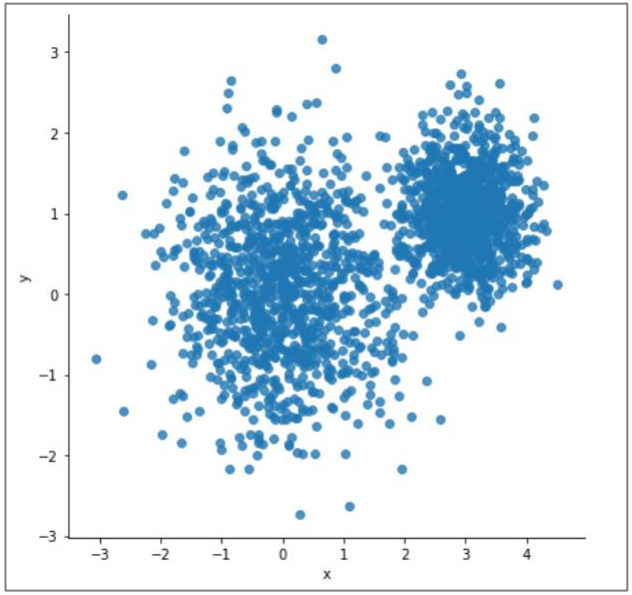
데이터 마이닝 개념과 기법

클러스터 분석: 기본 개념과 방법론

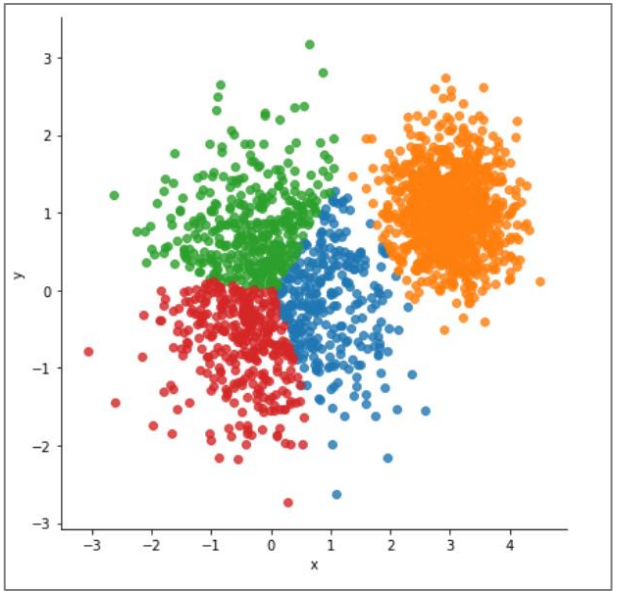
최우진

10.1 Cluster Analysis

- Cluster analysis(Clustering) : 데이터 오브젝트 집합을 부분집합으로 분할하는 것.
- Clustering은 데이터 안에서 기존에는 알 수 없었던 그룹을 발견 하려 할 때 유용



클러스터링 전



클러스터링 후

10.1 Cluster Analysis

- 같은 Cluster의 데이터는 특성이 유사하며, 서로 다른 Cluster의 데이터는 상이

  따라서 각 Cluster를 암묵적으로 'Class'로써 생각

- 이러한 관점에서 Cluster를 '자동 분류(automatic classification)'으로 부름
- 데이터 구획화, 아웃라이어 탐색에 이용
- 비지도학습 (클래스 라벨 정보 없이 분류 진행)

10.1.2 Requirements for Cluster Analysis

- 처리 규모(Scalability)

  대부분의 클러스터링 알고리즘은 작은 데이터세트에선 잘 작동하지만 그 결과가 편향되기가 쉬우므로, 대용량 데이터를 잘 처리해야 좋은 클러스터링 알고리즘이다.

- 다양한 데이터 형식에 대한 수용성(Ability to deal with different types of attributes)

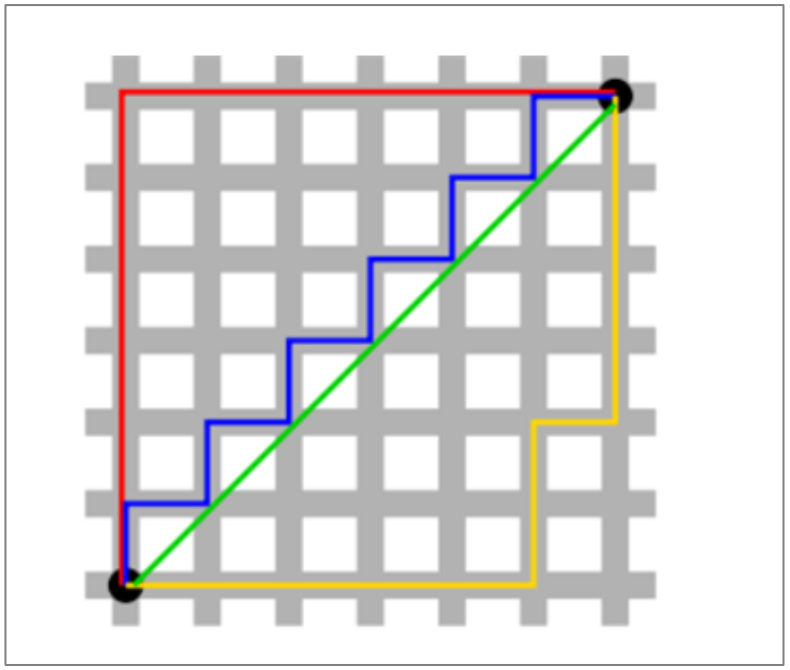
  정량적인 데이터, 정성적인 데이터 모두에서 잘 작동 해야한다.

- 클러스터 위상(Discovery of clusters with arbitrary shape)

  보통 '기하 거리(Euclidean)' 나 '맨하탄 거리(Manhattan)'를 바탕으로 클러스터를 결정한다. 이렇게 거리 측정법을 바탕으로 나누면 비슷한 크기와 밀도의 구형 집단을 나누기 쉽다. 따라서 원이나 구형이 아닌 다른 도형으로도 클러스터를 찾아줄 알고리즘이 필요하다.

10.1.2 Requirements for Cluster Analysis

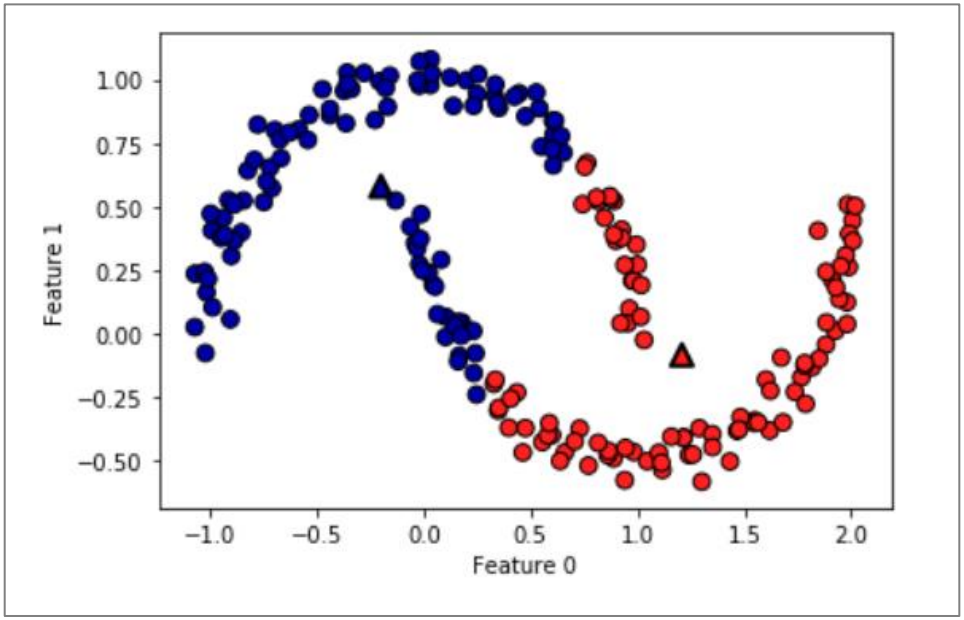
- ‘기하 거리(Euclidean)’ 와 ‘맨하탄 거리(Manhattan)’ 비교



–: 맨하탄 거리
–: 유클리디안 거리

10.1.2 Requirements for Cluster Analysis

- 클러스터 위상(Discovery of clusters with arbitrary shape)



10.1.2 Requirements for Cluster Analysis

- 입력 파라미터 결정에 필요한 대상 분야의 전문성(Requirements for domain knowledge to determine input parameters)

  클러스터링 알고리즘의 결과는 입력 파라미터에 좌우된다.

- 오차에 대한 안정성(Ability to deal with noisy data)

  실세계 데이터세트에는 아웃라이어(outlier), 결손 값(Missing), 알 수 없는 값, 오류 데이터 등이 존재할 수밖에 없다. 따라서 이러한 노이즈에 대해 안정적인 클러스터링 방법이 필요하다.

- 데이터 축적과 순서에 대한 통용성(Incremental clustering and insensitivity to input order)

  데이터 추가를 할 때마다 다시 계산하는 클러스터링이나 입력 데이터의 순서에 따라 다른 결과를 클러스터링은 좋지 못한 알고리즘이다. 따라서 데이터 축적에 대응할 수 있고, 입력 순서와 상관없이 결과를 보장하는 알고리즘이 필요하다.

10.1.2 Requirements for Cluster Analysis

- 고차원 데이터(Capability of clustering high-dimensionality data)

  클러스터링 알고리즘 대다수가 두 세 가지 속성 정도의 저차원 데이터 처리에는 좋은 성능을 띄지만, 고차원 공간의 데이터 오브젝트를 대상으로 클러스터를 분류하는 일은 무척 어렵다.

- 제약(Constraint-based clustering)

  실제 데이터에 클러스터링을 적용하려면 다양한 제약 조건이 있다.

- 가독성과 활용성(Interpretability and usability)

  클러스터링은 의미 해석과 적용에 중점을 두어야한다.

10.1.2 Requirements for Cluster Analysis

- 분할 기준(The partitioning criteria)

  데이터 오브젝트의 위계 구조 존재 여부

- 클러스터 분할(Separation of clusters)

  데이터 오브젝트를 상호 배타적 클러스터로 분할 여부(즉, 클러스터간 중복 여부)

10.1.2 Requirements for Cluster Analysis

- 유사성 측정법(Similarity measure)

  1. 거리, 밀도, 인접성에 따른 연결성으로 유사성 결정

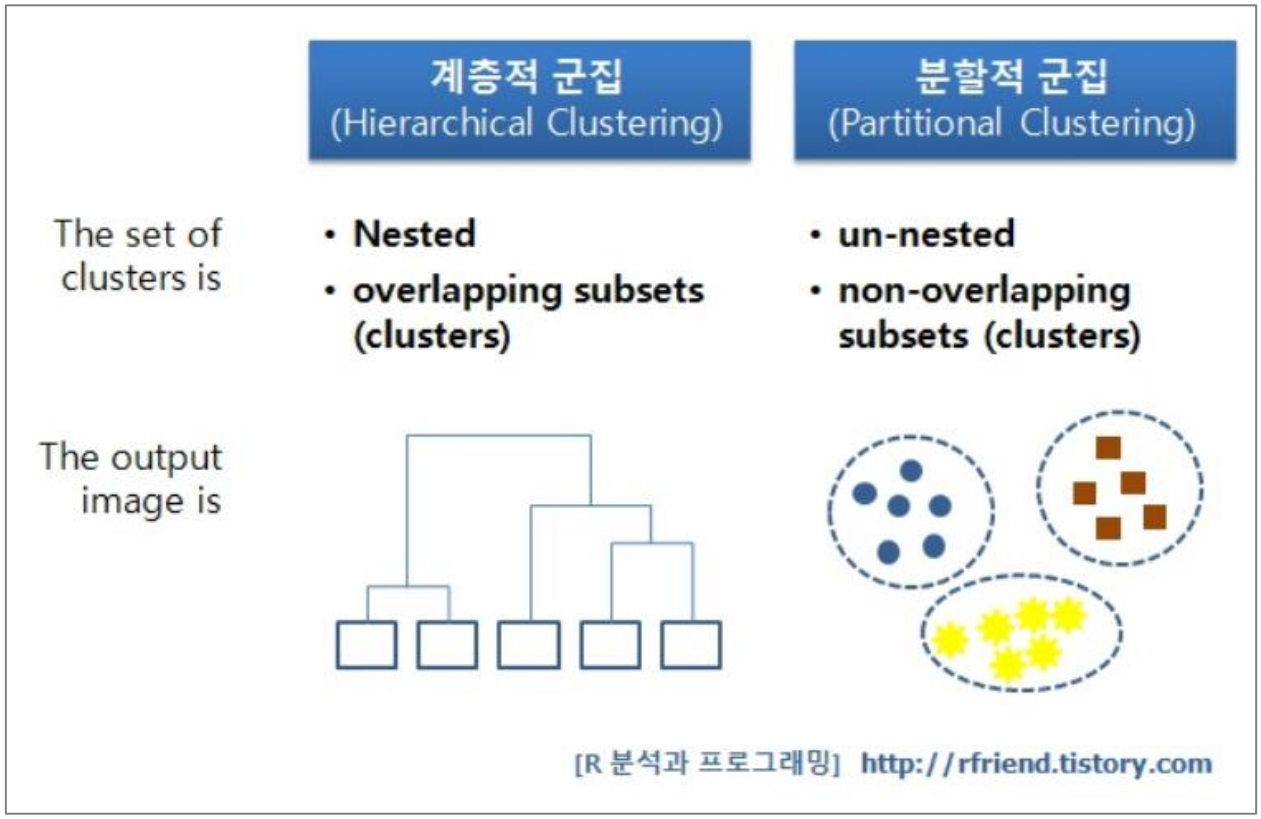
  2. 두 오브젝트 사이의 절대적인 거리 차이에 의존하지 않는 방법

- 클러스터링 공간(Clustering space)

  고차원 데이터인 경우에는 전체 차원 공간을 바탕으로 계산한 클러스터는
  노이즈 값이나 속성으로 좋지 않는 결과가 나올 수 있다. 따라서 특성을 선택하여
  낮은 차원의 공간에서 클러스터를 찾는 것이 더 좋은 결과를 가져온다.

10.1.3 Overview of Basic Clustering Methods

- 분할 클러스터링(Partitioning methods) 와 계층적 클러스터링(Hierarchical methods)



10.1.3 Overview of Basic Clustering Methods

- 분할 클러스터링(Partitioning methods)
  - n가지 데이터 오브젝트 세트를 k(k≤n)개의 파티션 부분집합(클러스터)으로 분할
  - 배타적 클러스터 분할 방식을 사용
  - 대부분의 분할 방법론은 거리를 중심으로 계산
  - 더 나은 분할을 만들어 내기 위해 반복적인 재배치 기법을 사용
  - 휴리스틱(heuristic) 기법을 사용

10.1.3 Overview of Basic Clustering Methods

- 계층적 클러스터링(Hierarchical methods)

  - 주어진 데이터 오브젝트를 구조적으로 분해해서 클러스터를 나누는 방식.

  - 단점: 한 단계의 계산이 끝나지 않으면 계산이 마무리 되지 않는다.

    이 클러스터링 도중에 잘못된 결정을 내렸을 경우 정정할 수 없다.

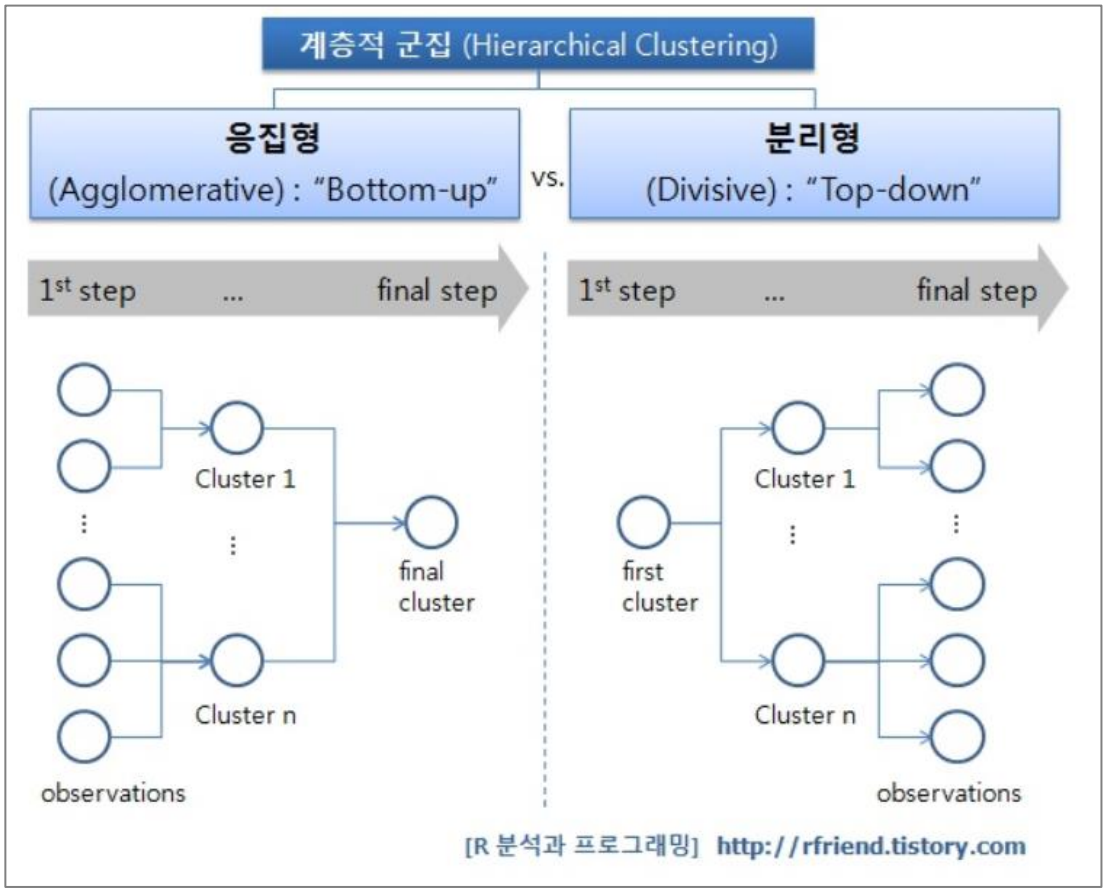
  - 장점: 이러한 경직성(첫번째 단점) 덕분에 가능한 조합 경우의 수를 고려할 필요가 없어 계산횟수를 줄여준다.

  - 거리, 밀도, 연속성 기반으로 나눌 수 있음.

  - 조적식(agglomerative)와 분할식(divisive)으로 구분.

10.1.3 Overview of Basic Clustering Methods

- 계층적 클러스터링(Hierarchical methods)



10.1.3 Overview of Basic Clustering Methods

- 조적식 방법(The agglomerative approach)

  - 상향식 기법(the bottom-up approach)

  1. 모든 오브젝트를 별개의 그룹으로 구성.
  2. 전체 그룹이 단 하나의 그룹을 구성할 때까지 차츰 인접한 오브젝트 혹은 그룹을 하나로 합침.

- 분할식 방법(The divisive approach)

  - 하향식 기법(the top-down approach)

  1. 모든 오브젝트를 하나의 클러스터로 구성.
  2. 모든 오브젝트가 하나씩 하나의 클러스터가 될 때까지, 하나의 클러스터를 작은 조각으로 쪼갬.

10.2 Partitioning Methods

- 분할 클러스터링

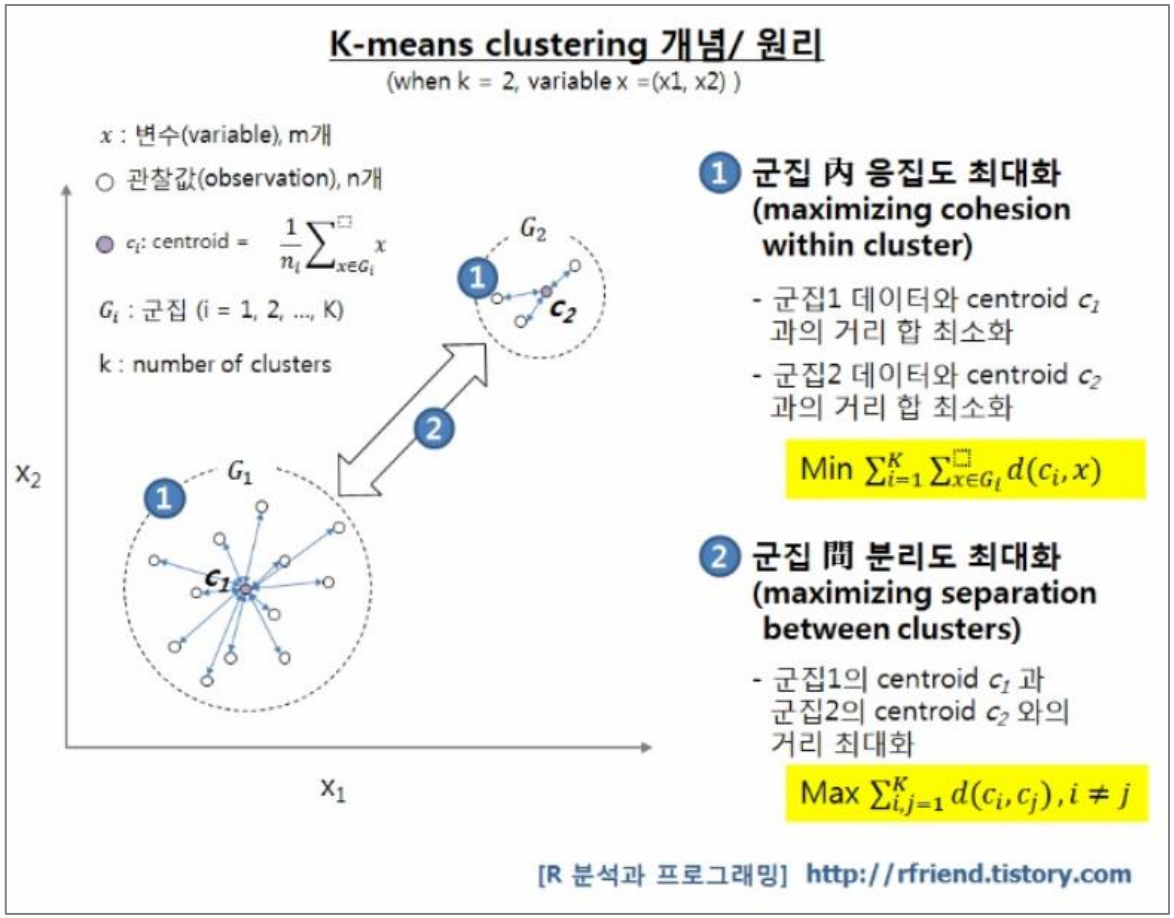
  - N개의 오브젝트로 구성된 입력 데이터세트 D에 대해 k개의 클러스터를 구성한다면, 분할 알고리즘은 오브젝트를 k개의 파티션($k \leq n$)으로 나눈다.

  - 이때 하나의 파티션이 하나의 클러스터에 해당한다.

  - 클러스터는 목적 함수 최적화를 통해 형성한다.

  - 종류: K-means, K-medoids

10.2.1 k-Means: A Centroid-Based Technique



10.2.1 k-Means: A Centroid-Based Technique

- 알고리즘: k-means. k-means 분할 알고리즘. 클러스터를 구성하는 오브젝트의 평균 값으로 클러스터의 중심을 나타낸다.
- 입력: (k: 클러스터의 개수, D: 데이터세트. N개의 오브젝트를 포함)
- 출력: k개의 클러스터 집합
- 방법:

  (1) 임의로 k개의 오브젝트를 선택해서 초기 클러스터의 중심으로 삼는다.

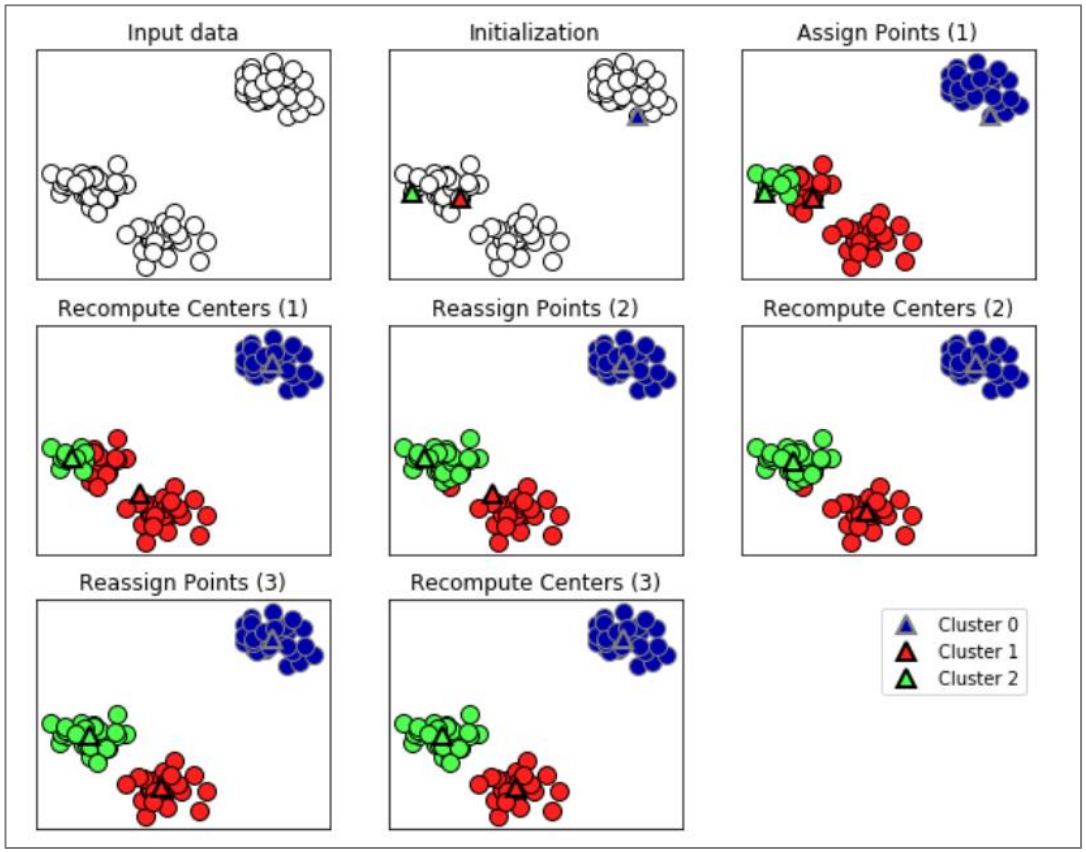
  (2) 모든 오브젝트를 각 클러스터의 중심과 비교해서 가장 유사한(가까운) 클러스터로 다시 배정한다.

  (3) 클러스터 중심을 다시 계산한다.
  즉, 클러스터에 포함된 모든 오브젝트의 평균 값을 계산한다.

  (4) 클러스터 배정에 변화가 없을 때까지 (2)~(3)을 반복.

10.2.1 k-Means: A Centroid-Based Technique

- 예시



10.2.2 k-Medoids: A Representative Object-Based Technique

- K-means는 평균 값을 구하는 연산을 수행하므로 잡음이나 이상치에 민감하다.

  이러한 단점을 해결하기 위해 나온 알고리즘이 k-medoids 알고리즘이다.

- K-medoids 알고리즘은 대표 값으로 오브젝트의 중심을 구하는 것이 아니라, 오브젝트 중에서 클러스터를 대표할 수 있는 가장 가까운 대표 값을 뽑는다.

- 대표로 뽑히지 않은 나머지 오브젝트는 가장 가까운 대표 오브젝트를 따라 해당 클러스터에 배정한다.

- K-medoids 식

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{p} \in C_i} dist(\boldsymbol{p}, \boldsymbol{o}_i),$$

10.2.2 k-Medoids: A Representative Object-Based Technique

- 알고리즘: k-medoids PAM 알고리즘. 중앙자 혹은 중앙 오브젝트를 바탕으로 클러스터를 분할한다.
- 입력: (k: 클러스터의 개수, D: n개의 오브젝트를 포함하고 있는 데이터세트)
- 출력: k 개의 클러스터 세트
- 방법:

  (1) D 안에서 임의의 k개 오브젝트를 선택, 초기 대표 오브젝트로 삼는다.

  (2) 대표가 아닌 나머지 오브젝트를 가장 가까운 대표의 클러스터로 배정한다.

  (3) 임의로 대표가 아닌 오브젝트 O_{random}을 선택한다.

  (4) 한 대표 오브젝트 O_j를 O_{random}과 교체했을 때 전체 비용 S를 계산한다.

  (5) if S < 0, O_j와 O_{random}을 교체해서 새로운 k 대표 오브젝트 세트를 구성한다.

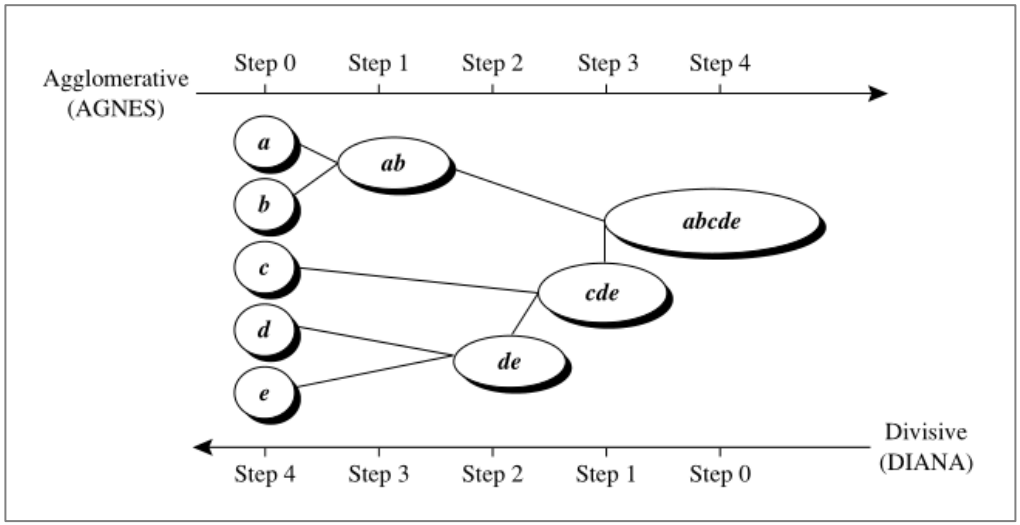
  (5) 클러스터 배정에 변화가 없을 때까지 (2)~(5)을 반복.

10.3 Hierarchical Methods

- 계층적 클러스터링은 데이터 오브젝트를 구조, 즉 '트리' 형태의 클러스터로 나눈다.
- '트리' 형태의 클러스터는 정리와 시각화에 편리하다.
- 종류: 조적식(agglomerative), 분할식(divisive)
- 클러스터링 품질을 높이는 방법은 다른 클러스터링 기법과 조합해서 '다중 단계 클러스터링'이다.
- 다중 단계 클러스터링의 종류: BIRCH기법, Chameleon 기법

10.3.1 Agglomerative versus Divisive Hierarchical Clustering

- 조적식 구조 클러스터링(Agglomerative) 알고리즘은 상향식이다.
- 분할식 구조 클러스터링(Divisive) 알고리즘은 하향식이다.



10.3.2 Distance Measures in Algorithmic Methods

- 가장 널리 쓰이는 4가지 클러스터 간 거리 측정법

Minimum distance: $$dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\} \quad (10.3)$$

Maximum distance: $$dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\} \quad (10.4)$$

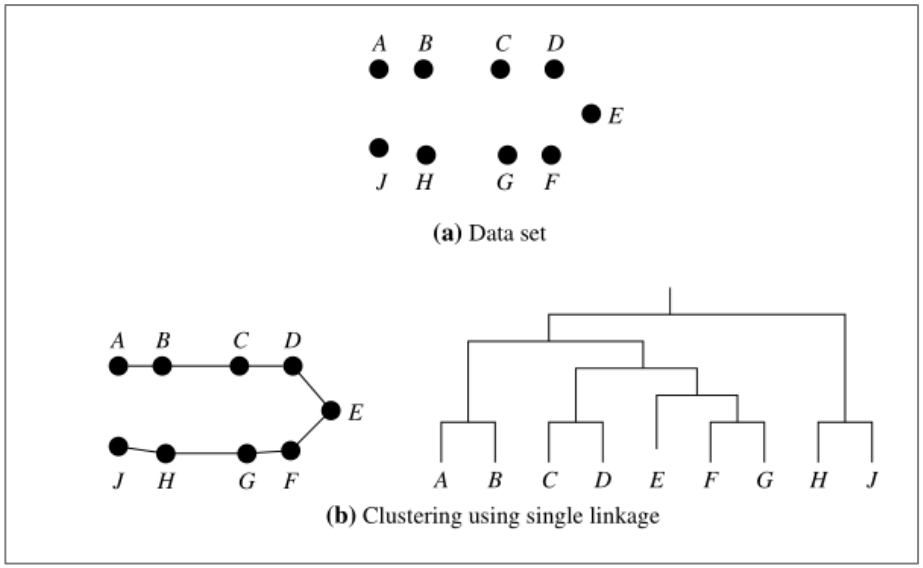
Mean distance: $$dist_{mean}(C_i, C_j) = |m_i - m_j| \quad (10.5)$$

Average distance: $$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'| \quad (10.6)$$

- $|p - p'|$: 두 오브젝트 p와 p′ 사이의 거리
- m_i: 클러스터의 평균값
- C_i: i번째 클러스터
- n_i: C_i 클러스터에 포함된 오브젝트의 개수

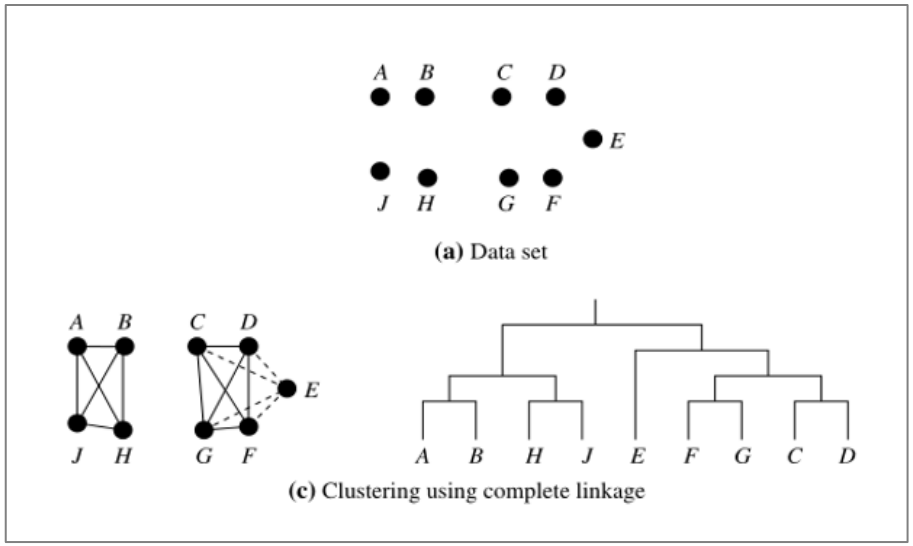
10.3.2 Distance Measures in Algorithmic Methods

- 알고리즘에서 '최소 거리'를 클러스터 간 거리 측정법으로 사용하는 알고리즘은 최인접 이웃 클러스터링 알고리즘(Nearest-neighbor clustering algorithm)이다.
- 이 알고리즘에서 가장 가까운 클러스터간 거리가 설정했던 Threshold값보다 클 때 클러스터링을 종료하면 '단연결 알고리즘(single-linkage)'이다.



10.3.2 Distance Measures in Algorithmic Methods

- 알고리즘에서 '최대 거리'를 클러스터 간 거리 측정법으로 사용하는 알고리즘은 가장 먼 이웃 클러스터링 알고리즘(Farthest-neighbor clustering algorithm)이다.
- 이 알고리즘에서 인접 클러스터 간의 최대 거리가 지정한 Threshold값보다 클 때 클러스터링을 종료하면 '전연결 알고리즘(complete-linkage)'이다.



클러스터 분석: 기본 개념과 방법론

감사합니다