CSCI4150U: Data Mining

K-Means and Hierarchical Clustering

Syed Naqvi
Student ID: 100590852

November 9, 2024

Abstract

blablabla

1 Introduction

1.1 Methodology

Datasets used in this analysis are sourced from the *UC Irvine Machine Learning Repository* and include:

- **Breast Cancer Wisconsin (Diagnostic)**
- **Waveform Database Generator (Version 1)**

We evaluate the following clustering algorithms:

- **K-means Clustering**
- **Hierarchical Clustering** (Single Link, Complete Link, and Group Average)

For K-means clustering, model k values range from 1 to 6 for the *Breast Cancer Wisconsin (Diagnostic)* dataset and 2 to 6 for the *Waveform Database Generator (Version 1)* dataset. Clustering performance is assessed using the Sum of Squared Errors (SSE) with Euclidean distance as the metric.

1.2 Preprocessing

To perform accurate clustering, we analyze feature ranges to determine the dataset most needing of standardization.

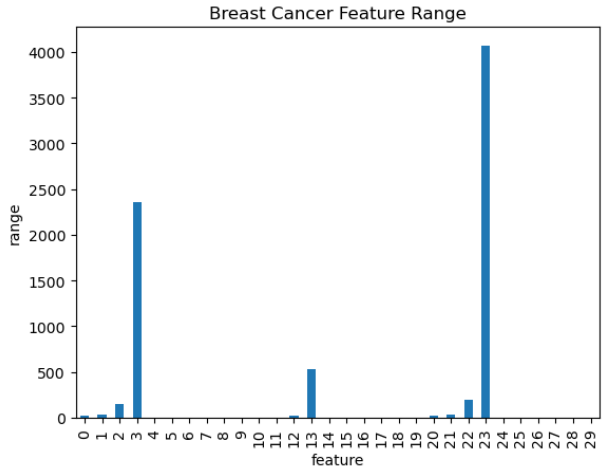

Figure 1: Pre-Standardized Feature Ranges (Breast Cancer Data)

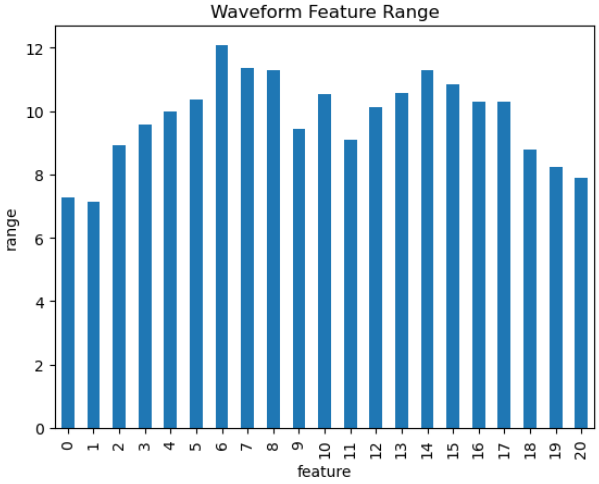

Figure 2: Pre-Standardized Feature Ranges (Waveform Data)

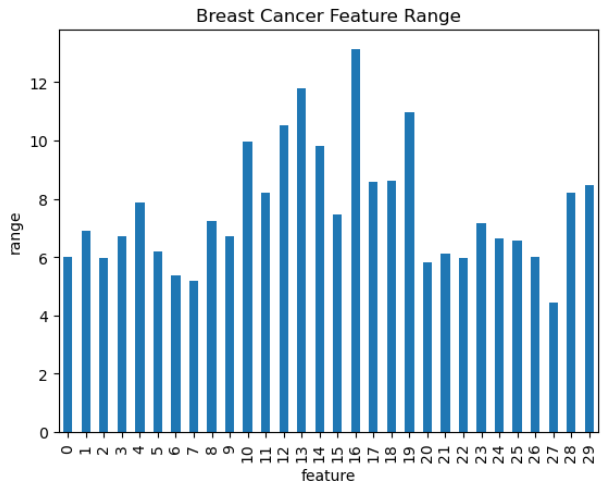

Figure 3: Post-Standardized Feature Ranges (Breast Cancer Data)

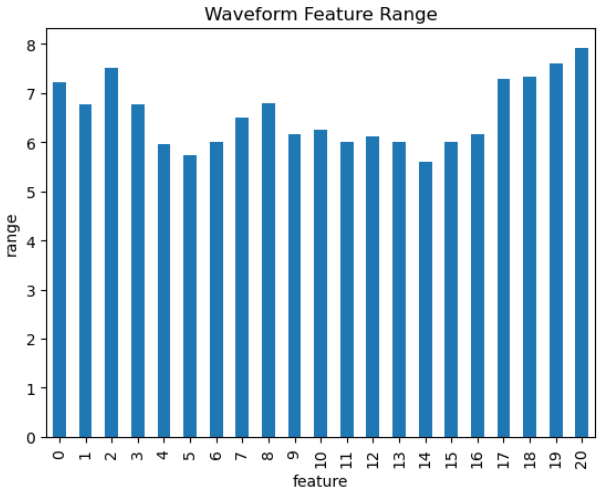

Figure 4: Post-Standardized Feature Ranges (Waveform Data)

The feature values have now been scaled and are significantly better suitable for clustering.

2 Part I: K-Means Clustering

2.1 Model Selection

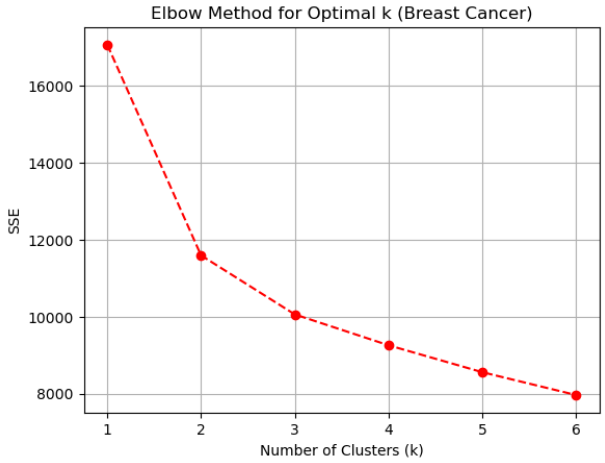

Figure 5: Elbow method for selecting optimal K-value (Breast Cancer Dataset)

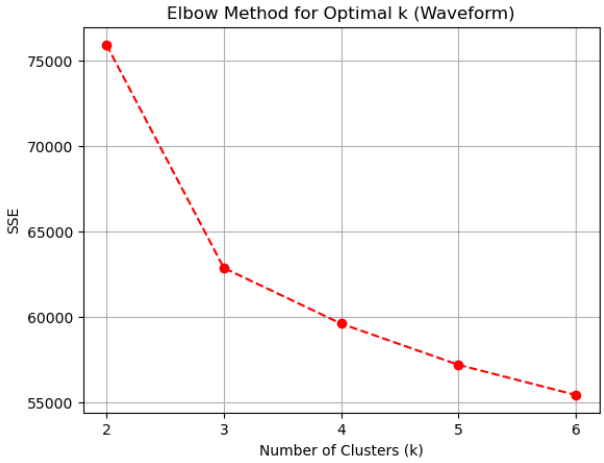

Figure 6: Elbow method for selecting optimal K-value (Waveform Dataset)

From the above figures, the ideal number of clusters is **k=2** for the Breast Cancer dataset and **k=3** for the Waveform dataset. This suggests there are likely 3 distinct waveforms in the Waveform dataset and two distinct tumor categorizations (malignant or benign) in the Breast Cancer dataset which is of course consistent with the actual number of unique labels in both datasets.

2.1.1 IBk Classification

Table 1 presents the accuracy of IBk classification for K values 1, 3, and 5 across the three datasets.

Table 1: Accuracy of IBk Classification

Dataset	**K=1**	**K=3**	**K=5**
letter	96.03	95.62	95.52
segment	97.14	96.02	95.06
waveform-5000	73.62	77.7	78.94

2.1.2 J48 and AdaBoostM1 Classification

Table 2 summarizes the accuracy results for J48 with M values of 2 and 4, and AdaBoostM1 with J48 (M=2) as the base classifier.

Table 2: Accuracy of J48 and AdaBoostM1 Classification

Dataset	**J48 (M=2)**	**J48 (M=4)**	**AdaBoostM1+J48 (M=2)**
letter	87.98	86.56	95.54
segment	96.93	96.06	98.53
waveform-5000	75.08	75.82	80.68

2.1.3 NaiveBayes Classification

Table 3 provides the accuracy results for NaiveBayes classification on each dataset.

Table 3: Accuracy of NaiveBayes Classification

Dataset	**Accuracy (%)**
letter	64.12
segment	80.22
waveform-5000	80.00

3 Part II: Clustering Task

3.1 Methodology

For the clustering task, we use SimpleKMeans to cluster each dataset with specified K values. We use the training set and evaluate each K-means model by measuring the Sum of Squared Errors (SSE). The chosen K values for each dataset are:

- **letter**: K = 11, 24, 38
- **segment**: K = 3, 5, 10
- **waveform-5000**: K = 2, 3, 5

3.2 Results

Table 4 shows the SSE results for SimpleKMeans clustering for each dataset with different K values.

Table 4: SSE of SimpleKMeans Clustering

Dataset	**K=K1**	**K=K2**	**K=K3**
letter	16513.56	9824.97	5132.69
segment	2173.14	1296.73	759.97
waveform-5000	5465.74	3846.65	3432.81

4 Conclusion

This report evaluates classification and clustering methods on three datasets using WEKA. IBk, J48, AdaBoostM1, and NaiveBayes are tested for classification, with AdaBoostM1 using J48 decision trees being the most accurate model across all datasets. SimpleKMeans clustering SSE is analyzed for different K values where the lowest error values are associated with the highest k values for each dataset.