Hierarchical Clustering in Dynamic Ecosystem

You

August 27, 2024

1 Hierarchical Clustering Overview

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It can be categorized into two main approaches: agglomerative and divisive hierarchical clustering.

1.1 Agglomerative Hierarchical Clustering

Agglomerative clustering, also known as bottom-up clustering, begins with each data point as its own cluster. The algorithm then merges the closest clusters iteratively until all data points belong to a single cluster or a stopping criterion is met. This process is typically implemented using a distance matrix and linkage criteria to determine which clusters to merge.

1.2 Divisive Hierarchical Clustering

Divisive clustering, or top-down clustering, starts with all data points in a single cluster. The algorithm then recursively splits the cluster into smaller clusters until each data point is in its own cluster or a stopping criterion is reached. This approach is less commonly used due to its computational complexity, especially for large datasets.

1.3 Differences Between Agglomerative and Divisive Clustering

The primary difference between agglomerative and divisive hierarchical clustering is in their approach to cluster formation. Agglomerative clustering builds clusters from the bottom-up, starting with individual points and merging them, while divisive clustering starts with a single cluster and recursively splits it. Agglomerative methods are generally more computationally feasible and are more widely used in practice compared to divisive methods.

2 Code

The code involves implementing Agglomerative Clustering from scratch in Python.

The dataset is first preprocessed by dropping unnecessary columns and onhot encoding categorical columns. Then the proximity matrix is calculated and a subset of it is taken to cluster and plot the dendrogram. The dendrogram was useful in visualizing how changes in the features of the animals affected their clustering.

It was observed that there were minor changes in the clustering that occurred as the time step increased. To better visualize this, an interactive tool was created using streamlit that allowed a user to observe the dendrogram at different time steps, add a new animal by inputting its features and also mutating an existing animal's features by its speciesID.

3 Checkpoint Questions

3.1 Explain the difference between agglomerative and divisive hierarchical clustering.

Agglomerative hierarchical clustering builds clusters from individual points, merging the closest clusters iteratively. Divisive hierarchical clustering starts with a single cluster and recursively splits it into smaller clusters. Agglomerative clustering is more commonly used due to its practicality in large datasets.

3.2 How does the choice of linkage criteria affect the formation of clusters? Provide examples.

Linkage criteria determine how distances between clusters are computed. Common criteria include single linkage (minimum distance between points in different clusters), complete linkage (maximum distance), and centroid linkage (distance between cluster centroids). For example, single linkage may result in "chaining" where clusters form long, thin shapes, while complete linkage tends to form more compact clusters.

3.3 Discuss the advantages and limitations of using hierarchical clustering for dynamic datasets that evolve over time.

Advantages include the ability to visualize the hierarchical structure of data and to handle datasets with different shapes and sizes. Limitations include high computational cost for large datasets and difficulties in updating clusters dynamically as new data arrives. Hierarchical clustering may also struggle with noisy data and outliers.

3.4 What are the computational challenges of hierarchical clustering for large datasets, and how can they be addressed?

Computational challenges include the high time complexity, which is generally $O(n^3)$ for agglomerative methods due to the need to compute and update the distance matrix. These challenges can be addressed by using more efficient data structures, approximating distances, or employing scalable algorithms such as BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).

3.5 Describe how dendrograms can be used to make decisions about the number of clusters in hierarchical clustering.

Dendrograms visually represent the hierarchical structure of clusters. By examining the dendrogram, one can choose a threshold level to cut the tree and decide on the number of clusters. For instance, a large vertical distance without merging in the dendrogram suggests an appropriate level to cut for cluster formation.

3.6 Propose a modification to the hierarchical clustering algorithm to better handle real-time data updates.

To handle real-time data updates, one could implement incremental clustering techniques where clusters are updated dynamically as new data arrives. Techniques like updating cluster centroids or maintaining a live distance matrix can help in accommodating new data without recalculating the entire clustering structure.

3.7 Discuss the ethical implications of using clustering algorithms in applications like wildlife management or conservation.

Clustering algorithms in wildlife management and conservation can help in monitoring and preserving species diversity. However, ethical considerations include ensuring that algorithms do not lead to biased or harmful decisions, such as prioritizing certain species over others or misrepresenting species distributions.

3.8 How would you modify your clustering approach to handle traits that are interdependent or correlated?

To handle correlated traits, dimensionality reduction techniques such as Principal Component Analysis (PCA) can be used to reduce the number of features while preserving variance. Additionally, incorporating methods that explicitly

account for feature correlations, such as distance metrics that consider feature interdependencies, can improve clustering results.

3.9 Explain how hierarchical clustering can be integrated with other machine learning techniques to improve ecosystem modeling.

Hierarchical clustering can be integrated with machine learning techniques such as supervised learning to enhance ecosystem modeling. For example, clustering results can be used as features for classification algorithms, or clustering can be combined with anomaly detection to identify outliers or emerging patterns in species evolution.

3.10 Compare and contrast hierarchical clustering with other clustering methods like k-means and DBSCAN in the context of evolving datasets.

Hierarchical clustering provides a comprehensive hierarchical structure but may be less scalable for large datasets. K-means clustering requires specifying the number of clusters and is sensitive to initial conditions but is computationally efficient. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) does not require specifying the number of clusters and can handle noise, but it may struggle with varying cluster densities. Each method has strengths and weaknesses depending on the characteristics of the evolving dataset.