Clustering for Disease Subtypes

Discover hidden patient subgroups with distinct characteristics

K-means

Fast, scalable clustering with predefined number of clusters

Cancer subtypes from gene expression

Hierarchical Clustering

Dendrogram-based, no need to specify K upfront

Patient stratification visualization

DBSCAN

Density-based, finds arbitrary shapes, handles outliers

Anomaly detection in clinical data

Consensus Clustering

Robust clustering through multiple runs and voting

Stable disease subtype identification

1. K-means Clustering

K-means is a partitioning method that divides data into K distinct, non-overlapping clusters. Each data point belongs to the cluster with the nearest mean (centroid). It's one of the most popular clustering algorithms due to its simplicity and efficiency.

Algorithm Steps:

1. Initialize K centroids randomly
2. Assign each point to nearest centroid
3. Recalculate centroids as cluster means
4. Repeat steps 2-3 until convergence

Key Characteristics:

- Requires specifying K (number of clusters) in advance
- Uses Euclidean distance for similarity
- Assumes spherical, equal-sized clusters
- Time complexity: O(n·K·i·d) where i = iterations, d = dimensions

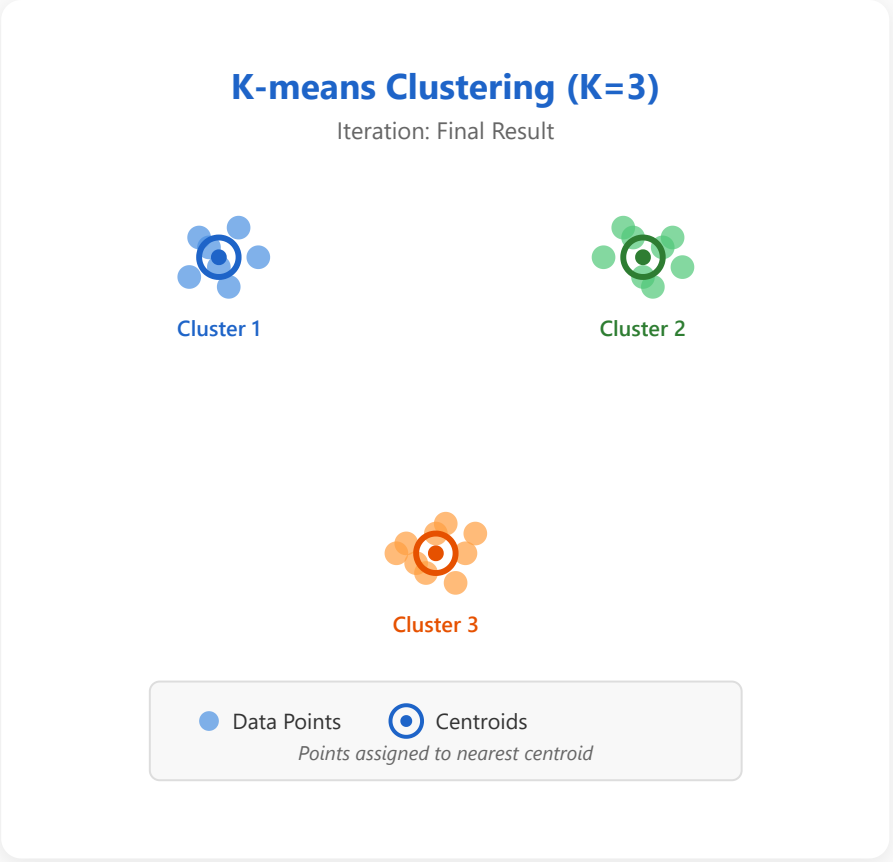
✓ Advantages

- Fast and scalable
- Easy to implement
- Works well with large datasets

✗ Limitations

- Sensitive to initialization
- Struggles with non-spherical shapes
- Affected by outliers

Clinical Example:



Identifying breast cancer subtypes (Luminal A, Luminal B, HER2+, Basal-like) based on gene expression profiles of ER, PR, and HER2 markers.

2. Hierarchical Clustering

Hierarchical clustering builds a tree-like structure (dendrogram) showing relationships between data points. It can be agglomerative (bottom-up) or divisive (top-down), with agglomerative being more common in biomedical applications.

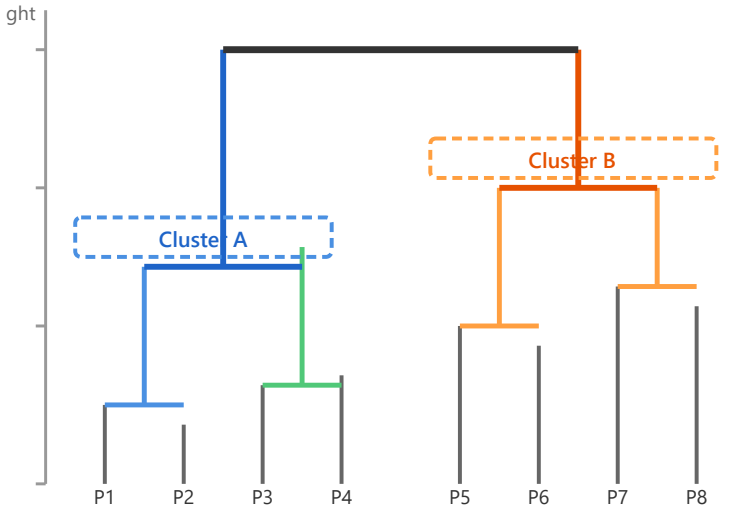
Algorithm Steps (Agglomerative):

1. Start with each point as its own cluster
2. Find the two most similar clusters
3. Merge them into a single cluster
4. Repeat until all points are in one cluster

Key Characteristics:

- No need to specify K in advance
- Creates hierarchical structure (dendrogram)
- Multiple linkage methods: single, complete, average, Ward

Hierarchical Clustering Dendrogram



Cut height determines number of clusters

- Time complexity: $O(n^3)$ or $O(n^2 \log n)$ with optimizations

Linkage Methods:

- **Single:** Minimum distance between clusters
- **Complete:** Maximum distance between clusters
- **Average:** Average distance between all pairs
- **Ward:** Minimizes within-cluster variance

✓ Advantages

- No pre-specified K needed
- Dendrogram shows relationships
- Deterministic results

✗ Limitations

- Computationally expensive
- Doesn't scale to large datasets
- Sensitive to noise/outliers

Clinical Example:

Patient stratification based on multiple clinical variables (age, lab values, symptoms) to visualize patient similarity and identify natural groupings for personalized treatment.

3. DBSCAN (Density-Based Spatial Clustering)

DBSCAN identifies clusters based on density - regions where points are closely packed together. Unlike K-means, it can find arbitrarily-shaped clusters and automatically identifies outliers as noise points.

Algorithm Steps:

1. Pick a random unvisited point
2. Find all points within ε distance (neighbors)
3. If neighbors ≥ minPts, start a new cluster
4. Recursively add density-reachable points
5. Mark isolated points as noise/outliers

Key Parameters:

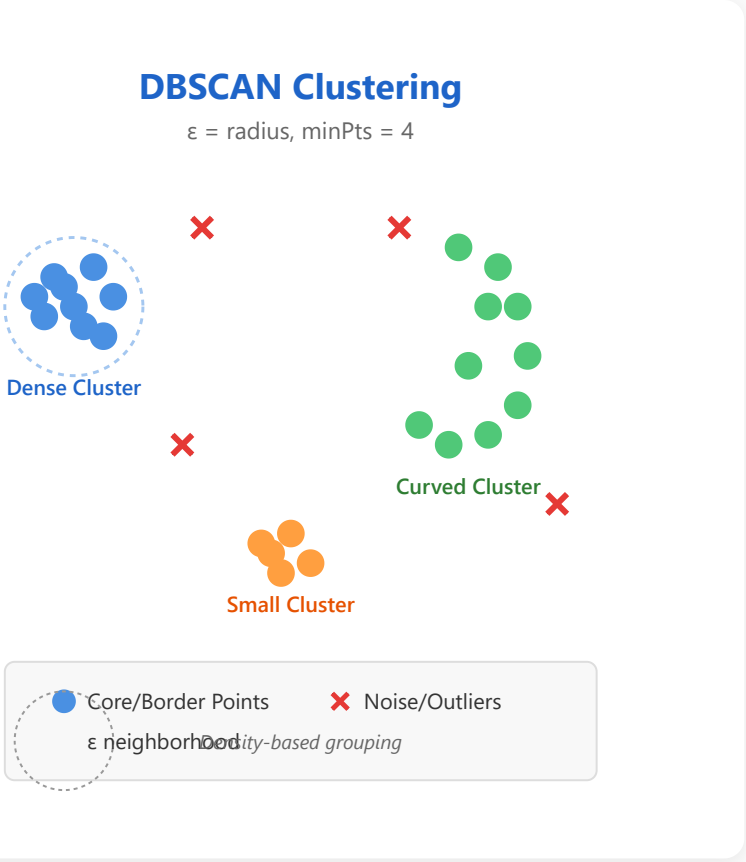
- **ε (epsilon):** Maximum distance for neighborhood
- **minPts:** Minimum points to form dense region
- **Core point:** Has ≥ minPts within ε distance
- **Border point:** In neighborhood of core point
- **Noise point:** Neither core nor border

✓ Advantages

- Finds arbitrary-shaped clusters
- Identifies outliers/noise
- No need to specify K
- Robust to outliers

✗ Limitations

- Struggles with varying densities
- Sensitive to ε and minPts
- High-dimensional challenges



Clinical Example:

Detecting unusual patient clusters in electronic health records that may represent rare disease phenotypes or adverse drug reactions, while identifying outlier cases that need special attention.

4. Consensus Clustering

Consensus clustering improves clustering stability and reliability by running multiple clustering iterations with resampled data, then combining results through voting. It provides confidence measures for cluster assignments.

Algorithm Steps:

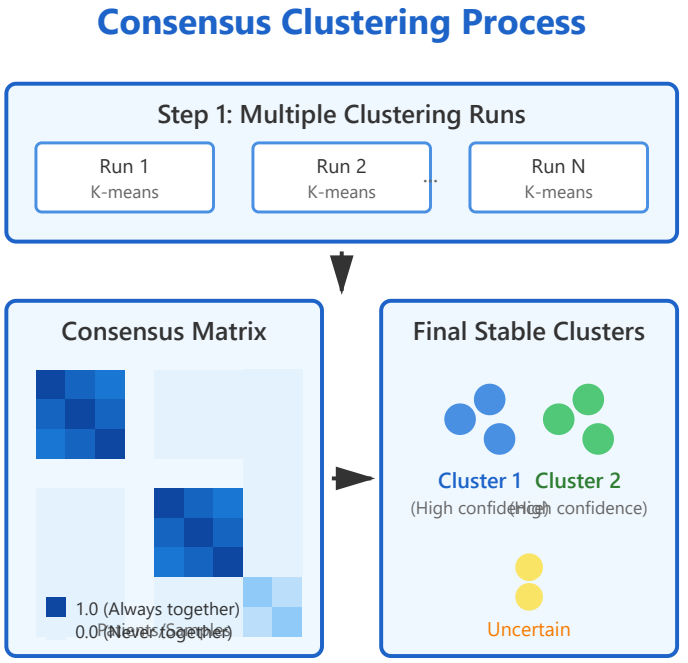
1. Resample data multiple times (e.g., 1000 runs)
2. Apply base clustering algorithm (e.g., K-means) to each sample
3. Record co-clustering frequency for all pairs
4. Build consensus matrix from frequencies
5. Apply final clustering to consensus matrix

Key Characteristics:

- Meta-clustering approach (combines multiple runs)
- Provides stability assessment via consensus matrix
- Helps determine optimal number of clusters
- Reduces sensitivity to initialization and sampling
- Can use any base clustering algorithm

Consensus Matrix:

- Entry (i,j) = proportion of runs where i and j are in same cluster
- Values near 1: strong evidence of co-clustering
- Values near 0: strong evidence of separation



Stability across multiple runs increases confidence

Typical: 100-1000 iterations with 80% resampling

- Intermediate values: uncertain assignments

✓ Advantages

- More robust and stable results
- Provides confidence measures
- Helps determine optimal K
- Reduces algorithm bias

✗ Limitations

- Computationally intensive
- Requires many iterations
- More complex to implement

Clinical Example:

Identifying stable molecular subtypes in heterogeneous cancers (e.g., glioblastoma) where robust classification is critical for treatment decisions and prognosis prediction.

Method Comparison Summary

Method	Pre-specify K?	Scalability	Cluster Shape	Best For
K-means	✓ Yes	★★★ High	Spherical	Large datasets, well-separated groups
Hierarchical	✗ No	★ Low	Any	Visualization, small-medium datasets

Method	Pre-specify K?	Scalability	Cluster Shape	Best For
DBSCAN	✗ No	⭐⭐ Medium	Arbitrary	Outlier detection, irregular shapes
Consensus	✓ Yes	⭐ Low	Depends on base	Stability, critical decisions