머신러닝 - 군집화

2021

1. 군집화

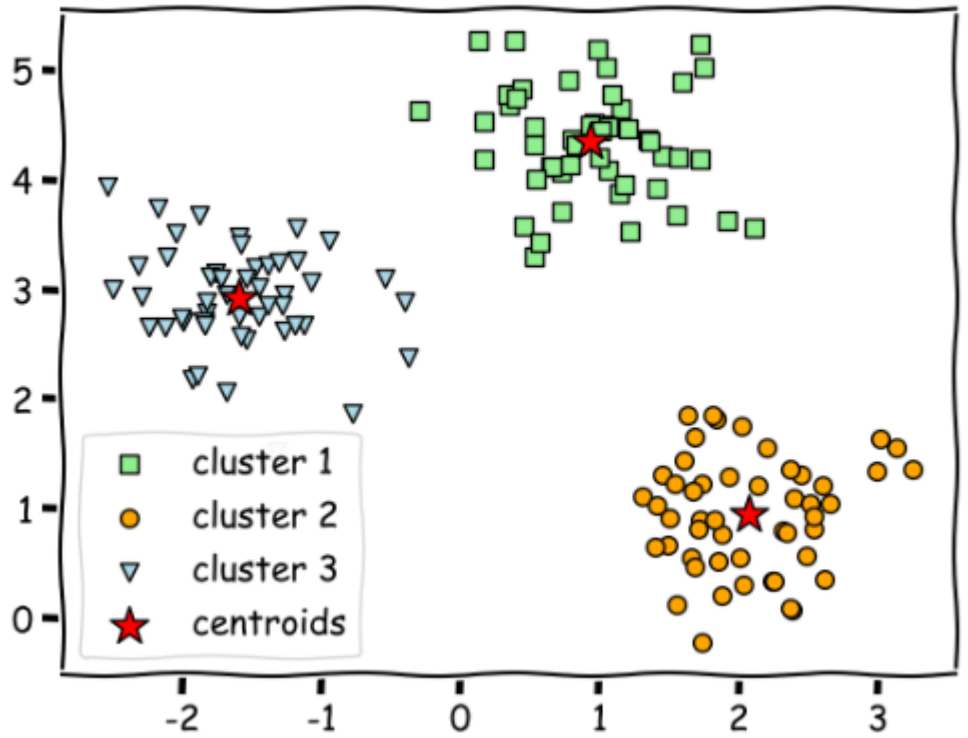
❖ 개념

- 비지도학습의 대표적인 기술
- X에 대한 레이블이 지정 되어있지 않은 데이터를 그룹핑하는 분석 알고리즘
- 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단의 대표할 수 있는 중심점을 찾는 것으로 데이터 마이닝의 한 방법
- 클러스터란 비슷한 특성을 가진 데이터들의 집단
- 데이터의 특성이 다르면 다른 클러스터에 속해야 함

2. 종류

❖ K 평균 (K Means)

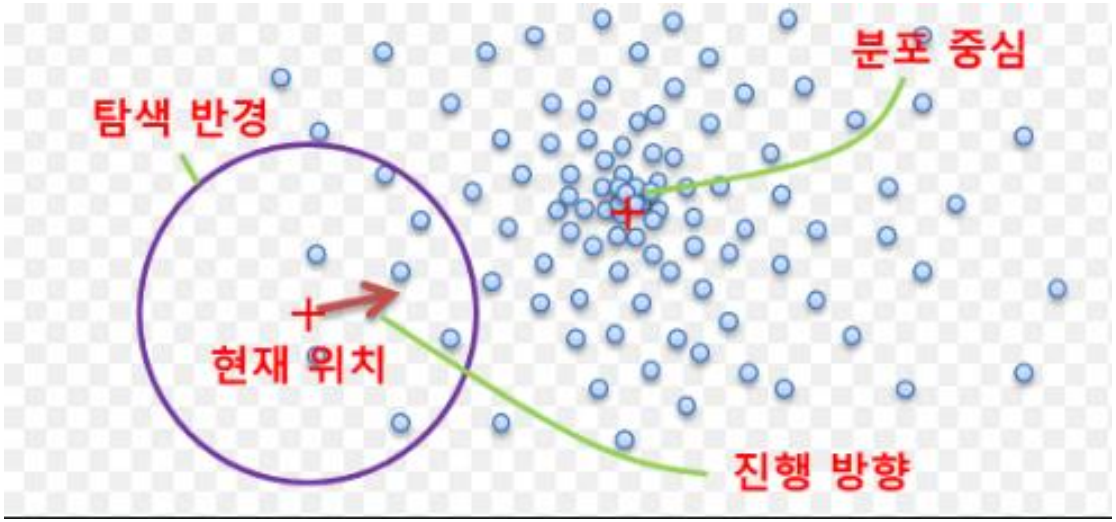
- 군집 중심점(centroid)이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법
- 선택된 포인트의 평균지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택,
- 다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행



2. 종류

❖ 평균이동(Mean Shift)

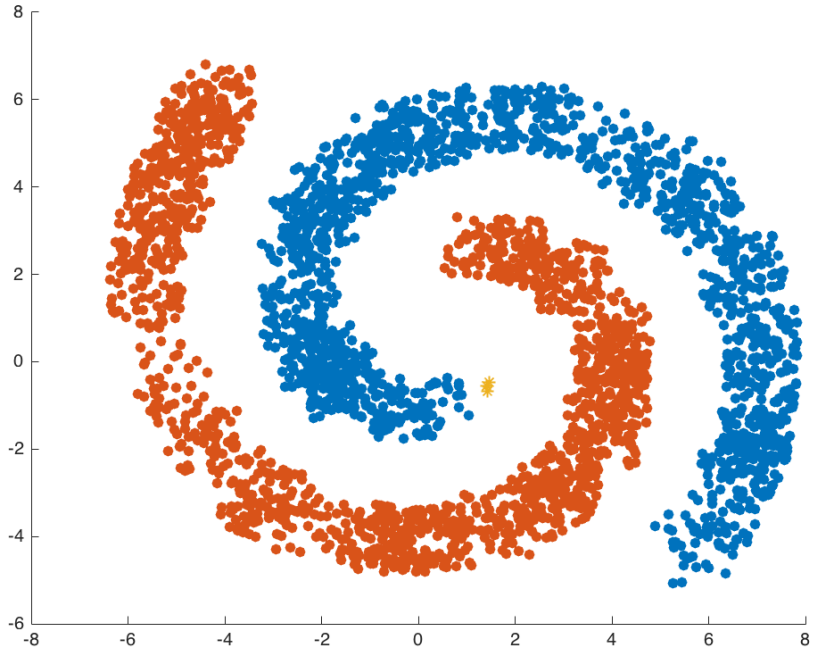
- 거리 중심이 아니라 데이터가 모여있는 **밀도가 가장 높은쪽**으로 군집 중심점을 이동하면서 군집화를
- 컴퓨터 비전 영역에서 이미지나 영상 데이터에서 특정 개체를 구분하거나 움직임을 추적하는데 뛰어난 역할을 수행하는 알고리즘
- 컴퓨터 비전 영역에서 잘 사용됨



2. 종류

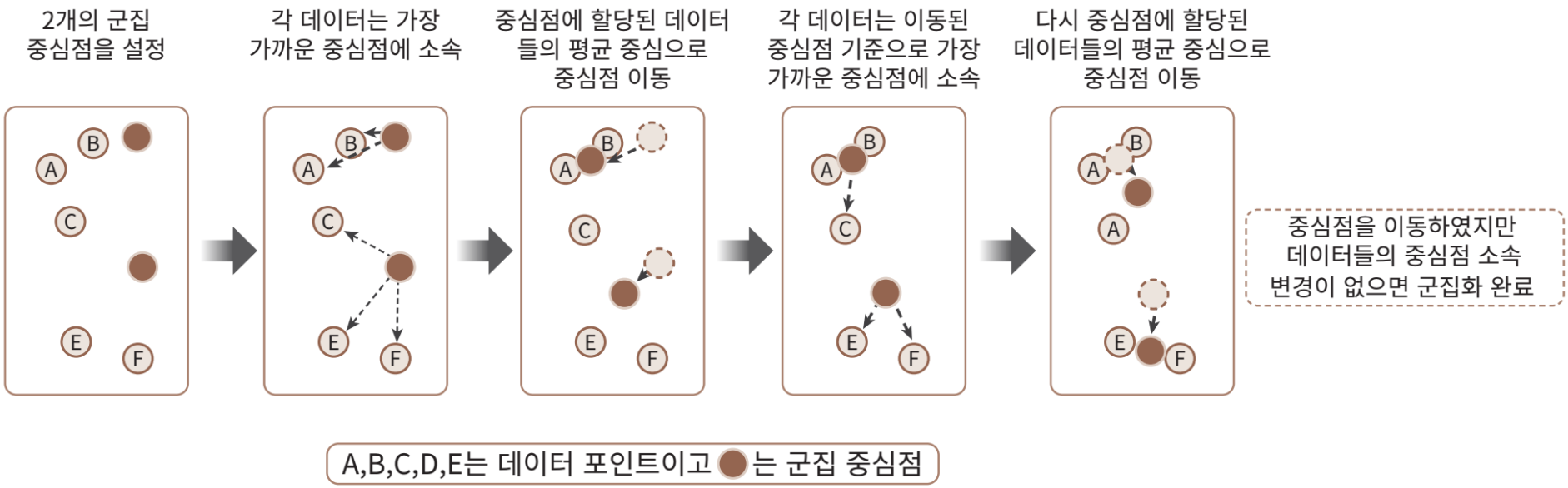
❖ DBSCAN(Density Based Spatial Clustering of Applications with Noise)

- 밀도 기반 군집화의 대표적 예
- 간단하고 직관적인 알고리즘으로 되어 있음에도 데이터의 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화가 가능
- 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행함



3. K-Means 알고리즘

❖ K 평균 (K Means)



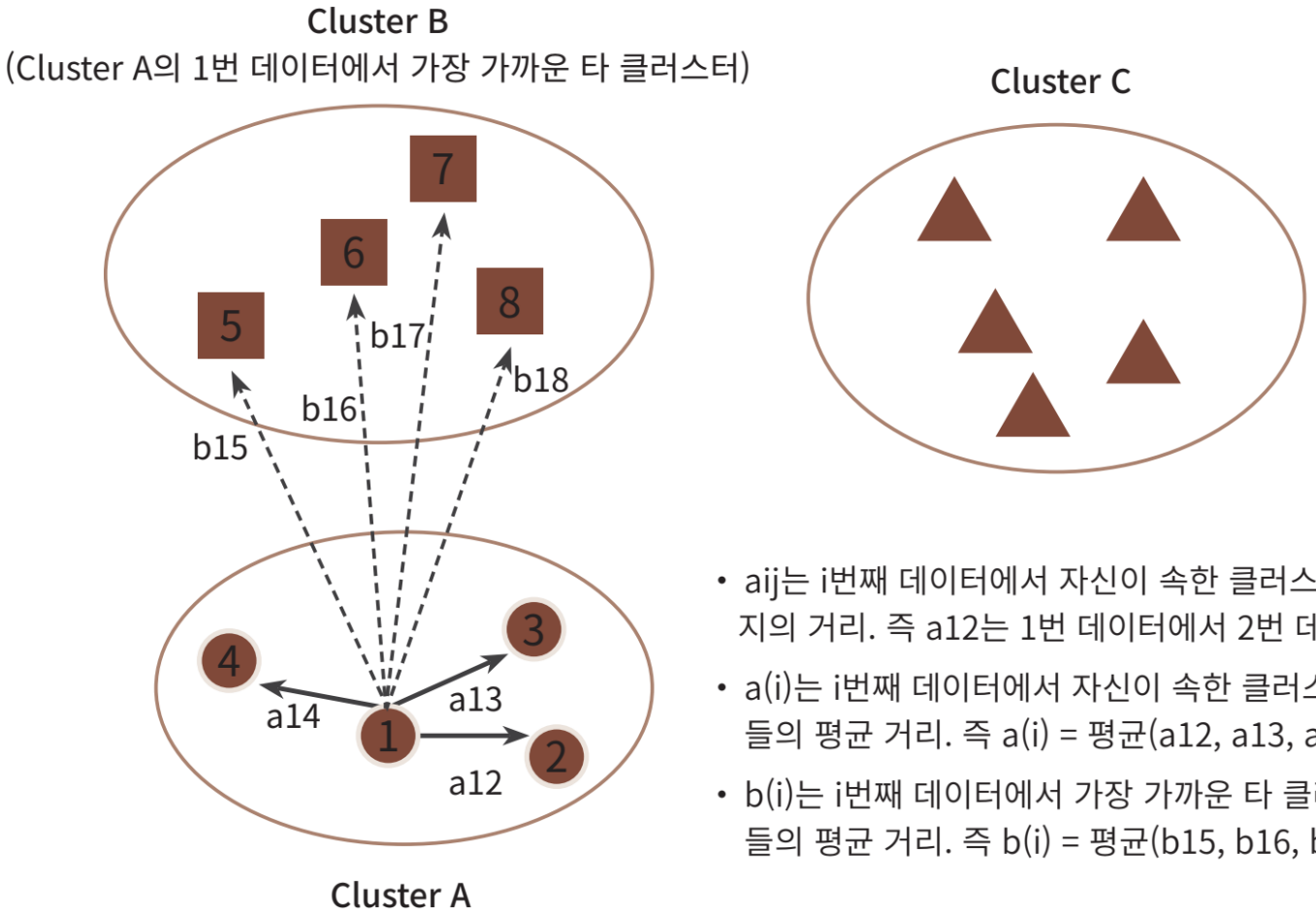
3. K-Means 알고리즘

❖ 사이킷런 K-Means 클래스

- n_clusters
- init : default 값은 'k-means++'
  1. 가지고 있는 데이터 포인트 중에서 무작위로 1개를 선택하여 그것을 첫번째 중심점으로 지정
  2. 나머지 데이터 포인트들에 대해 그 첫번째 중심점까지의 거리를 계산
  3. 두번째 중심점은 각 점들로부터 거리비례 확률에 따라 선택
     즉, 이미 지정된 중심점으로부터 최대한 먼 곳에 배치된 데이터포인트를 그 다음 중심점으로 지정
  4. 중심점이 k개가 될 때까지 2, 3번을 반복
- max_iter : default 값은 300

4. 군집 평가

❖ 실루엣 분석

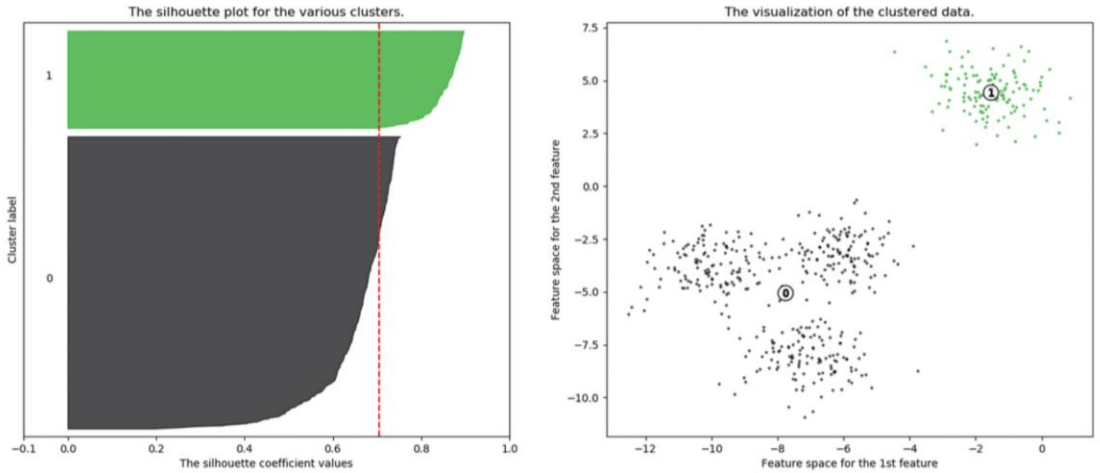


- aij는 i번째 데이터에서 자신이 속한 클러스터내의 다른 데이터 포인트까지의 거리. 즉 a12는 1번 데이터에서 2번 데이터까지의 거리
- a(i)는 i번째 데이터에서 자신이 속한 클러스터내의 다른 데이터 포인트들의 평균 거리. 즉 a(i) = 평균(a12, a13, a13)
- b(i)는 i번째 데이터에서 가장 가까운 타 클러스내의 다른 데이터 포인트들의 평균 거리. 즉 b(i) = 평균(b15, b16, b17, b18)

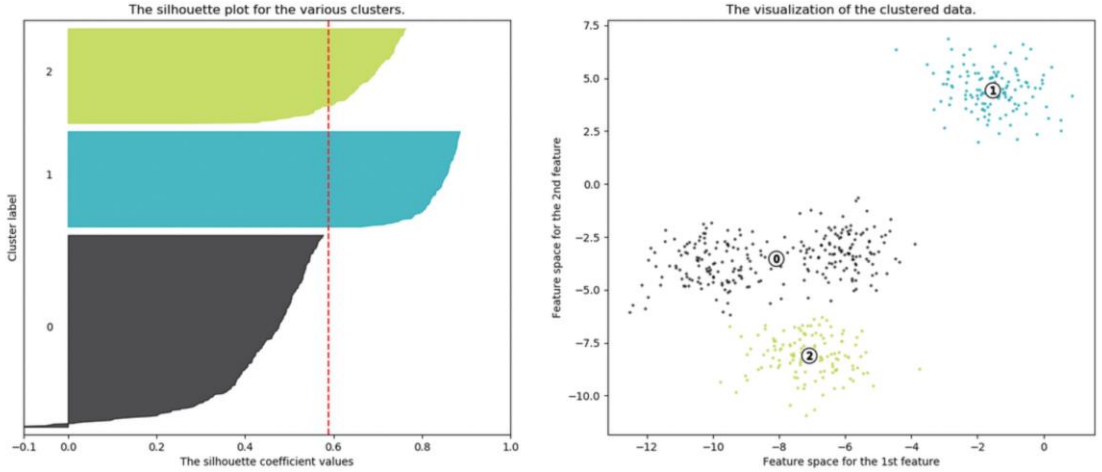
$$s(i) = \frac{(b(i) - a(i))}{(\max(a(i), b(i))}$$

4. 군집 평가

❖ 군집 2개 – 평균 실루엣 계수 0.704



❖ 군집 3개 – 평균 실루엣 계수 0.588



4. 군집 평가

❖ 군집 4개 – 평균 실루엣 계수 0.654