# K-Means Clustering - Anuran Calls Dataset

Subham Ghosh

November 5, 2024

## 1 Introduction

This report explores K-Means clustering applied to the Anuran Calls Dataset, consisting of MFCC coefficients of frog calls. The objective is to group frog species based on their acoustic features.

## 2 Data Preprocessing and Exploration

### 2.1 Exploratory Data Analysis

The dataset is analyzed for missing values, outliers, and feature distributions. Data scaling is applied using standardization to improve clustering performance.

### 2.2 Feature Engineering

Additional features are created using polynomial and interaction terms, with the aim of enhancing cluster separation.

## 3 K-Means Clustering

### 3.1 Elbow Method

The Elbow Method is used to determine the optimal number of clusters, where the optimal value of $K$ is chosen based on the point where inertia starts to decrease linearly.

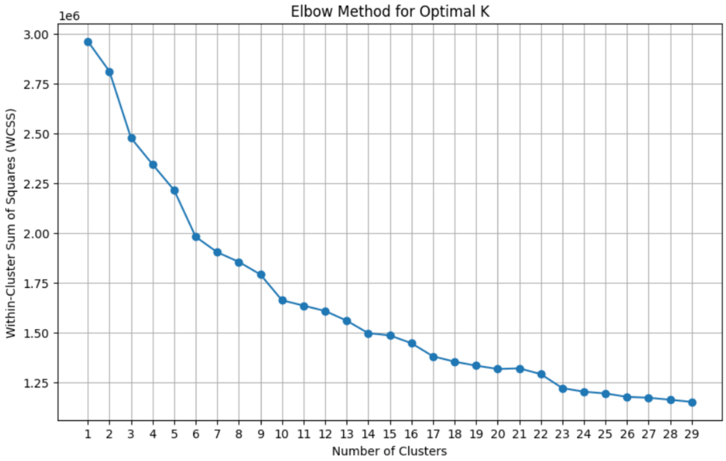

Figure 1: Elbow method for optimal K

### 3.2 Silhouette Score Evaluation

The silhouette score is calculated for each clustering result to assess the quality of clusters.

### 3.3 Cluster Initialization

A comparison of K-Means with random initialization vs. k-means++ initialization is presented, with silhouette scores reported for each method.

## 4 Cluster Visualization

### 4.1 Dimensionality Reduction

Principal Component Analysis (PCA) is applied to reduce dimensionality for visualization purposes.

### 4.2 Cluster Plots

Clusters are visualized using 2D scatter plots, enabling an intuitive understanding of cluster structure and separation.

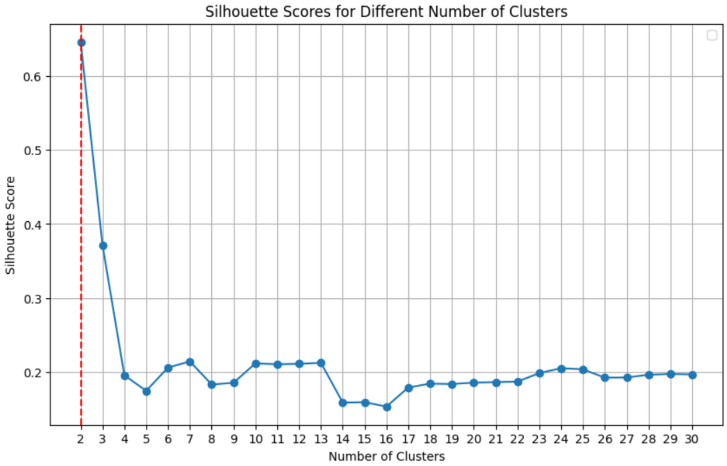

Figure 2: Silhouette Scores for different number of clusters

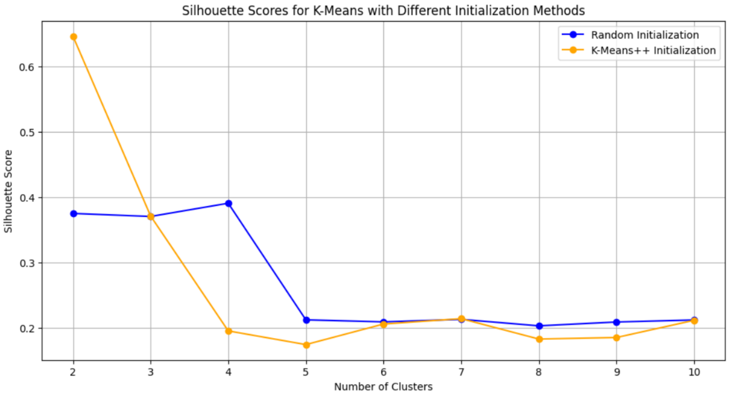

Figure 3: Silhouette Scores

# 5 Cluster Evaluation Metrics

Additional metrics, including the Davies-Bouldin Index and Calinski-Harabasz Index, are calculated for a thorough evaluation of cluster quality. These metrics validate the optimal number of clusters suggested by the Elbow Method and silhouette score.

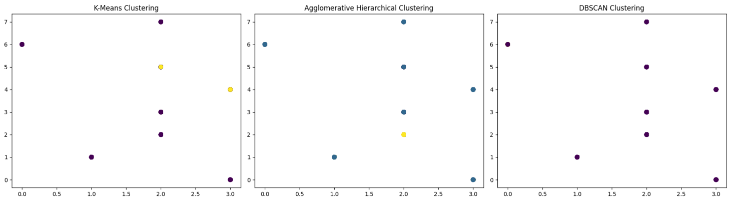

Figure 4: Clustering Results

# 6 Comparison with Other Clustering Algorithms

Agglomerative Clustering and DBSCAN are applied to the dataset. A comparison with K-Means is presented, discussing each algorithm's strengths and weaknesses for this dataset.

# 7 Analysis and Report

## 7.1 Summary of Clustering Process

The overall clustering process is summarized, including insights from cluster visualizations and evaluation metrics.

## 7.2 Limitations of K-Means and Other Algorithms

The limitations of K-Means and other clustering algorithms are discussed in terms of their applicability to this dataset, especially concerning noise sensitivity and cluster shape assumptions.

# 8 Conclusion

K-Means provided effective clustering of frog calls, validated through multiple evaluation metrics. Future work could explore alternative clustering algorithms or further optimize feature engineering for improved separation.