Clustering the Iris Dataset Using the SEMMA Methodology

Aditya Rajpurohit

October 12, 2024

Abstract

The Iris dataset is one of the most well-known datasets in the field of machine learning. In this research, we apply the SEMMA methodology (Sample, Explore, Modify, Model, and Assess) to perform K-Means clustering on the Iris dataset. The study aims to identify natural groupings within the data by implementing K-Means clustering, evaluating the results using the Elbow Method and Silhouette Score, and visualizing the clusters. The SEMMA methodology provides a structured framework for data analysis and model building.

1 Introduction

The Iris dataset, introduced by Fisher in 1936, contains 150 observations of iris flowers from three species: Setosa, Versicolor, and Virginica. Each observation includes four features: sepal length, sepal width, petal length, and petal width. Clustering is a powerful technique for discovering hidden patterns and relationships in datasets. The SEMMA methodology, which stands for Sample, Explore, Modify, Model, and Assess, offers a systematic approach to data analysis and model building.

2 SEMMA Methodology

The SEMMA methodology consists of five steps:

1. **Sample:** Load and inspect the dataset.
2. **Explore:** Understand the distribution of data through statistical summaries and visualization.
3. **Modify:** Prepare the data for clustering by normalization and feature selection.
4. **Model:** Apply K-Means clustering to identify natural groupings in the dataset.
5. **Assess:** Evaluate the clustering results using metrics such as the Silhouette Score and visualize the clusters.

3 Data Sampling

The dataset contains 150 records of iris flowers with four features and a target variable indicating the species. In this step, the dataset is loaded and basic information such as column names, data types, and missing values is checked. No missing values were found, making the data suitable for further analysis.

```
# Python Code
import pandas as pd
iris_df = pd.read_csv(’IRIS.csv’)
iris_df.info()
```

4 Data Exploration

Exploration involves generating statistical summaries and visualizations to understand the distribution and relationships between features. A pair plot is used to visualize relationships between variables, and a correlation heatmap helps identify multicollinearity among features.

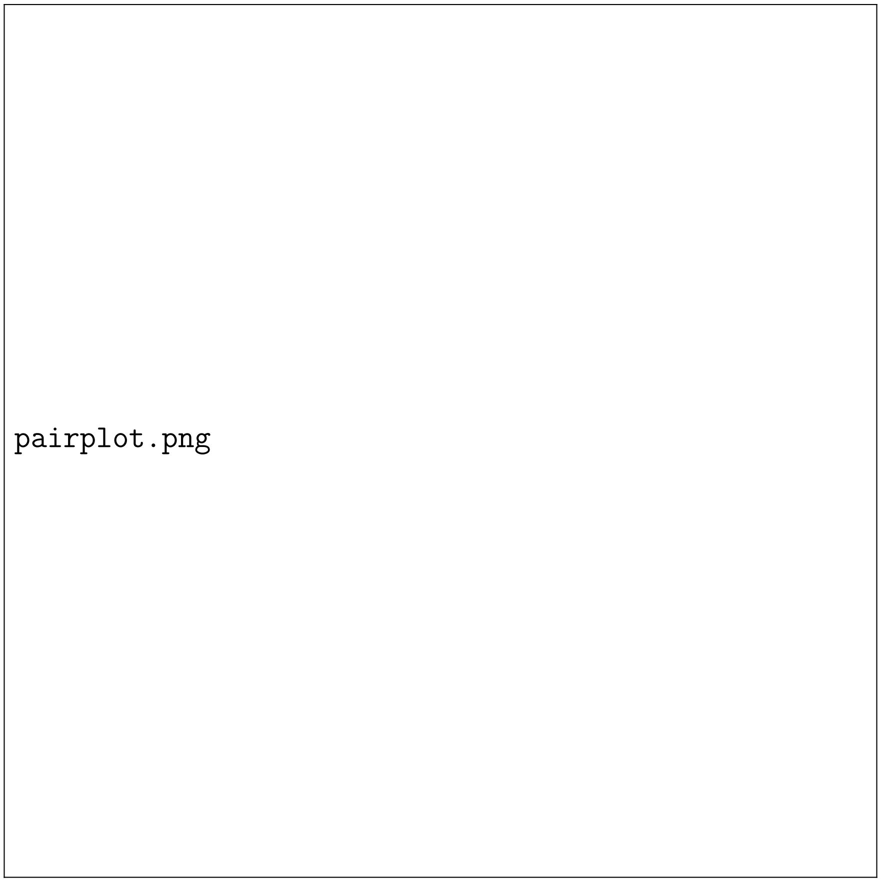

Figure 1: Pair plot of Iris dataset

The pair plot (Fig. 1) reveals clear separations among some species based on petal length and width, indicating potential clusters.

5 Data Modification

Normalization is performed to ensure that all features contribute equally to the clustering process. The target column is removed before applying K-Means clustering.

```
# Python Code
from sklearn.preprocessing import StandardScaler
X = iris_df.drop('species', axis=1)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

6 Modeling

K-Means clustering is applied to the normalized data. The optimal number of clusters is determined using the Elbow Method and Silhouette Score.

```
# Python Code
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

cluster_range = range(2, 11)
inertia_list = []
silhouette_list = []

for k in cluster_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    inertia_list.append(kmeans.inertia_)
    silhouette_list.append(silhouette_score(X_scaled, kmeans.labels_))
```

elbow_silhouette.png

Figure 2: Elbow Method and Silhouette Score for different cluster sizes

The Elbow Method and Silhouette Score (Fig. 2) suggest that three clusters are optimal, aligning well with the original classification.

7 Model Assessment

The clustering results are evaluated using the Silhouette Score and visualized through a pair plot with the predicted clusters.

```
# Python Code
optimal_k = 3
kmeans_optimal = KMeans(n_clusters=optimal_k, random_state=42)
```

```
iris_df[’Cluster’] = kmeans_optimal.fit_predict(X_scaled)


# Silhouette Score
silhouette_avg = silhouette_score(X_scaled, iris_df[’Cluster’])
print(f’Silhouette Score: {silhouette_avg}’)
```

cluster_pairplot.png

Figure 3: Clustered pair plot of Iris dataset

The clustered pair plot (Fig. 3) shows a good separation among clusters, with a silhouette score indicating satisfactory clustering quality.

8 Conclusion

This research demonstrated the application of the SEMMA methodology to the Iris dataset using K-Means clustering. The study identified three natural clusters that closely align with the actual species classification. The SEMMA approach provides a structured framework that enhances model-building and evaluation processes. Future work may explore alternative clustering algorithms or apply this approach to other datasets.

References

[1] Fisher, R.A. (1936). ”The use of multiple measurements in taxonomic problems.” *Annals of Eugenics*, 7(2), 179-188.

[2] SAS Institute Inc. (1998). ”Introducing the SEMMA Methodology.” SAS Institute White Paper.