Hierarchical Clustering

Your Name

October 24, 2024

1 Hierarchical Clustering - Part 1

1.1 Main Concepts

1. **Introduction to Hierarchical Clustering**: Hierarchical clustering creates a hierarchy of clusters, as opposed to flat clustering, which returns an unstructured set of clusters. This clustering method is valuable because it does not require specifying the number of clusters in advance and produces a more informative, structured output. There are two main types of hierarchical clustering: **agglomerative** (bottom-up) and **divisive** (top-down).

2. **Agglomerative Hierarchical Clustering (HAC)**:

   - In **agglomerative clustering**, each document is initially treated as its own cluster. Clusters are then iteratively merged based on similarity until all documents belong to one large cluster. This is a bottom-up approach.
   - A **dendrogram** is used to visualize this hierarchy, where each horizontal line represents the merging of clusters, and the height of the line reflects the similarity between the clusters merged. This allows us to trace back the merging process, from individual documents to larger clusters.

3. **Monotonicity in Hierarchical Clustering**: An essential assumption in hierarchical clustering is **monotonicity**, meaning that as clusters merge, the combination similarities should decrease or remain the same. A violation of this principle leads to an **inversion**, which indicates that the clustering may not reflect the best possible merges at every step.

4. **Comparison with Flat Clustering**: Hierarchical clustering is often used when flat clustering fails to provide enough structure, or when a prespecified number of clusters is undesirable. While hierarchical clustering produces more meaningful clusters, it is computationally more expensive compared to flat clustering methods like K-means.

1.2 Deep Explanation of Key Topics

1. **Hierarchical Agglomerative Clustering (HAC)**:

   - **Agglomerative** methods start with each document as a separate cluster and merge clusters step by step based on their similarity. This is contrasted with **divisive** methods, where all documents start in one cluster and are recursively split.
   - **Dendrogram**: A tree diagram that represents the order in which clusters are merged. The height of each branch reflects the level of similarity at which the clusters are joined. The tree can be "cut" at a chosen height to yield a desired number of clusters.
   - **Monotonicity**: As clusters merge, the similarity (or proximity) between clusters must either decrease or stay the same. This ensures that each step represents the most logical grouping based on the data.

2. **Similarity Measures in HAC**: The clustering process depends on how similarity is measured between clusters. HAC typically uses one of the following approaches:

   - **Single-link** clustering merges clusters based on the closest pair of elements from each cluster.
   - **Complete-link** clustering merges clusters based on the most distant pair of elements.
   - **Group-average** clustering averages similarities across all elements.
   - **Centroid similarity** measures similarity between the centers (centroids) of clusters.

3. **Efficiency and Complexity**: Hierarchical clustering, while powerful, is computationally expensive. Its complexity is at least quadratic ($O(N^2)$) in terms of the number of documents, which can make it impractical for very large datasets. Flat clustering algorithms like K-means, which have linear complexity ($O(N)$), are more efficient but lack the structure and flexibility of hierarchical clustering.

   This first part introduces the foundational concepts of hierarchical clustering, focusing primarily on agglomerative methods and their visualization through dendrograms. The next part will explore different hierarchical clustering methods, such as single-link, complete-link, and group-average clustering, as well as their optimality and computational challenges.

2 Hierarchical Clustering - Part 2

2.1 Main Concepts

(a) **Different Methods in Hierarchical Clustering**:

   - **Single-link clustering (Minimum linkage)**: This method merges two clusters based on the closest distance between any pair of points in the clusters. It tends to form elongated, chain-like clusters.

- **Complete-link clustering (Maximum linkage)**: Merges clusters based on the largest distance between any pair of points in the clusters. It produces compact clusters with tighter boundaries.
- **Group-average clustering (Average linkage)**: Uses the average similarity between all pairs of elements from the two clusters. This method strikes a balance between single-link and complete-link approaches.
- **Centroid-based clustering**: Measures the distance between the centroids (average points) of clusters. Merging is based on the centroid similarity.

(b) **Challenges in Hierarchical Clustering**:

- **Optimality**: Since hierarchical clustering is greedy, it makes local decisions at each step without considering the global clustering structure. This can result in suboptimal clusters, especially with single-link and complete-link methods.
- **Inversions in Dendrograms**: Inversions occur when clusters at higher levels in the dendrogram appear to be more similar than clusters at lower levels, violating monotonicity. This can create confusion in interpreting cluster quality.

(c) **Evaluating Clusters**:

- **Cophenetic Correlation Coefficient**: Measures how well a dendrogram preserves the pairwise distances between the original data points. A higher value indicates that the clustering better reflects the original data structure.
- **Silhouette Score**: Helps assess how well-defined clusters are by considering the similarity of points within clusters compared to points in other clusters. It ranges from -1 to 1, where a higher score indicates better-defined clusters.
- **Internal Validity Measures**: These assess the cohesiveness of clusters using internal criteria like compactness and separation between clusters.

(d) **Non-monotonicity and Fixes**:

- **Non-monotonicity** occurs when clusters at higher levels in the hierarchy are formed with higher similarity values than clusters at lower levels. This suggests that some cluster merges may not follow the best possible path in terms of similarity structure.
- **Fixing Inversions**: Various methods, such as constrained clustering, can help avoid or minimize inversions. These methods enforce additional constraints, ensuring that the hierarchical clustering process adheres to the monotonicity principle more strictly.

2.2 Deep Explanation of Key Topics

(a) **Single-link vs. Complete-link Clustering**:

- **Single-link clustering** merges clusters by the smallest distance between two points (nearest neighbor). It is sensitive to outliers and can form long, chain-like clusters, leading to less compact clusters. While this can capture some types of structures, such as connected components, it often suffers from the **"chaining effect"**.
- **Complete-link clustering**, on the other hand, merges clusters based on the farthest points (furthest neighbor). This method is more robust against outliers and tends to form compact, well-separated clusters. However, it can be sensitive to the shapes of the data and may fail when clusters are not spherical.

(b) **Group-average Clustering**:

- Group-average clustering strikes a balance between single-link and complete-link methods by using the average similarity between all points in two clusters. This makes it less sensitive to outliers than single-link and more adaptable to clusters with irregular shapes than complete-link.
- The group-average approach is often used in practice because it avoids the extremes of chaining and over-compactness, providing a middle ground that works well in various situations.

(c) **Centroid-based Clustering**:

- In this method, clusters are merged based on the distance between their centroids (the average position of all points in a cluster). It focuses on the overall center of the clusters, making it particularly useful for spherical data distributions.
- However, centroid-based clustering can lead to non-intuitive merges when clusters are not convex or when they have unequal sizes and densities, making it crucial to evaluate the nature of the data when using this method.

(d) **Evaluating Clusters**:

- The **cophenetic correlation coefficient** measures the degree to which the clustering structure is maintained in the dendrogram compared to the original distance matrix. A higher coefficient indicates better preservation of the distance relationships, making it a useful metric for evaluating the quality of hierarchical clusters.
- The **silhouette score** is a popular internal validation method for evaluating clustering quality. It measures how similar an object is to its own cluster compared to other clusters. A silhouette score closer to +1 indicates that the object is well-clustered, while a score near 0 suggests that the object is on the border of two clusters, and a negative score indicates incorrect clustering.
- Internal validity measures focus on the compactness and separation of clusters, allowing for objective evaluation and comparison of different clustering solutions.