계층적 군집화 심층 분석 보고서

Hierarchical Clustering

1. 개요 및 핵심 철학

계층적 군집화(Hierarchical Clustering)는 데이터 간의 유사도를 바탕으로 마치 대진표나 가계도 같은 **'트리 구조(Dendrogram)'**를 형성하며 군집을 나누는 방법입니다. K-Means와 달리 처음에 군집의 개수(K)를 정할 필요가 없으며, 데이터가 어떻게 묶이는지 계층적인 구조를 시각적으로 확인할 수 있다는 큰 장점이 있습니다.

핵심 컨셉: "덴드로그램(Dendrogram)"

"숲을 보면서 나무를 자른다."
데이터들이 하나씩 합쳐지는 과정을 트리 형태로 기록합니다. 분석가는 이 트리의 높이(유사도 거리)를 보고 적절한 위치에서 잘라서(Cut) 원하는 개수의 군집을 얻어낼 수 있습니다.

2. 상세 작동 메커니즘

가장 많이 쓰이는 **병합(Agglomerative) 방식**은 'Bottom-up' 접근법을 따릅니다.

STEP 1. 개별 군집 시작

처음에는 모든 데이터 포인트 하나하나를 각각의 독립된 클러스터로 간주합니다. (N개의 클러스터)

STEP 2. 가장 가까운 군집 병합

거리 행렬(Distance Matrix)을 계산하여 가장 가까운 두 군집을 찾아 하나로 합칩니다. 이때 군집 간 거리를 재는 방법(Linkage)이 중요합니다.

STEP 3. 반복 및 트리 형성

모든 데이터가 하나의 거대한 군집이 될 때까지 이 과정을 반복하며, 병합 순서와 거리를 기록하여 덴드로그램을 완성합니다.

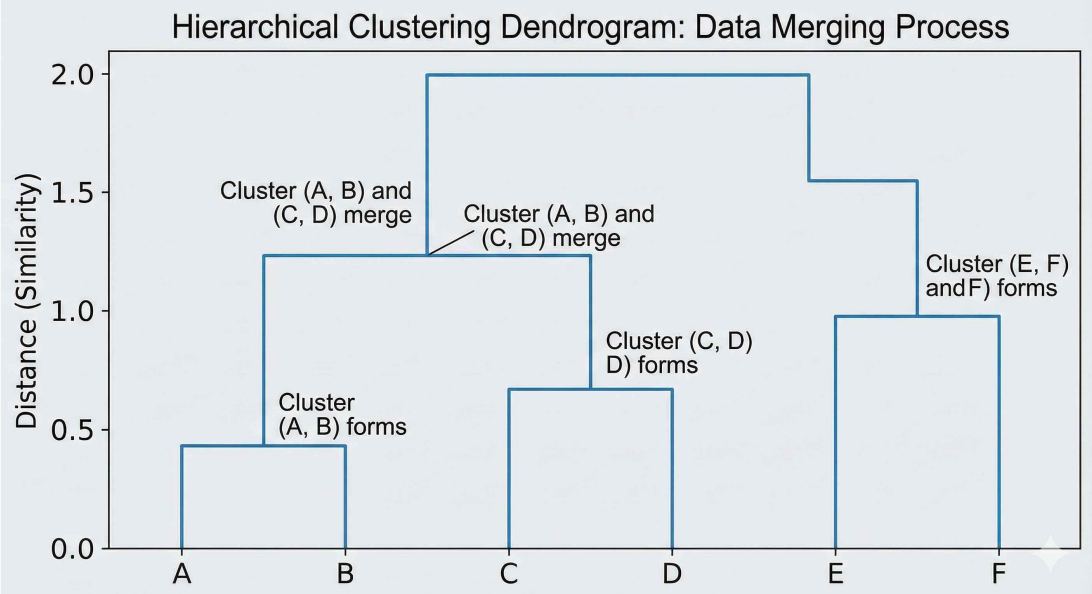

그림 1: 데이터 병합 과정을 보여주는 덴드로그램

⚖️ 3. K-Means vs 계층적 군집화

두 모델은 군집을 형성하는 접근 방식과 결과물의 형태가 완전히 다릅니다.

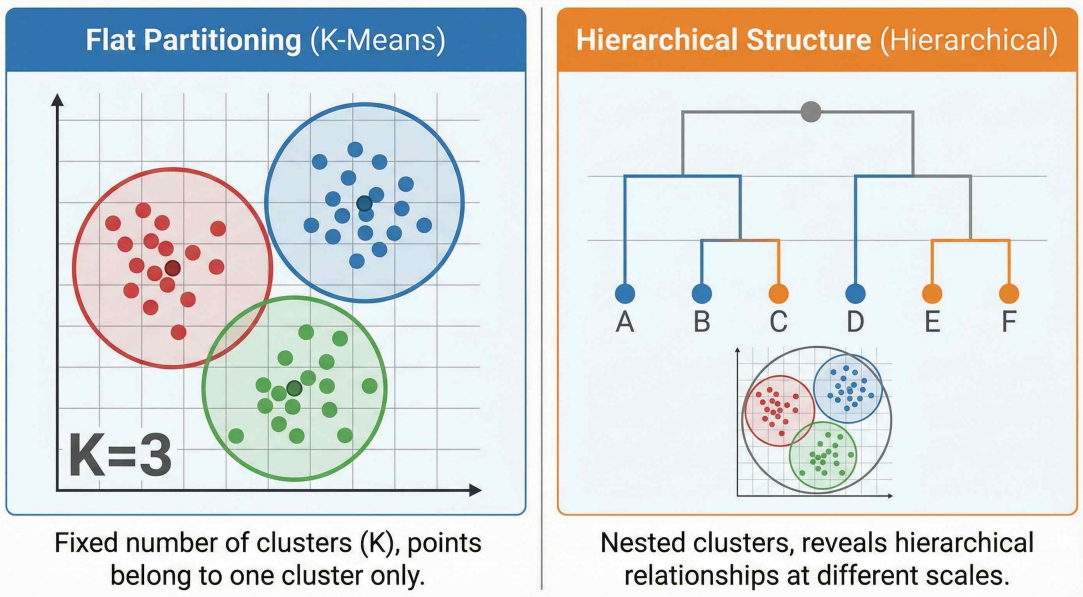

그림 2: 평면적 분할(K-Means) vs 계층적 구조(Hierarchical)

항목	K-Means Clustering	Hierarchical Clustering
사전 설정	**K(군집 수)를 미리 정해야 함**	미리 정할 필요 없음 (나중에 결정)
결과물	군집 레이블 (0, 1, 2...)	**덴드로그램 (트리 구조)**
계산 복잡도	빠름 ($O(N)$), 대용량 가능	**매우 느림** ($O(N^3)$ 또는 $O(N^2)$), 대용량 불가
재현성	초기값에 따라 결과 달라짐	**항상 동일한 결과** (Deterministic)

4. 장단점 심층 분석

장점 (Pros)

- ✓ **시각적 직관성:** 덴드로그램을 통해 데이터의 군집 구조를 한눈에 파악할 수 있어 설명력이 매우 좋습니다.
- ✓ **K 결정의 유연성:** 분석 후 트리를 어디서 자를지 결정하면 되므로, 최적의 K를 찾기 수월합니다.

단점 및 주의사항

- ✗ **치명적인 속도:** 모든 데이터 간의 거리를 계산해야 하므로, 데이터가 수천 개만 넘어가도 계산 시간이 기하급수적으로 늘어납니다.
- ✗ **노이즈 민감성:** 한 번 잘못 병합되면 되돌릴 수 없어서, 이상치나 노이즈 데이터에 결과가 크게 왜곡될 수 있습니다.

🏁 5. 결론 (Conclusion)

> 계층적 군집화는
> **"데이터의 족보(Family Tree)"**를
> 그려주는 분석 도구입니다.

최종 요약

- ✔ 데이터 양이 적고(Small Data), 군집 구조의 시각화가 중요할 때 사용합니다.
- ✔ 연결 방식(Linkage: Ward, Average, Complete) 선택이 결과에 큰 영향을 줍니다.
- ✔ 대용량 데이터라면 K-Means나 DBSCAN을 사용하는 것이 정신 건강에 좋습니다.