Université Paris-Est Créteil

UFR de Sciences et Technologie

Assignment of Practice Lab: Scipy Dendrogram

Master 2 Optique, Image, Vision, Multimédia parcours SIM

Yaghmoracen BELMIR

Academic year 2023-2024

Introduction

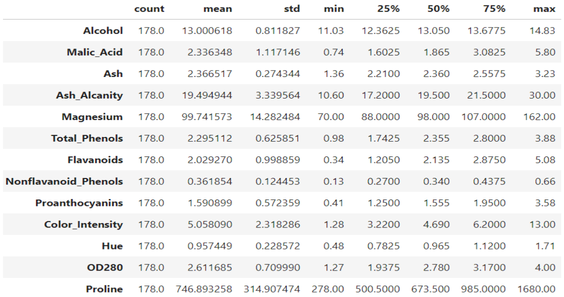
Dendrograms, a key component of hierarchical clustering, provide an insightful visual representation of the relationships and structures within a dataset. Hierarchical clustering is a powerful technique for organizing data into a hierarchy of clusters, which allows for a more detailed understanding of the data's inherent structure. Dendrograms, in particular, offer a graphical means of representing these hierarchical relationships by displaying the order and distances at which data points or clusters are merged. This report explores the significance of dendrograms in the context of hierarchical clustering, highlighting their role in revealing the natural divisions and similarities among data elements. By examining the intricacies of dendrograms, we gain valuable insights into the underlying structures of datasets and their applications across various fields.

Dataset description

This dataset is adapted from the Wine Data Set[1] obtained from the UCI Machine Learning Repository. The original dataset includes information about three types of wine. However, for this unsupervised learning analysis, the information regarding wine types was removed.

The dataset contains measurements of 13 constituents in wines, originating from three different cultivars. The attributes included are as follows:

1. **Alcohol**: Percentage of alcohol in the wine. This parameter measures the quantitative content of ethanol in the wine and affects its strength.
2. **Malic Acid**: Amount of malic acid in the wine. Malic acid imparts freshness and brightness to the wine.
3. **Ash**: Amount of mineral substances (ash) in the wine after water evaporation and residue incineration. It reflects the wine's mineral content.
4. **Alcalinity of Ash**: Alkalinity of ash in the wine. Alkalinity measures the wine's pH level and influences its taste characteristics.
5. **Magnesium**: Amount of magnesium in the wine. Magnesium is one of the trace elements that can affect the taste and aroma of the wine.
6. **Total Phenols**: Total amount of phenolic compounds in the wine. Phenols are antioxidants and can influence the taste and color of the wine.
7. **Flavanoids**: Amount of flavonoids in the wine. Flavanoids are also phenolic compounds and can contribute to the taste and color of the wine while possessing antioxidant properties.
8. **Nonflavanoid Phenols**: Amount of nonflavonoid phenolic compounds in the wine.
9. **Proanthocyanins**: Amount of proanthocyanidins in the wine. Proanthocyanidins also belong to the group of phenolic compounds.
10. **Color Intensity**: Wine's color intensity, measured as light absorption at a specific wavelength. This parameter is related to the wine's color depth.
11. **Hue**: Wine hue, measured on a color scale. This value can vary from orange to purple and is related to the wine's color subtlety.
12. **OD280/OD315 of Diluted Wines**: Optical density of the wine at a specific wavelength. This parameter may be related to the wine's anthocyanin content (pigments that give wine its red color).
13. **Proline**: Amount of the amino acid proline in the wine. Proline can influence the texture and structure of the wine.



	count	mean	std	min	25%	50%	75%	max
Alcohol	178.0	13.000618	0.811827	11.03	12.3625	13.050	13.6775	14.83
Malic_Acid	178.0	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.80
Ash	178.0	2.366517	0.274344	1.36	2.2100	2.360	2.5575	3.23
Ash_Alcanity	178.0	19.494944	3.339564	10.60	17.2000	19.500	21.5000	30.00
Magnesium	178.0	99.741573	14.282484	70.00	88.0000	98.000	107.0000	162.00
Total_Phenols	178.0	2.295112	0.625851	0.98	1.7425	2.355	2.8000	3.88
Flavanoids	178.0	2.029270	0.998859	0.34	1.2050	2.135	2.8750	5.08
Nonflavanoid_Phenols	178.0	0.361854	0.124453	0.13	0.2700	0.340	0.4375	0.66
Proanthocyanins	178.0	1.590899	0.572359	0.41	1.2500	1.555	1.9500	3.58
Color_Intensity	178.0	5.058090	2.318286	1.28	3.2200	4.690	6.2000	13.00
Hue	178.0	0.957449	0.228572	0.48	0.7825	0.965	1.1200	1.71
OD280	178.0	2.611685	0.709990	1.27	1.9375	2.780	3.1700	4.00
Proline	178.0	746.893258	314.907474	278.00	500.5000	673.500	985.0000	1680.00

Figure 1: Data set description

Methodology

In this section, we outline the steps taken to analyze the dataset and determine the optimal number of clusters. The analysis was performed using a combination of hierarchical clustering, silhouette analysis, and the identification of key clustering features.

Data Inspection and Preprocessing

The first step involved examining the dataset for any missing or incomplete values. It was confirmed that the dataset was complete and ready for analysis.

Hierarchical Clustering with Dendrograms

To identify an appropriate method for hierarchical clustering, we explored various linkage methods and distance metrics.

Linkage Methods	**Distance Metrics**
single	euclidean
complete	chebyshev
average	correlation
weighted	

Table 1: Considered Linkage Methods and Distance Metrics

Note :

$$d_{\text{euclidean}}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$d_{\text{chebyshev}}(p,q) = \max_i(|p_i - q_i|)$$

$$d_{\text{correlation}}(p,q) = 1 - \frac{\text{cov}(p,q)}{(\text{std}(p) \cdot \text{std}(q))}$$

Hierarchical clustering was performed for each possible combination of linkage method and distance metric.

The quality of the clustering was assessed using the CCC[1] and the depth of the R[2] matrix. To understand the impact of the depth parameter in the inconsistent matrix, we set it to 5 for all combinations, and it was noted that changes in the depth value affected the inconsistency coefficient.

The R matrix quantifies the inconsistency in the hierarchical clustering dendrogram. It is computed using the “inconsistent” function in Python’s SciPy library. The formula is:

$$R(i) = \frac{h(i) - c}{t(i)}$$

Where: $R(i)$ is the inconsistency coefficient for node i. $h(i)$ is the height of node i in the dendrogram. c is the height of the cut that forms node i. $t(i)$ is a factor, typically 3, to adjust the threshold for considering node i inconsistent.

The combination of method “weighted” and metric “euclidean” produced the highest CCC value of 0.8066, indicating the best combination for clustering.

Determining the Number of Clusters

With the optimal combination selected, we applied the ’fcluster’ method to determine the number of clusters (k). We tested a range of k values from 2 to 14. For each k value, we calculated the silhouette score to evaluate the quality of the clusters. The silhouette score measures the similarity of data points within clusters and helps identify the most suitable number of clusters.

The k value that resulted in the highest silhouette score was selected as the optimal number of clusters. In our analysis, k = 3 yielded the highest silhouette score, indicating that the dataset could be best clustered into three groups.

Results[2]

We can observe various dendrograms generated for every combination of linkage methods and distance metrics mentioned above:

[1] CCC (Cophenetic Correlation Coefficient): Measures how well a hierarchical clustering preserves original data distances. Higher CCC values indicate better clustering.

[2] R Matrix: Represents the inconsistency of merges in hierarchical clustering, often used to identify meaningful clusters by analyzing inconsistencies.

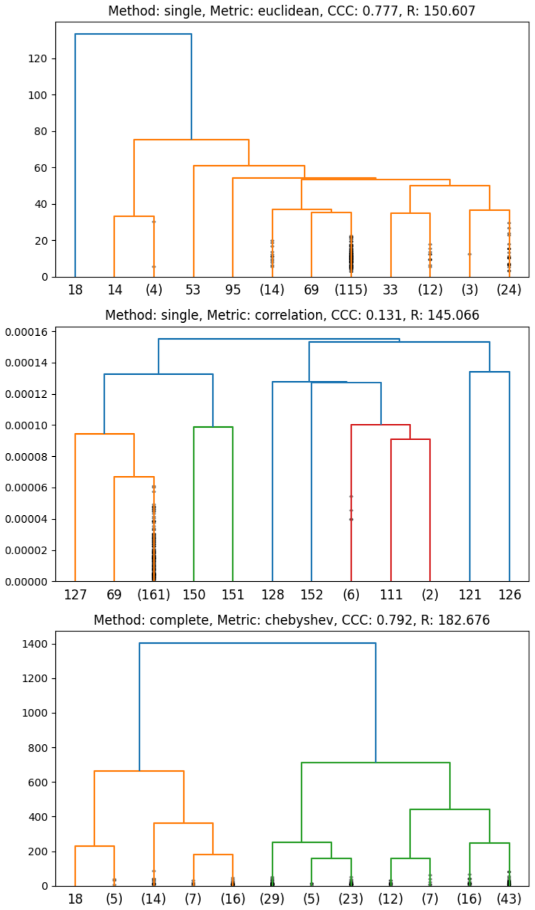


Figure 2: Three different combination sample.

After selecting the optimal cluster with a CCC value of 0.8066, we can visually divide it into 2, 3, or up to 14 clusters.

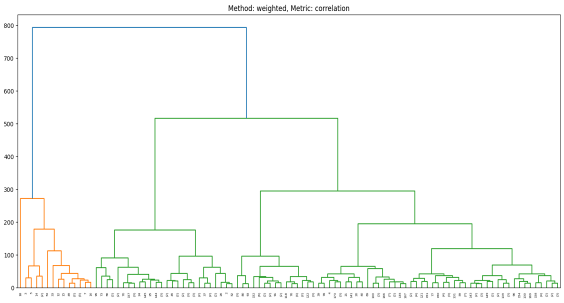


Figure 3: Dendrogram with highest value of CCC.

After utilizing the plot to determine the K value with the highest silhouette value, we observed that k = 3 exhibited the highest value, as shown in the figure 4.

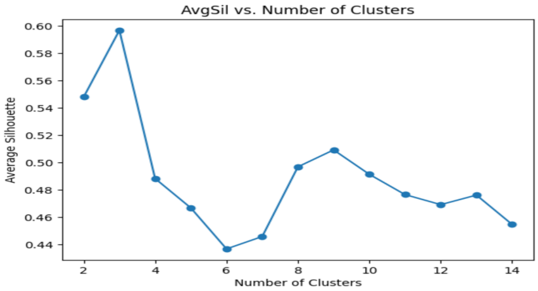


Figure 4: Various k values with their corresponding silhouette values.

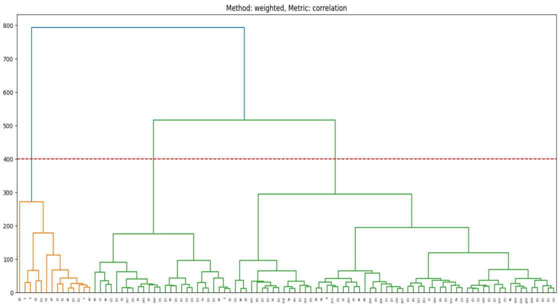


Figure 5: Cut off k = 3 for the clustering.

Conclusion

The significance of the findings highlights the effectiveness of hierarchical clustering in categorizing the wine dataset. The dataset originally consisted of three types of wines, and the clustering results, particularly the optimal choice of k = 3, align with this initial classification. This outcome suggests that hierarchical clustering has the potential to accurately identify the inherent structure within the data, reinforcing its relevance to the wine dataset.

These findings offer valuable insights not only in terms of data exploration but also in practical applications. The clustering results can be used for wine quality assessment, targeted marketing, or even the development of new wine varieties. Understanding the inherent groupings within the dataset can aid in making informed decisions in the wine industry, such as product recommendations, quality control, and targeted advertising.

In summary, the hierarchical clustering results reinforce the dataset's inherent structure, providing a foundation for various applications within the wine industry.

Bibliography

[1] UCI Machine Learning Repository. *Wine*. `https://archive.ics.uci.edu/ml/datasets/wine`.

[2] BELMIR Yaghmoracen. *THE CODE*. `https://www.kaggle.com/yaghmoracenbelmir/agglomerative-clustering-dendrogram`.