MDM-2024 Homework 2

Cuong Nguyen (101559968), Petteri Raita (909635),
and Raihan Gafur (101555441)

October 7, 2024

Contents

List of Figures

1 Methods

In this section, we describe the methodology employed in the analysis, including feature extraction, distance metric formulation, feature selection, clustering process, and evaluation of clustering quality.

1.1 Feature Extraction

Feature extraction involves creating new variables that capture the underlying patterns in the data and preparing the features for clustering.

1.1.1 Numerical Features

The numerical features in the dataset were initially represented in ranges. To simplify and standardize the data, the **mean** of each range was taken, resulting in a single representative value per feature. This helped in reducing the complexity and allowed for straight-forward comparisons between data points.

Additionally, several new features were engineered to capture relationships between different measurements. Specifically:

- **Body Mass Index (BMI):** This was computed as:

$$\text{BMI} = \frac{\text{weight}}{\text{length}^2}$$

  BMI provides a standardized ratio of weight to length, useful for characterizing the relative bulk of the subjects.

- **Wing Span Index (WSI):** This was computed as:

$$\text{WSI} = \frac{\text{wspan}}{\text{length}}$$

  WSI represents the proportion between wing span and length, giving insight into the morphological characteristics.

These new features, along with other ratios such as **AR** (Aspect Ratio) and **wload** (Wing Load), were selected for clustering. These ratios collectively capture critical characteristics of the subjects, aiding in distinguishing between different groups.

1.1.2 Categorical Features - Color Features

For the categorical color features, a semantic approach was adopted to improve the effectiveness of clustering. Instead of treating all colors as distinct categories, the similarity between color shades was considered. A hierarchy of colors was developed in which shades with similar base colors were grouped. For example, "light grey" and "bluish grey" were treated as being closer to each other since they shared the main color, "grey." This approach ensured that the color feature was represented in a way that captured inherent similarities, making the distance metric between instances more meaningful.

```
def color_extraction(color_str: str) -> list:
    '''
    Extract modifier and main_color from a given color string
    e.g. "light grey" -> "light", "grey"
    '''
    color_elements = color_str.split(' ')
    if len(color_elements) == 1:
        return ['', color_str]

    return color_elements

def color_dist(color1: str, color2: str) -> float:
    '''
    Return color distance between the two given color strings.
    This custom color distance includes two components: modifier
    distance and main color distance, where each component has its
    own weight.

    D = modifier_weight * modifier_distance + main_color_weight * main_color_distance,
    modifier_weight + main_color_weight = 1
    '''
    modifier1, main_color1 = color_extraction(color1)
    modifier2, main_color2 = color_extraction(color2)

    main_color_dist = main_color_rgb_dist(main_color1, main_color2)
    modifier_dist = levenshtein_dist(modifier1, modifier2)

    return (1-MAIN_COLOR_DIST_WEIGHT)*modifier_dist + MAIN_COLOR_DIST_WEIGHT*main_color_dist
```

1.2 Combined Distance Measure

For the numerical features, Euclidean distance was utilized to measure pairwise distances, as specified in the exercise requirements. In contrast, pairwise distances for categorical color features were calculated using a custom distance metric. First, the color value was separated into the main color and its shade. The distance between two main colors was determined by the Euclidean distance between their respective RGB values. For color shades, the Levenshtein distance was employed, which measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one word into another. In this case, color shades were represented as strings, such as "light", "dark", or "bluish". Consequently, the overall pairwise color distance was a weighted combination of the main color distance and the color shade distance, with a weight of 0.8 assigned to the main color distance. This weighting reflects the primary role of the main color component in the visual appearance.

The clustering process required a combined distance measure. For that reason, a custom function was used to incorporate both numerical and categorical features using the eq (1).

$$\text{sim}(x, y) = \lambda \cdot \frac{\text{numSim}}{\sigma_n} + (1 - \lambda) \cdot \frac{\text{catSim}}{\sigma_c} \tag{1}$$

Here, the value of λ has been calculated based of on both numerical and categorical features:

$$\lambda = \frac{\text{no. numerical features}}{\text{no. numerical features} + \text{no. categorical features}} \tag{2}$$

1.3 Feature Selection

During feature selection, it was crucial to identify the most informative features while discarding redundant or less relevant ones. The final feature set included four numerical features: **AR**, **wload**, **BMI**, and **WSI**, and two color features: **belly** and **back**. These features were selected because they represented critical morphological and functional characteristics of the subjects. Features like weight, length, and wspan were dropped because **BMI** was calculated using weight and length, and **WSI** was derived from wspan and length, making their inclusion redundant.

1.4 Clustering

1.4.1 Method

Agglomerative Hierarchical Clustering was chosen as the primary clustering method. This approach is well-suited for datasets where the underlying relationships are unknown, as it allows for visualizing the hierarchical relationships between data points through a dendrogram.

1.4.2 Linkage Metrics

Various linkage metrics have been used, such as **single**, **complete**, and **average** linkage, to determine the optimal grouping. Linkage metrics **ward** has been dropped from the linkage because ward does not work on pre-computed distance matrix [1].

1.4.3 Number of Clusters

A range of numbers of clusters (range(5, 13)), was considered to get the best combination of clusters.

1.4.4 Implementation

The implementation of the clustering was done using a standard machine learning library such as `scikit-learn` in Python. The library's efficient implementation of hierarchical clustering methods would allow for straightforward experimentation with different linkage metrics and cluster numbers.

1.5 Evaluation

For each combination of features, different linkage methods and numbers of clusters were evaluated using agglomerative hierarchical clustering. The `AgglomerativeClustering` algorithm was applied with the precomputed distance matrix, varying both the number of clusters and the linkage method to determine the optimal clustering configuration. During this process, configurations that produced singleton clusters (clusters containing only a single data point) were skipped, as they excessively impact the normalized mutual information (NMI) score.

2 Results

2.1 NMI

The different clusterings were evaluated for all the different feature combinations, linkage methods, and number of clusters. The scoring for each individual combination of features was based on the NMI of the cluster with the real labels (the biological groupings). However, the best NMI value included 12 clusters of which there were one or more singleton clusters. Since NMI is excessively influenced by the singleton clusters, those were filtered out.

As a result, the highest NMI score for a clustering was 0.727. This score was provided by a combination of 11 cluster with the selected features being AR, wload, WSI, and the back color. The clustering used a complete linkage. Comparing this clustering to the other top candidates in terms of NMI, the common features were very much the same as with the chosen clustering. Only a few combinations dropped the AR or the wload. Out of the best 10 combinations, none used BMI as a feature. This is indicative of the statistically low information provided by the BMI feature. The number of clusters was between 9 and 12, which is a relatively small difference.

numerical_features	color_features	n_clusters	linkage	NMI
[AR, wload, WSI]	[back]	11	complete	0.727361
[wload, WSI]	[back]	10	complete	0.709282
[AR, wload, WSI]	[back]	10	complete	0.697048
[AR, wload, WSI]	[back, belly]	11	complete	0.683040
[AR, BMI, WSI]	[back]	12	complete	0.681305
[AR, WSI]	[back]	11	complete	0.680133
[AR, wload, WSI]	[back]	9	complete	0.677938
[AR, wload, WSI]	[back, belly]	10	complete	0.667347
[AR, BMI, WSI]	[back]	11	complete	0.666402
[AR, wload, WSI]	[belly]	10	complete	0.658753

Figure 1: *NMI Values*

The normalized mutual information score was calculated through the Sklearn library, which uses the following formula for the calculation of the mutual information score. The normalization of the score uses the entropies of the classes and labels to normalize the MI value to fall between 0 and 1.

The Mutual Information is a measure of the similarity between a cluster and the biological grouping. In the code, the true labels for NMI score were the 'group' column of the data. The predicted labels were found through fitting the clustering method to the data.

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \left(\frac{N|U_i \cap V_j|}{|U_i||V_j|} \right)$$

2.2 Analysis of the cluster contents

Comparing the clustering results with the actual groups and flying types, it is evident that there are clear common features among the clusters. For example, the cluster contents align well with the flying types, notably clusters 0, 1, and 2, which include only type C, B, and C birds, respectively. Cluster 0 contains only species from the podicipedidae family, while cluster 2 exclusively includes species from the gaviidae family. The biological families are well separated, indicating effective clustering performance.

Regarding the features, we notice that all birds in cluster 8 are dappled brown. Another observation is that all birds in cluster 7 have a WSI (Wingspan Index) of more than 0.9, which is extremely high. This cluster includes hawks such as ruskosuohaukka and haarahaukka, demonstrating a great clustering. WSI is the wingspan divided by the length of the bird, and a high WSI indicates that these birds have a large wingspan compared to their body length. This is intuitive since hawks are excellent fliers. Future research should look at the intercluster distance and the similarities of different bird groups. An interesting hypothetical research question would be to look at what are the closest bird groups to the duck family. In our experiment almost all ducks were in the cluster 8. Additionally, the evaluation of the cluster contents could be made more systematic with an algorithm that would categorize the birds inside a cluster to correspond with certain biological trademarks.

The clusters were not placed in a dendogram, but the linkage metric dendogram is shown in the Figure 2. The dendogram has the actual similar biological groups denoted by the same color. As seen from the groupings of the colors, the dendogram accurately groups birds of the same biological group together. The dendogram is created using the complete linkage with the pairwise distances.

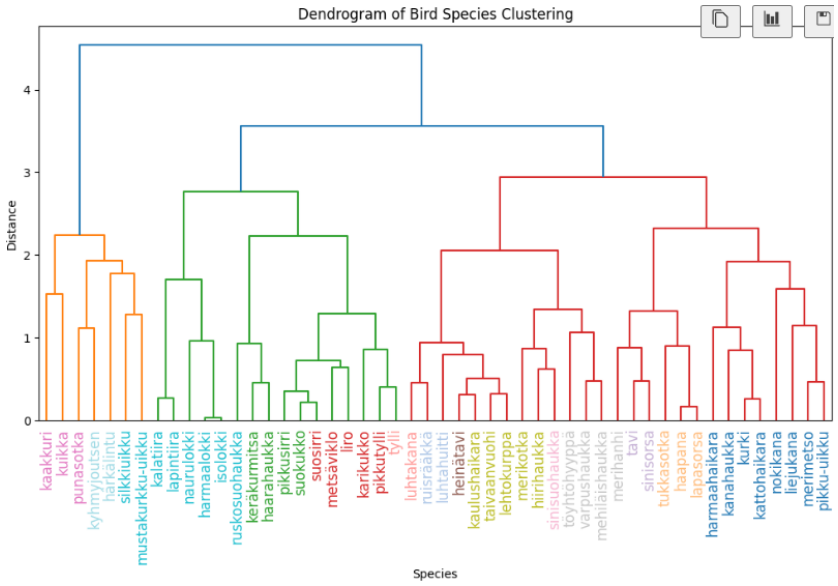

Figure 2: *Dendogram*

3 Conclusions

The clustering analysis effectively identified key features that align well with biological groupings and flying types of birds. The highest normalized mutual information (**NMI**) score achieved was 0.727, using a combination of 11 clusters with features including aspect ratio (**AR**), wload, wingspan index (**WSI**), and back color, and employing complete linkage. Importantly, **BMI** was consistently excluded from the top clustering combinations, suggesting its low informative value.

Analysis of the cluster contents revealed strong biological correlations. For instance, clusters 0, 1, and 2 aligned exclusively with type C, B, and C birds, respectively, and separated biological families such as podicipedidae and gaviidae effectively. Hawks, grouped in cluster 7, displayed a high WSI value, reinforcing their flying prowess. Similarly, birds in cluster 8 were all dappled brown, showing distinct feature-based clustering.

The results highlight that features like **AR**, **wload**, and **WSI** are significant in distinguishing bird groups. Future research could explore inter-cluster relationships and develop systematic methods to categorize cluster contents based on biological traits. The dendrogram further confirmed the accuracy of these groupings, visually clustering birds of the same biological families together using complete linkage.

References

[1] Sklearn. *Agglomerative Clustering*. 2024. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html (visited on 10/06/2024).

Appendices

The code is publicly accessible at https://github.com/ancuongnguyen07/CS-E4650/tree/master/hw2