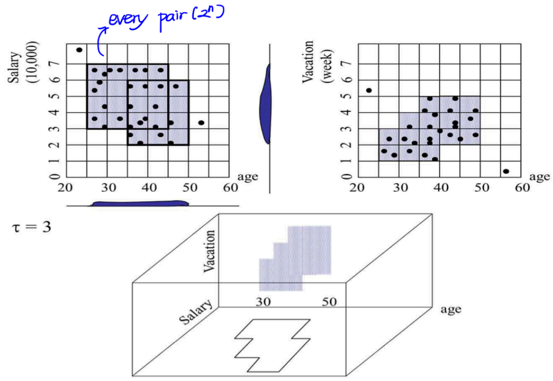
2-4-(2). CLIQUE

Clustering In Quest ; subspace clustering (high-dimensional clustering은 어려우므로)

- Agrawal, Gehrke, Gunopulos, Raghavan [SIGMOD'98]:
  - Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space.
- CLIQUE can be considered as both density-based and grid-based.
  - It partitions each dimension into the same number of equal length unit.
  - A unit is non-overlapping rectangular area in a subspace.
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input parameter.
  - A cluster is a maximal set of connected dense units within a subspace.

Grid based

Density based



- Identify the subspaces that contain clusters using the *Apriori* principle. (refer to Apriori algorithm in Wiki.)
  - Partition each dimension and find dense units on each dimension.
  - Merge two dimensions, each of which has dense units.
    - When merging two dimensions the size of each unit reduces.
    - If no unit becomes dense, no need to extend the subspace.

A B C D E

↓ (dense units)만 남겨서 합치기

BC BD BE

↓ (dense units)

BDE CDE

- Identify clusters:
  - Determine dense units in all subspaces of interests.
  - Determine connected dense units in all subspaces of interests.
- Properties
  - Find the subspaces of the highest dimensionality where clusters exist.
  - Find clusters of arbitrary shape without presuming any canonical data distribution.
  - Scales linearly with the size of input.
  - Theoretically exponential but practically scalable to some extent w.r.t. the number of dimensions.

3. Clustering summary

- Clustering groups objects based on their similarity and has wide applications.
- Clustering methods can be categorized into partitioning, hierarchical, density-based, grid-based, and model-based methods.
- Measures of distance between clusters are variant and determine the shape of clusters.

장단점

- Partitioning methods (e.g. k-means) are simple and efficient but produce clusters of spherical shape and equivalent size.
- Hierarchical methods produce multi-level clusters (or dendrogram) of convex or non-convex shape but not scale well (at least $O(n^2)$).
  Avg Single-link
- Density-based methods (e.g. DBSCAN) and grid-based methods (e.g. CLIQUE) produce clusters of arbitrary shape and detect outliers as byproduct but need a careful tuning of parameters.
- Clustering is usually done for low-dimensional data and is hard in high-dimensional space, because many important distance metrics (e.g. Euclidean distance) become not meaningful in high-dimensional space.
  다 멀어져버린다!
- An interesting research topic is clustering high dimensional data based on *deep learning* approaches.