HIERARCHICAL CLUSTERING 핵심 원리 및 활용

거리 기반, 덴드로그램, 연결 기준

Hierarchical Clustering 개요

Hierarchical Clustering(계층적 군집화)는 데이터를 단계적으로 묶거나(병합형) 쪼개며(분할형) 군집 구조를 트리 형태(덴드로그램)로 표현하는 비지도 학습 기법입니다. 전체 병합/분할 과정을 먼저 만든 뒤, 원하는 높이에서 트리를 잘라 군집 개수를 정할 수 있어 해석성과 유연성이 뛰어납니다.

핵심 원리

계층적 군집화는 “군집 간 거리”를 정의하고, 그 거리가 가장 작은(또는 특정 기준이 최소인) 군집쌍을 반복적으로 병합/분할하여 덴드로그램을 구성합니다.

- 병합형(Agglomerative): 각 샘플을 1개 군집으로 시작해 가까운 군집끼리 순차적으로 합칩니다.
- 분할형(Divisive): 전체를 1개 군집으로 시작해 분할을 반복하며 내려갑니다.
- 연결 기준(Linkage): 군집 간 거리를 정하는 규칙(single/complete/average/ward 등).
- 거리 척도(Distance): 유클리드/맨해튼/코사인 등. 스케일 표준화 여부에 따라 결과가 달라집니다.
- 컷(Cut): 덴드로그램을 k개 군집 또는 거리 임계값 기준으로 잘라 최종 라벨을 얻습니다.

주요 장점

덴드로그램 기반 해석

병합/분할의 전 과정을 트리로 보여주기 때문에, 군집이 어떻게 합쳐졌는지 근거를 시각적으로 확인할 수 있습니다. 모델 설명이 필요한 분석에서 특히 유용합니다.

군집 수를 미리 고정하지 않음

K-means처럼 k를 먼저 정하지 않아도 됩니다. 덴드로그램을 만든 뒤 업무 목적에 맞는 높이에서 잘라 군집 수를 결정할 수 있습니다.

연결 기준 선택으로 형태 제어

single-linkage는 '사슬'처럼 길게 이어붙는 군집이 생기기 쉽고, complete/average는 더 균형 잡힌 군집을 만드는 경향이 있습니다. Ward는 군집 내 분산 증가가 최소가 되도록 병합해 실무에서 자주 선택됩니다.

다양한 도메인에 바로 적용

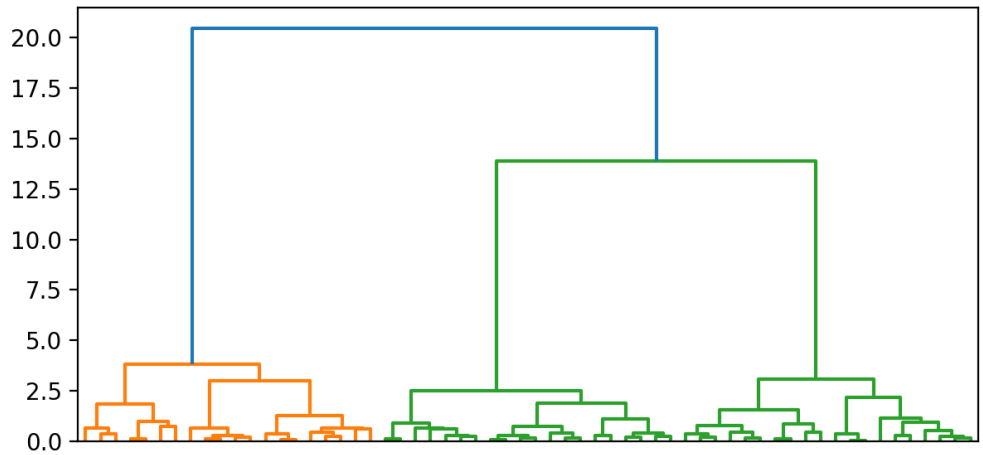
문서 임베딩, 고객 세그먼트, 유전자 발현, 이미지 특징 등 '거리/유사도'를 정의할 수 있는 거의 모든 데이터에 적용 가능합니다.

실무 팁: 전처리와 스케일링

계층적 군집화는 거리 기반이므로 표준화/정규화가 매우 중요합니다. 고차원 데이터에서는 PCA/UMAP 같은 차원 축소 후 군집화를 적용하면 노이즈를 줄이고 덴드로그램 해석이 쉬워질 수 있습니다.

덴드로그램 시각화

덴드로그램은 각 병합 단계의 거리(높이)를 보여줍니다. 큰 점프(급격한 높이 증가)가 나타나는 지점을 기준으로 자르면 군집 간 분리가 비교적 명확한 k를 선택할 수 있습니다.



설정/하이퍼파라미터 튜닝

계층적 군집화는 모델 파라미터가 많지 않지만, 거리 척도와 linkage 선택에 따라 결과가 크게 달라집니다. 실루엣 점수 등 내부 지표를 활용해 여러 조합을 비교하고, 업무 해석 관점에서 가장 일관된 군집 구조를 선택하는 것이 좋습니다.

설정	설명	일반적인 범위	영향
metric	거리 척도	euclidean manhattan cosine	유사도 정의가 바뀌면 군집 형태가 크게 변화
linkage	군집 간 거리 정의	ward / average / complete / single	응집/분리 성향과 chaining 여부에 영향
n_clusters	최종 군집 개수	2-20 (문제에 따라)	작을수록 거친 분류, 클수록 세분화
distance_threshold	병합 거리 임계값	데이터 스케일 의존	작을수록 많은 군집, 클수록 적은 군집
preprocessing	스케일링/차원축소	standardize, PCA/UMAP	노이즈 감소, 거리 왜곡 완화
criterion	컷 기준	k 또는 거리	해석 용이성과 재현성에 영향

튜닝/선택 방법으로는 (1) linkage·metric 조합 비교, (2) 실루엣/칼린스키-하라바즈 같은 내부 지표 평가, (3) 덴드로그램의 큰 거리 점프 구간을 활용한 컷 기준 설정 등이 있습니다.

결론

Hierarchical Clustering은 덴드로그램을 통해 군집 구조를 해석 가능한 형태로 제공하며, 군집 수를 사후에 유연하게 결정할 수 있는 강력한 비지도 학습 방법입니다. 거리 척도와 linkage 선택, 그리고 스케일링/차원축소 같은 전처리를 적절히 적용하면 다양한 도메인에서 신뢰도 높은 군집 결과를 얻을 수 있습니다.

참고 문헌 (추천)

- Ward Jr., J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. JASA.
- Johnson, S. C. (1967). Hierarchical Clustering Schemes. Psychometrika.
- Sibson, R. (1973). SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method. The Computer Journal.
- Murtagh, F., & Contreras, P. (2011). Methods of Hierarchical Clustering. arXiv:1105.0121.
- M llner, D. (2011). Modern Hierarchical, Agglomerative Clustering Algorithms. arXiv:1109.2378.
- M llner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. JSS.
- Eisen, M. B., et al. (1998). Cluster analysis and display of genome-wide expression patterns. PNAS.
- Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. JCGS.