Analysis of Penguin Species with K-Means Clustering

Yale University - Current Topics in Applied Machine Learning
Filipa Monteiro
March 11th, 2024

Introduction

Analysis of morphological features of species alongside genetics is fundamental in the study of evolution. The combined knowledge of the latter provides spatial-temporal information that is necessary for the reconstruction of the evolutionary history of species. Divergent evolution of a species into its subspecies is amongst the events that are of particular interest to study given that differing morphological features arise for individuals of the same species – individuals with extremelly high genetic similarity - due to exposure to different environmental and social pressures. While genetic analysis can provide insight into what species a fossil belongs to, the variations in the morphological aspects are what allow one to hypotheisze what types of events may have led to the divergence and what type of habitat each set of individuals of the species may have habitted in[1,2]. Therefore, it is of crutial importance to be able to analyze large fossil data that accounts for varied measurements of a species' features and speculate how many differing subspecies the species may have diverged into due to allopatric speciation[2]. This work intends to analyze if k-means clustering would be a helpful tool in hypothesizing the number of subspecies that arise in divergence events via analysis of morphological features' clustering. Given that no allopatric speciation-specific datasets were found, a dataset of features containg data from multiple penguin species is considered for analysis. Extrapolation of the clustering results between different species vs subspecies is considered given the similarity of the two concepts in morphological terms.

Background Work

Machine learning algorithms have been used to conduct analysis of evolution-related events for many years now. Algorithms like minimal representation size clustering (MRSC)[3] – a genetic algorithm[4], support vector machines (SVM)[5], k-means clustering[6] and other clustering algorithms have been used for multiple purposes from analysis of evolutionary

speciation[4] to morphometric quantification of taxa[5] and gene expression analysis[6,7,8]. K-means clustering has mainly been a prominent algorithm in the gene expression domain of biological systems but has not been, according to this research's findings, widely applied for studies in the realm of morphological analysis within evolutionary biology. Given the k-means clustering's reputation as a simple, flexible and computationally inexpensive algorithm, the latter has been, however, widely used in multiple domains that require clustering[9] and was selected as the method of this research.

Method Description

Data Processing

The pre-processed dataset was collected from "Kaggle", a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC (https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species). The dataset selected has a collection of penguin features coming from multiple penguin species – amount of species not specified. The features considered for analysis of the penguins species for clustering purposes include culmen - upper ridge of a bill - length (mm), culmen depth (mm), flipper length (mm), body mass (g) and sex (male vs female). Processing of the initial dataset included oulier exclusion as well as exclusion of datapoints whose measurement were not assigned to a sex (sex not specified in raw dataset). Data scalling was also conducted for all features of the processed data – code analysis of the head of the scalled data provides insights on the range of values associated with each feature and was resorted to in further analysis of the data with regards to the results.

Model: K-Means Clustering

K-means clustering is a distance-based unsupervised clustering algorithm that aims to cluster data points into K clusters, where each point belongs to the cluster with the nearest mean – the so called, cluster centroid. The goal of this algorithm is to iteratively minimize the within-cluster sum of squared distances from each point to its assigned centroid. The algorithm can be succinctly explained as follows[10,11]:

1. *Initialization*: Randomly select K data points as initial cluster centroids;
2. *Assignment Step*: Assign each data point to the nearest centroid using the Euclidean distance metric $d(x_i, c_j) = \sqrt{\sum_{n=1}^{N}(x_{i,n} - c_{j,n})^2}$, where x_i represents a given data point, c_j denotes the centroid of cluster j, and N represents the dimensionality of the data;
3. *Update Step*: Recalculate the centroids based on the mean of the data points assigned to each cluster $c_j = \frac{1}{|S_j|}\sum_{x_i \in S_j} x_i$, where x_i is a given data point, c_j denotes the centroid of cluster j, and S_j represents the set of data points assigned to cluster j;

The algorithm undergoes multiple iterations until convergence is achieved. Ideally, convergence will be associated with the achievement of negligible centroid changes in position; however, for more elaborate and complex systems that may not show easy of convergence, an "artifiial convergence" may be user-set – a maximum number of iterations is defined.

Selection of Optimal K-value

In order to select an ideal k-value for the dataset selected, many methods have been proposed. The most widely used one is deseignated the "Elbow Method" – a graphical method that "shows the within-cluster-sum-of-square (WCSS) values on the y-axis corresponding to the different values of K (on the x-axis)"; the optimal k-value is accepted as that in which an elbow inflection point is observed in the elbow graph[12]. However, in the majority of the datasets, there isn't a clear inflection point and thus, alternative methods have also been explored. One of the proposed methods that seems to be also popular is the "Silhouette Score":

$$Silhouette\ Score = \frac{b-a}{\max(a,b)}$$

,where a is average intra-cluster distance, meaning the average distance between each point within a cluster and b is average inter-cluster distance - the average distance between all clusters.

The latter score presents a range of classification [-1,1] with -1 meaning points incorrectly assigned to cluster, 0 meaning overlapping clusters, and 1 meaning perfect point assignment to cluster with easily distinguishable clusters[12]. To note that both optimization methods were explored for this research.

Additonal Analysis: Pairwise Relationships Plot and PCA

While analysis via k-means clustering is essential to try and undertsnad the data, further analysis of these clusters via Principal Component Analysis (PCA) and dataset trends' analysis

via a pairwise relationships plot allowed further insight into the data. PCA was conducted to try and determine data features that had higher importance in the determinning cluster separation. An analysis of the pairwise relationiships plot intended to help better understand the dataset and provide some a priori knowledge of potential “obvious” feature separation that might influence clustering.

Results and Discussion

Pairwise Relationships Plot and PCA

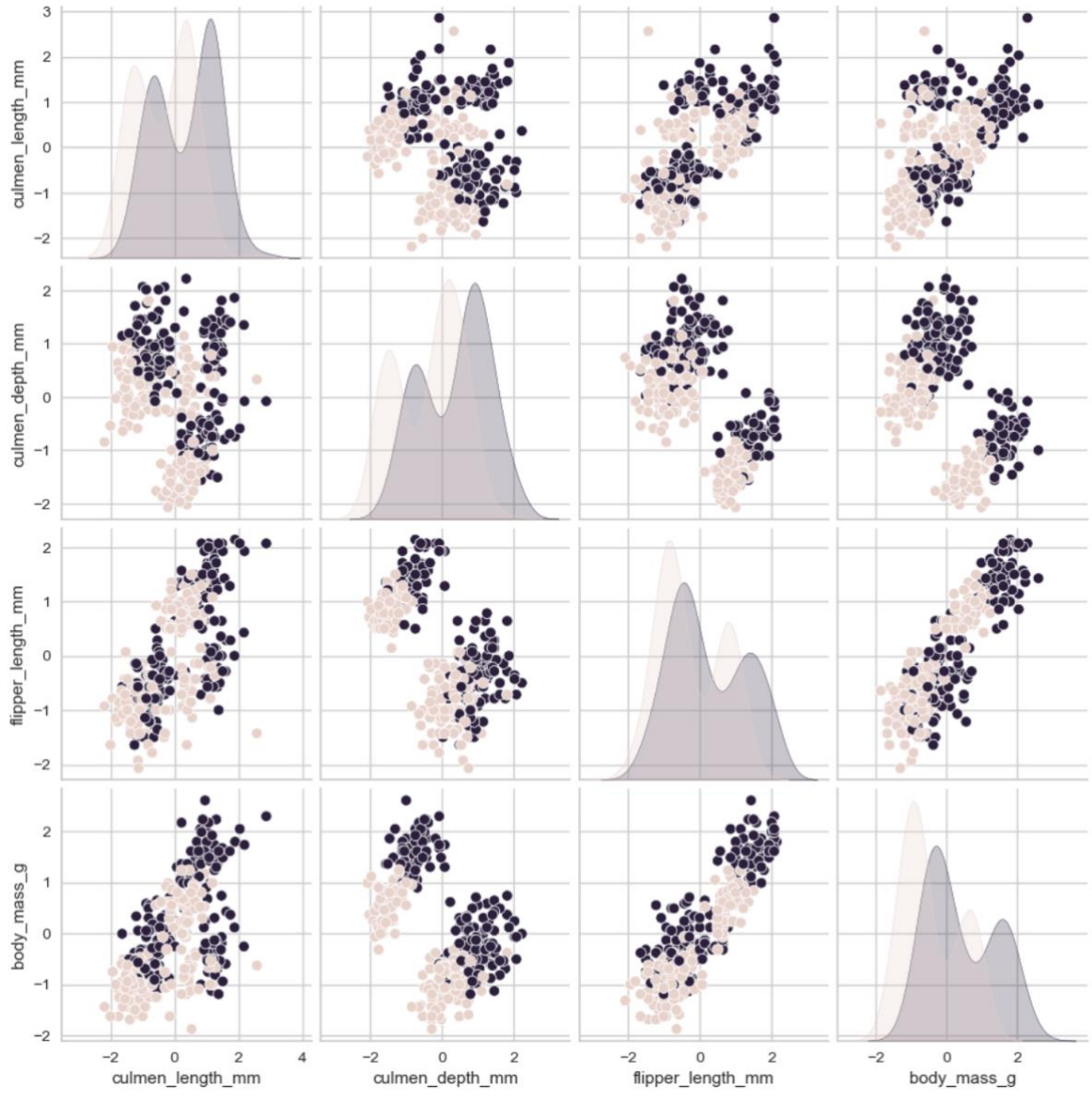

Figure 1. Pairwise Relationships - analysis via the python data visualization library Seaborn

Analysis of pairwise relationships showed clear separation between sexes – male showing in dark purple with a standardized value of 0.993994 and female showing in pink with a standardized value of -1.006042. Analysis of the diagonal patterns show two bimodal curves – one for each sex - that seems to indicate two populations may be present, with significant feature variations between sexes. Comparison of the relationships above seem to show a direct relation between body mass and flippler length, body mass and culmen length, and flippler length and culmen length. Inverse relations seem to be present between flippler length and culmen depth, body mass and culmen depth, and culmen depth and culmen length. Given this pattern, it is expected that a population with larger body mass will have larger flipper length and culmen length and smaller culmen depth, while a population with lower body mass will have lower flipper length and lower culmen length but larger culmen depth. Afterwards, PCA analysis considering an explained variance ratio of 0.25 determined presence of 2 principal components that would explain the dataset analyzed. Analysis of the ranges of values for PCA 1 and PCA 2 showed the ranges]-3,1[and]0,2[, respectively. Comparison of these ranges with those associated with each of the features under analysis allowed the speculation that PCA 2 is associated with the culmen depth given that the latter feature is the only one that has its values spaning the whole range denoted, making culmen depth a main feature in the determination of the differences between the clusters and potential a key differing feature between penguin species. There wasn't a specific feature that seemed to span the whole PCA 1 range and thus, a combination of features is thought to be associated with the latter. This data seems to be in line with the pairwise relationship analysis noted above. This poses a hypothesis that k-clustering could potentially lead to 2 or 4 clusters – 2 if sex is disregarded and the centroid is solemnly focused on sex-independent features, and 4 if sex is considered, since each population would be subdivided into its male and female populations with differing centroids that would arise from the significant sex-related differences.

Optimal K-value Selection

The Elbow Method showed an optimal k-value of 4; however, graphical visualization did not show a clear inflection point and thus, the silhouette score was also determined. The latter showed an optimal k-value of 4 with a silhouette score of 0.509. PCA analysis considering an explained variance ratio of 0.25 determined presence of 2 principal components that would explain the dataset analyzed. The latter k-value was considered for k-means clustering.

K-Means Clustering

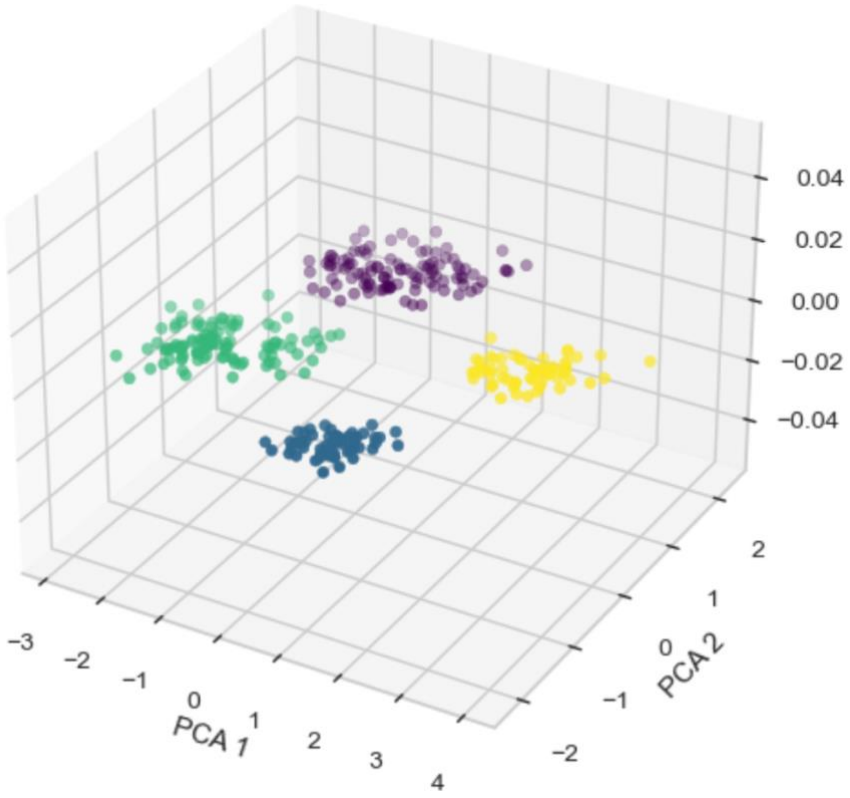

Figure 2. K-Means Clustering (K=4) with PCA Explained Variance Ratio >0.25 (n=2)

K-means clustering with k-value=4 considering PCA with n=2 was plotted and showed 4 distinct clusters. Considering the pairwise relationship analysis of the raw data, in which males tends to show higher values for all features in comparison to female, and also considering

our previous hypothesis that populations would have its female and male portions subdivided if there was a significant sex-driven feature average mean within each population that might lead to two diverging centroids, the clusters observed seem to be distinguishable as follows: yellow and purple clusters associated with population 1 (higher sex-independent culmen depth) and blue and green clusters associated with populaton 2 (lower sex-independent culmen depth). Under the scope of this research, each population would directly correlate to a different penguin species. In divergence studies, one would hope that these findings would be able to separate much more closely related/less dispar changes in features to be able to show clustering representative of subspecied that occurred due to allopatric speciation. Further conclusions cannot be withdrawn from this clustering without additional analysis due to the ambiguous nature of the PCA 1 component.

Conclusion

Overall, this study showed that k-means clustering has great potential for being used as an allopatric speciation analysis algorithm based on morphological features' clustering. While the results presented in this research were those for clustering of two different species, the results can be extrapolated and expected similar for subspecies differentiation. Nevertheless, given that the degree of difference in features between different species is potentially much more significant than that between subspecies, slight variations to the k-means clustering method may be necesssary.

References

1. Earth, D. O., & Agriculture, B. O. (2019). Evaluating the taxonomic status of the Mexican gray Wolf and the red Wolf. In *National Academies Press eBooks*. https://doi.org/10.17226/25351
2. Murdock, J. and Yaeger, L.S. (2011) 'Identifying species by genetic clustering', *ECAL 2011: The 11th European Conference on Artificial Life* [Preprint]. doi:10.7551/978-0-262-29714-1-ch087.
3. Hacaoğlu, C. and Sanderson, A.C. (1995) 'Evolutionary speciation using minimal representation size clustering', *Evolutionary Programming IV*, pp. 187–204. doi:10.7551/mitpress/2887.003.0022.
4. Bartz-Beielstein, T., Branke, J., Mehnen, J. and Mersmann, O. (2014), Evolutionary Algorithms. WIREs Data Mining Knowl Discov, 4: 178-195. https://doi.org/10.1002/widm.1124
5. Van Bocxlaer, B., & Schultheiß, R. (2010). Comparison of morphometric techniques for shapes with few homologous landmarks based on machine-learning approaches to biological discrimination. *Paleobiology*, *36*(3), 497–515. http://www.jstor.org/stable/40792302
6. Botía, J.A. *et al.* (2017) 'An additional K-means clustering step improves the biological features of WGCNA gene co-expression networks', *BMC Systems Biology*, 11(47). doi:10.1186/s12918-017-0420-6.
7. Lu, Y. *et al.* (2004) 'Incremental genetic k-means algorithm and its application in gene expression data analysis', *BMC Bioinformatics*, 5(172). doi:10.1186/1471-2105-5-172.
8. Oyelade, J. *et al.* (2016) 'Clustering algorithms: Their application to gene expression data', *Bioinformatics and Biology Insights*, 10. doi:10.4137/bbi.s38316.
9. Ikotun, A.M. *et al.* (2023) 'K-means Clustering Algorithms: A comprehensive review, variants analysis, and advances in the era of Big Data', *Information Sciences*, 622, pp. 178–210. doi:10.1016/j.ins.2022.11.139.
10. *CS221*. (n.d.). https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
11. MacQueen, J.B. (2011) 'Some methods for classification and analysis of multivariate observations', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), pp. 281–297. doi:10.7551/978-0-262-29714-1-ch087.
12. Tomar, A. (2024) *Stop using elbow method in k-means clustering, Built In*. Available at: https://builtin.com/data-science/elbow-method# (Accessed: 04 March 2024).