COMPARATIVE ANALYSIS OF CLUSTERING METHODS ON BREAST CANCER, DIGIT, AND IRIS DATASETS

Abstract

The three popular clustering techniques K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and agglomerative clustering are examined and compared in this study using three different datasets: Breast Cancer, Digit, and Iris. A crucial unsupervised machine learning method known as clustering divides linked data points into distinct clusters based on their inherent patterns and similarities. K-means clustering is used for the Breast Cancer dataset to separate benign and malignant cases based on patient feature profiles. To find groups of handwritten digits that are similar to one another and find possible outliers, DBSCAN is performed to the Digit dataset. On the Iris dataset, aggregative clustering is used to examine correlations between various iris flower species based on their characteristics.

The evaluation's findings shed light on how well each strategy clustered the relevant datasets. For the Breast Cancer dataset, K-means obtains relatively well-defined clusters, however DBSCAN finds it difficult to discover meaningful clusters in the high-dimensional Digit dataset, resulting in a low silhouette score. In the Iris dataset, a good degree of cluster separation is shown through agglomerative clustering.

The study emphasises how important it is to comprehend dataset features and choose the right clustering methods and settings in order to get useful insights. Additionally, it highlights how crucial it is to investigate various clustering methods for a variety of datasets in order to make sensible judgements for practical applications.

Table of Contents

Table of Figures

Introduction

Unsupervised machine learning's primary clustering approach groups related data points into separate clusters based on their inherent patterns and similarities. In this context, K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and agglomerative clustering are three widely used clustering techniques that will be examined and discussed. These methods are essential for resolving the clustering issue for the three different datasets, Iris, Digit, and Breast Cancer.

K Means

The popular partition-based clustering technique K-means is renowned for being straightforward and effective. In order to reduce the squared Euclidean distance between the data points and the centroid, it seeks to split the data into K clusters, each of which is represented by its centroid. K-means will attempt to group patients with similar feature profiles for the Breast Cancer dataset, perhaps helping to discriminate between benign and malignant instances. We may learn more about the Breast Cancer dataset's capacity to identify distinctive patient characteristics, perhaps aiding in cancer diagnosis and therapy planning.

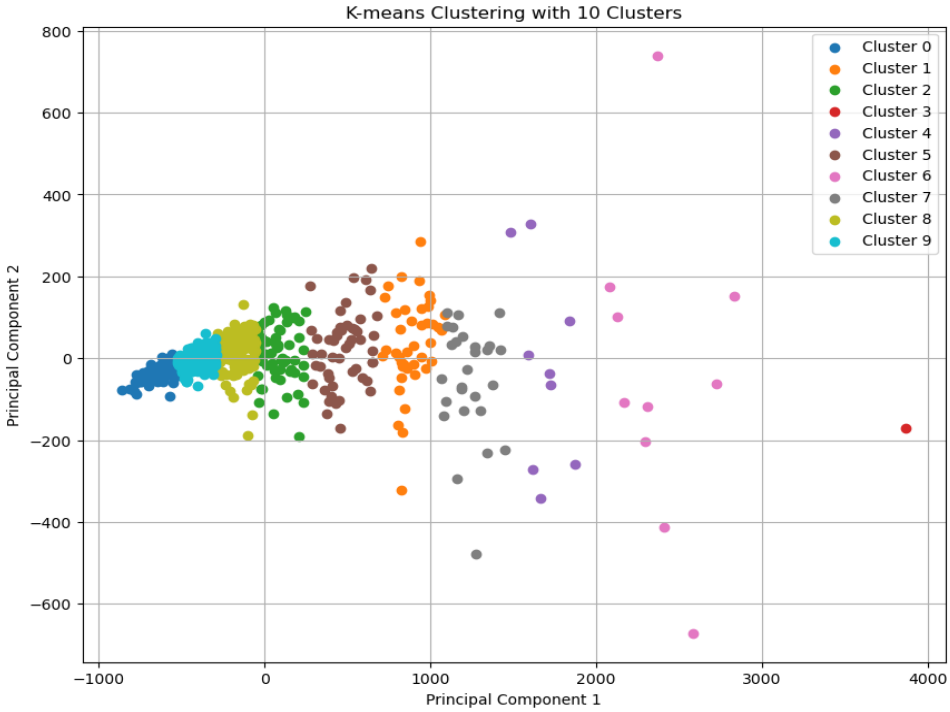

Figure 1: K Means Clustering with 10 clusters

DBSCAN

DBSCAN is a density-based clustering technique that is excellent at finding clusters of various sizes and shapes in data and at recognising outliers as noise. DBSCAN successfully distinguishes between dense and sparse regions by evaluating data density in local neighbourhoods, which enables it to identify

complicated data structures. DBSCAN may be used to find clusters of handwritten digits that are similar by applying it on the Digit dataset. Additionally, it has the ability to automatically identify abnormalities that might be related to ambiguous or outlier samples. This investigation allows us to see how well DBSCAN's density-based strategy handles digit identification and noise detection.

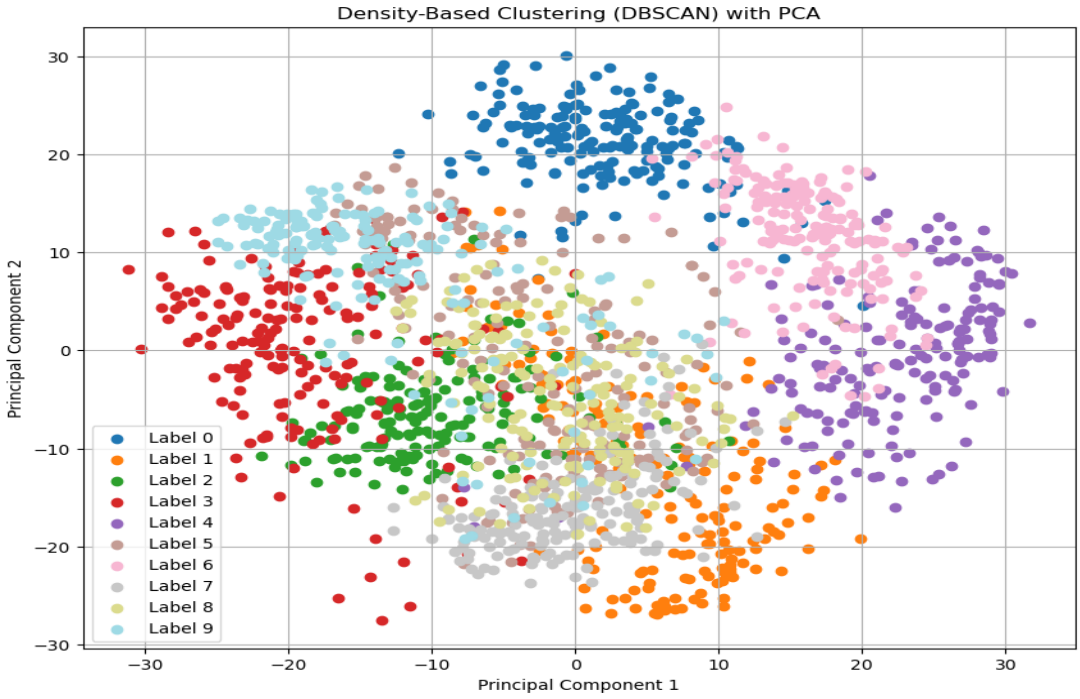

Figure 2: Density Based Clustering (DBSCAN) with PCA

Agglomerative Clustering

Each data point begins as its own cluster in the hierarchical bottom-up technique known as agglomerative clustering, and when other clusters get closer to it, they gradually merge with it. It produces a dendrogram, a tree-like structure that gives a visual representation of the hierarchy of the data. Agglomerative Clustering will try to identify associations between various species of iris blossoms based on their characteristics in the context of the Iris dataset. The hierarchical structure of the dataset may be better understood using this technique, which can also highlight the similarities and differences between the various iris species.

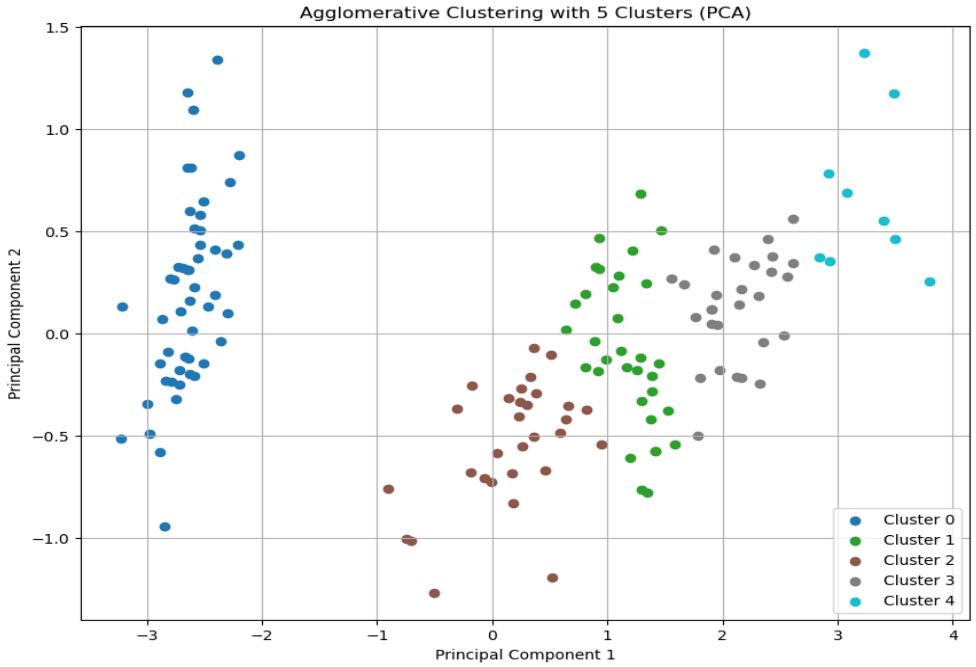

Figure 3: Agglomerative Clustering with Clusters (PCA)

Comparison of Results

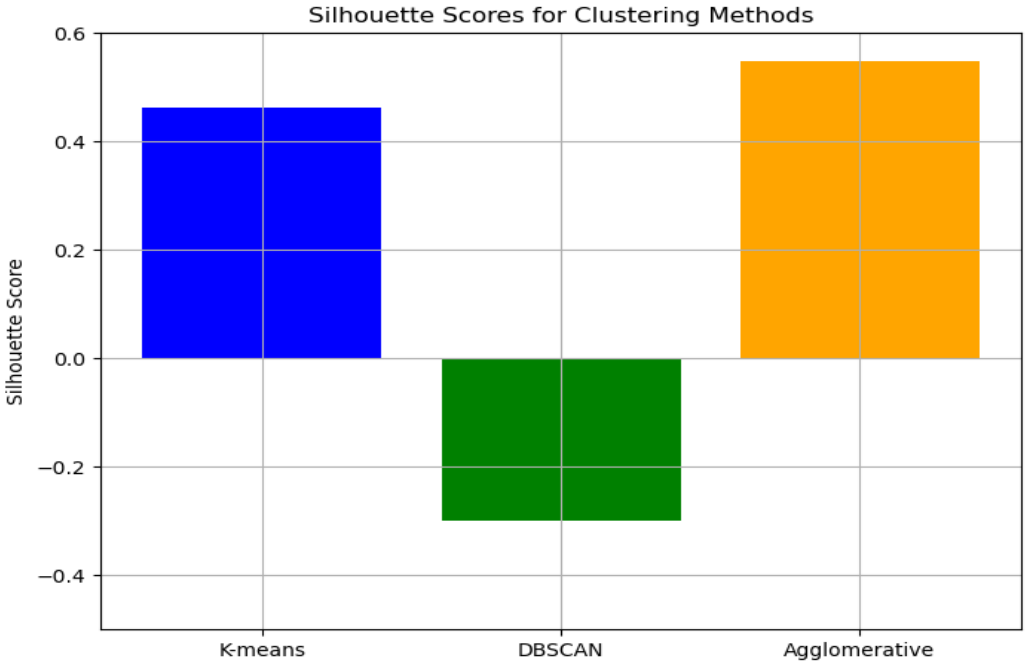

Figure 4: Sihouette Scores for Clustering Methods

Part 1: K-means Clustering of a Breast Cancer Dataset

Used Principal Component Analysis (PCA) to decrease the dimensionality of the 30 original characteristics in the Breast Cancer dataset to 2 principal components. The data were then divided into 10 groups using the K-means clustering technique. The outcomes were as follows:

K-means Silhouette Score: 0.46276849398282627

K-means Inertia: 35.843585237258345

The clarity of the clusters is gauged by the silhouette score. A score around 0 denotes overlapping or poorly formed clusters, whereas a score near 1 denotes well-separated clusters. The clusters may not be completely different in this example because the silhouette score of 0.46 shows a considerable separation between the clusters.

The within-cluster sum of squared distances, sometimes referred to as the inertia, gauges how compact the clusters are. More compact and well separated clusters are indicated by lower inertia levels. It is vital to keep in mind that this value cannot be immediately interpreted without a reference, despite the fact that the inertia of 35.84 shows that the data points inside each cluster are rather near to their cluster's centroid.

Part 2: Digits Dataset with DBSCAN Clustering

Instead of using K-means for clustering in the Digits dataset, we employed DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The density of the data points used in DBSCAN, a density-based approach, is used to identify clusters. The outcomes are as follows:

DBSCAN Silhouette Score: -0.29948366468262777 The DBSCAN silhouette score is negative, meaning that many of the data points may be ignored as noise or outliers and that clusters are poorly formed. The high dimensionality of the Digits dataset makes it difficult for DBSCAN to properly find clusters, hence it is not a good fit for such datasets. A low silhouette score in this instance indicates that DBSCAN was unsuccessful in locating significant clusters in the Digits dataset and is thus unsuitable for use with this particular dataset.

Part 3: Agglomerative Clustering Using the Iris Dataset

We used the Agglomerative Clustering technique with 5 clusters to apply hierarchical clustering to the Iris dataset. Based on data closeness, hierarchical clustering creates a tree-like structure of layered clusters. These are the outcomes that were attained:

Silhouette Score for Agglomerative Clustering: 0.5487843719847739

Agglomerative Clustering's silhouette score is 0.55, which shows that the clusters are only fairly well-separated. The Iris dataset's significant clusters may have been found via agglomerative clustering, according to the positive silhouette score. It is important to remember that the choice of distance metric and connection criterion might have an impact on how well Agglomerative Clustering performs.

Conclusion

Finally, the outcomes of the clustering techniques differed among the three datasets. The Breast Cancer dataset's K-means clustering produced a few relatively well-defined clusters, suggesting considerable distinction between the data points. On the other hand, DBSCAN on the Digits dataset struggled to properly handle high-dimensional data since it was unable to find any significant clusters, which resulted in a negative silhouette score. Agglomerative Clustering on the Iris dataset, which finally obtained a reasonable level of cluster separation, suggests that it was able to recognise separate groups among the iris species.

The clustering method chosen and its settings have a big influence on how well the results turn out. In order to derive actionable insights from the data, it is crucial to comprehend the dataset's features, take into account its dimensionality, and experiment with various clustering techniques.

References

D. Deng, "DBSCAN Clustering Algorithm Based on Density," *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Hefei, China, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.

D. Jain, M. Singh and A. K. Sharma, "Performance enhancement of DBSCAN density based clustering algorithm in data mining," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017, pp. 1559-1564, doi: 10.1109/ICECDS.2017.8389708.

F. AlMahamid and K. Grolinger, "Agglomerative Hierarchical Clustering with Dynamic Time Warping for Household Load Curve Clustering," *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Halifax, NS, Canada, 2022, pp. 241-247, doi: 10.1109/CCECE49351.2022.9918481.

H. Xu, S. Yao, Q. Li and Z. Ye, "An Improved K-means Clustering Algorithm," 2020 IEEE 5th International Symposium on Smart and Wireless Systems within the Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS), Dortmund, Germany, 2020, pp. 1-5, doi: 10.1109/IDAACS-SWS50031.2020.9297060.

K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

M. H. Chehreghani, "Reliable Agglomerative Clustering," *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534228.