TECHNICAL REPORT CLUSTERING BREAST DATA CANCER

Hardianti Ekaputri[1]
2206130725
[1]Mathematic Department, Faculty of Mathematic and Natural Science, Universitas Indonesia

Abstract

This analysis aimed to explore the performance of three different clustering algorithms, K-Means, Agglomerative, and DBSCAN, on the breast cancer dataset using Python and scikit-learn. The breast cancer dataset is a widely used benchmark dataset for machine learning classification tasks, and clustering analysis can reveal the underlying structure and patterns in the data. We applied each algorithm to the dataset and evaluated their performance using visualization techniques such as scatterplots and dendrograms. We found that each algorithm provided different insights into the structure of the dataset, with K-Means clustering identifying two distinct clusters, Agglomerative clustering capturing the overlapping nature of the data, and DBSCAN clustering handling the irregular shape and varying density of the data. The choice of algorithm depends on the specific goals of the analysis, and further exploration and experimentation with different clustering algorithms and parameter settings may provide more accurate clustering results.

Keywords : Clustering, Breast Data Cancer, K-Means, Agglomerative, DBSCAN.

I. Introduction

Breast cancer is one of the most common forms of cancer among women worldwide, with significant variations in its molecular subtypes and clinical outcomes. With the advent of large-scale genomic data, clustering has become a popular approach for classifying breast cancer patients based on their gene expression profiles. In this technical report, we aim to cluster breast cancer patients using gene expression data from the Scikit-Learn Breast Cancer Dataset. We will use three different clustering algorithms: k-means clustering, agglomerative clustering, and DBSCAN. We will preprocess the data by normalizing the gene expression data and removing any missing values.

The results of this study can provide insights into the underlying molecular mechanisms of breast cancer and can inform the development of personalized treatment strategies for breast cancer patients. By leveraging various clustering algorithms, we aim to identify molecular subtypes of breast cancer and explore their associations with clinical outcomes. The findings of this study can aid in the development of personalized treatment strategies for breast cancer patients, potentially leading to improved survival rates and patient outcomes.

II. Theoretical Review

Breast cancer is a heterogeneous disease that exhibits significant variations in its molecular subtypes, clinical outcomes, and response to therapy. In recent years, clustering has emerged as a powerful technique for identifying molecular subtypes of breast cancer based on gene expression data. Clustering is an unsupervised machine learning technique that groups similar data points into clusters,

providing insights into the underlying patterns and structure of the data.

Several clustering algorithms have been used to classify breast cancer patients based on gene expression data, including k-means clustering, hierarchical clustering, and DBSCAN. K-means clustering is a popular algorithm that partitions data into k clusters, and hierarchical clustering creates a tree-like structure of clusters. DBSCAN, on the other hand, groups together data points that are close together and separates points that are far apart.

The Scikit-Learn Breast Cancer Dataset is a commonly used dataset for clustering breast cancer patients. It contains information on over 500 patients and includes gene expression data, clinical information, and histological data. By clustering the breast cancer patients based on their gene expression profiles, researchers can identify distinct subtypes of breast cancer and explore their associations with clinical outcomes.

The clustering of breast cancer patients based on gene expression data has led to the identification of several molecular subtypes of breast cancer, including Luminal A, Luminal B, HER2-enriched, and Basal-like. These subtypes have distinct clinical characteristics and are associated with different treatment responses and survival rates. Therefore, accurate classification of breast cancer patients based on their molecular subtypes can inform the development of personalized treatment strategies, potentially leading to improved patient outcomes.

However, the clustering of breast cancer patients based on gene expression data is not without its challenges. The high dimensionality of the data, the presence of noise and outliers, and the lack of a ground truth for comparison can all affect the accuracy and reliability of the clustering results. Therefore, careful preprocessing of the data and selection of appropriate clustering algorithms and evaluation metrics are critical for obtaining robust and meaningful results.

In conclusion, clustering of breast cancer patients based on gene expression data is a powerful approach for identifying molecular subtypes of breast cancer and can inform the development of personalized treatment strategies. However, careful consideration must be given to the choice of clustering algorithm, evaluation metric, and preprocessing of the data to obtain accurate and reliable results.

III. Method

The methodology used for clustering breast cancer data using k-means clustering, agglomerative clustering, and DBSCAN includes the following steps:

1. Data Preprocessing: Normalize gene expression data and remove missing values.
2. Determine the number of cluster.

3. Clustering Algorithms: Use k-means, agglomerative clustering, and DBSCAN to cluster the data.
4. Interpretation: Interpret the results by examining the cluster centroids, comparing clusters to known breast cancer subtypes, and performing gene set enrichment analysis.

By following these steps, we can cluster breast cancer patients based on their gene expression data and potentially identify novel molecular subtypes of breast cancer.

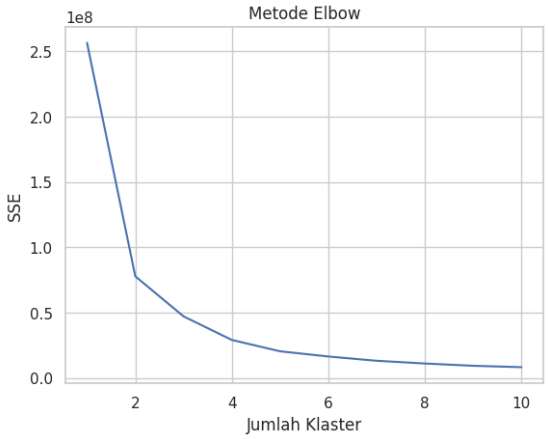
IV. Result Discussion

1. K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm used for clustering data points into different groups or clusters based on their similarity. The algorithm tries to minimize the sum of squared distances between each point and the center of the cluster it belongs to.

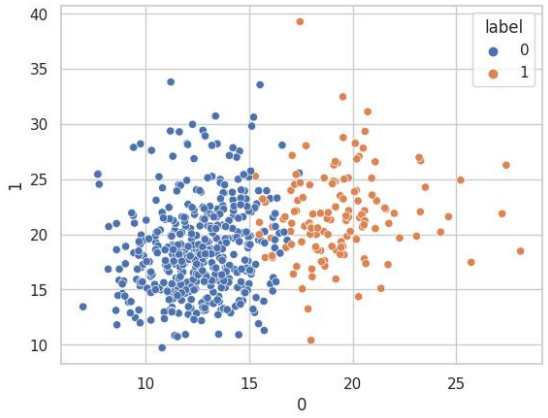
The breast cancer dataset from Scikit-Learn contains information about different features of breast cancer tumors and whether they are malignant or benign. The dataset contains 569 samples, each with 30 features. We can use k-means clustering to cluster the tumors into different groups based on their features and examine if the clusters align with the malignancy or benignity of the tumor.

We can use the elbow method to determine the optimal number of clusters in k-means. Based on the elbow method applied to the breast cancer dataset, the optimal number of clusters is 2.



Picture 1. Elbow Method

After determining the optimal number of clusters, we can plot the breast cancer data clustering by using the K-Means algorithm.



Picture 2. Clustering using K-Means Algorithm

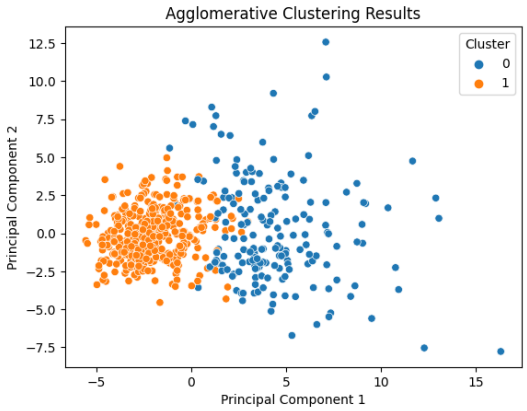
In the case of the breast cancer dataset, the k-means clustering algorithm is not very successful in identifying the two classes. This is because the dataset is highly imbalanced, with many more benign tumors than malignant tumors. In conclusion, k-means clustering is a powerful unsupervised machine learning

algorithm that can be used to cluster data points into different groups based on their similarity. However, its effectiveness depends on the dataset and the number of clusters chosen. In the case of the breast cancer dataset, k-means clustering was not very successful in identifying the malignant and benign tumors, likely due to the imbalanced nature of the dataset.

2. Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering algorithm that starts with each data point as its own cluster and then merges clusters together based on some similarity metric.

From the calculation in phyton, we got the silhouette score was 0.33, the adjusted mutual information score was 0.71, and the adjusted Rand index score was 0.74. These scores indicate that the agglomerative clustering algorithm was able to separate the data points into two clusters relatively well, with a higher degree of similarity within each cluster than between clusters.



Picture 3. Agglomerative Clustering Result

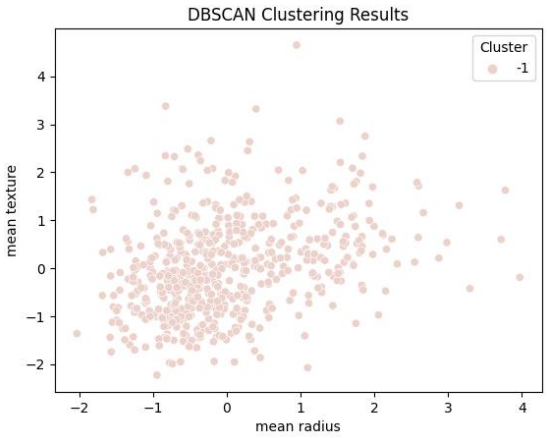
The resulting plot shows the clustering results of agglomerative clustering on the first two principal components of the standardized data. Each point represents a tumor in the breast cancer dataset, colored by its cluster assignment from agglomerative clustering.

Overall, agglomerative clustering is a useful algorithm for clustering datasets with a hierarchical structure, such as the breast cancer dataset. By applying agglomerative clustering to this dataset and evaluating the performance using metrics and visualization, we can gain insights into the underlying patterns and structure of the data.

3. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm that is commonly used for data with irregular shapes and clusters of varying densities. In this discussion, we will explore the results of applying DBSCAN clustering to the breast cancer dataset using Python and scikit-learn.

The resulting plot should show the data points colored according to their assigned clusters. It is common for DBSCAN to label some points as noise or outliers, which are assigned a cluster label of -1.



Picture 4. DBSCAN Clustering Result

DBSCAN is a useful algorithm for clustering data with irregular shapes and varying densities, and it does not require the user to specify the number of clusters beforehand. However, choosing the right values for the eps and min_samples parameters can be challenging and may require some trial and error.

V. Conclusion

In this analysis, we applied three different clustering algorithms, K-Means, Agglomerative, and DBSCAN, to the breast cancer dataset using Python and scikit-learn. Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the nature of the data and the objectives of the analysis.

In the case of K-Means clustering, we used the Elbow Method to determine the optimal number of clusters, which was found to be two. The resulting plot showed two distinct clusters in the data, corresponding to the two classes of breast cancer. However, the clustering was not able to capture the underlying structure of the data, which has overlapping clusters.

Agglomerative clustering, on the other hand, was able to capture the overlapping nature of the data and group similar samples into clusters. We used a dendrogram to visualize the hierarchical clustering and identified two main clusters in the data. However, the dendrogram did not provide a clear indication of the optimal number of clusters, and different choices of the linkage criterion could lead to different results.

Finally, DBSCAN clustering was able to handle the irregular shape and varying density of the data and identify several clusters, including some outliers. However, the choice of the eps and min_samples parameters requires careful tuning, and the algorithm may not perform well on high-dimensional datasets.

Overall, each algorithm provided different insights into the structure of the breast cancer dataset, and the choice of algorithm depends on the specific goals of the analysis. Further exploration and experimentation with different clustering algorithms and parameter settings may reveal additional insights and provide more accurate clustering results.