AI351 - Machine Learning Lab

Taiwanese Bankruptcy Statistics

Ali Hasan Khan
Artificial Intelligence(FCSE)
Ghulam Ishaq Khan Institute of Engineering and Technology
Topi,Swabi, Pakistan
u2021079@giki.edu.pk

I. Introduction

The application is primarily focused on the application of various clustering techniques on a dataset. Let's delve into the details and understand its functionalities and operations.

- Data Preprocessing:

We began by importing necessary libraries, including pandas, numpy, and matplotlib, and then loads a dataset from a CSV file. It drops a specific column ('Net Income Flag') from the dataset. Further preprocessing steps involve handling missing values by replacing them with the mean of the corresponding column. The data is then standardized using the 'StandardScaler' from the 'sklearn.preprocessing' module.

- Data Visualization and Dimensionality Reduction:

WE employed two popular dimensionality reduction techniques, namely Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), to reduce the dimensionality of the data to two dimensions. It then visualizes the reduced data using scatter plots to analyze the clusters' distribution in the reduced feature space.

- K-Means Clustering:

We applied the K-Means clustering algorithm to the preprocessed and scaled data. It initializes the K-Means algorithm with a specified number of clusters and generates cluster labels for each data point. We also calculates various clustering evaluation metrics, including Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Normalized Mutual Info, Adjusted Rand Index, Adjusted Mutual Info, V-Measure, Completeness Score, and Homogeneity Score, to assess the performance of the clustering algorithm.

- Generalized Clustering Evaluation:

Then we defines a function to evaluate various clustering metrics for a given set of clustering models. It iterates over different clustering algorithms, including K-Means, Agglomerative Clustering, and Spectral Clustering, and assesses the models' performance using the defined evaluation metrics.

- Visualization of Clustering Results:

The script provides a function to visualize the clustering results for specific clustering models, such as DBSCAN and Mean Shift. The function plots the clusters formed by the respective clustering algorithms, enabling the visual interpretation of the clustered data points.

- Affinity Propagation Clustering:

We also implements the Affinity Propagation clustering algorithm, specifying the damping and preference parameters. It fits the scaled data to the Affinity Propagation model and retrieves the cluster labels assigned to each data point.

- Further Dimensionality Reduction Visualization:

The script leverages PCA and t-SNE once again to visualize the clustered data points in the reduced two-dimensional feature space, facilitating a better understanding of the clusters' distribution and separation based on the assigned labels.

Overall, We demonstrates the implementation and evaluation of various clustering techniques, along with visualization and assessment of the clustering results. It provides valuable insights into the dataset's underlying structure and the effectiveness of different clustering algorithms in identifying distinct clusters within the data.

II. TSNE and PCA

In this we snippet focuses on the implementation of dimensionality reduction techniques, namely Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), on the scaled data. Let's discuss these operations in detail:

- Dimensionality Reduction using PCA and t-SNE:
- Principal Component Analysis (PCA):

PCA is a widely used linear dimensionality reduction technique that aims to capture the maximum variance in the data by projecting it onto a lower-dimensional subspace. Also PCA is applied to the 'scaled-data' with the specified number of components as 2, transforming the data into a 2-dimensional space ('pcaresult').

- t-distributed Stochastic Neighbor Embedding (t-SNE):

t-SNE is a nonlinear dimensionality reduction technique that focuses on preserving the local structure of data points in a lower-dimensional space. Here, the t-SNE algorithm is applied to the 'scaled-data' with 2 components, generating a 2-dimensional representation of the data ('tsne-result'). Data Visualization:

- PCA Visualization:

The first subplot displays the data points transformed using PCA, with the x-axis representing the first principal component

978-1-6654-5992-1/22/$31.00 ©2022 IEEE

and the y-axis representing the second principal component. The scatter plot illustrates the distribution of data points in the 2-dimensional PCA space.

- t-SNE Visualization:

The second subplot shows the data points transformed using t-SNE, with the x-axis and y-axis representing the two t-SNE components. This scatter plot depicts the distribution of data points in the 2-dimensional t-SNE space.

- Interpretation:

The visualization of the data in the reduced 2-dimensional space provides insights into the data's underlying structure and potential patterns. It enables the understanding of the data's distribution and the relationships between data points in a lower-dimensional representation. Both PCA and t-SNE play a crucial role in exploratory data analysis, as they facilitate the visualization of complex datasets, especially when dealing with high-dimensional data.

III. SCORES

Certainly! We calculates various evaluation metrics to assess the performance of the clustering algorithm used. Understanding these metrics and their implications can provide valuable insights into the quality of the clustering process and the efficacy of the algorithm in accurately partitioning the data points.

We utilizes various evaluation metrics that aid in assessing the clustering results. These metrics are crucial in determining the effectiveness of the clustering algorithm in separating and grouping the data points.

- Silhouette Score

This metric quantifies the cohesion and separation of clusters. It computes the mean silhouette coefficient for all samples, indicating how similar each sample is to its own cluster compared to other clusters. A higher silhouette score reflects better-defined clusters.

- Calinski-Harabasz Score

Also known as the Variance Ratio Criterion, this metric calculates the ratio of the sum of between-cluster dispersion and the within-cluster dispersion. A higher score suggests better-defined, dense, and well-separated clusters.

- Davies-Bouldin Score

This metric measures the average similarity between each cluster's elements. It provides information on the clustering's compactness and separation, where lower values indicate better clustering.

- Normalized Mutual Info (NMI)

NMI measures the mutual information between two clusterings, considering the ground truth. It is a normalized value between 0 and 1, where higher values indicate a better clustering.

- Adjusted Rand Index (ARI)

ARI computes the similarity between two clusterings, considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The ARI value ranges between -1 and 1, where 1 indicates perfect matching.

- Adjusted Mutual Info (AMI)

AMI measures the agreement between two clusterings, accounting for chance. It normalizes the mutual information score to account for the chance of seeing a particular clustering just by random.

- V-Measure V-Measure

is the harmonic mean of homogeneity and completeness. It provides a balanced measure that considers both aspects in the clustering process.

- Completeness Score:

This metric measures the completeness of the clustering with respect to the true labels. It indicates how many samples of the same class are present in a given cluster.

- Homogeneity Score

Homogeneity indicates whether clusters contain only data points from a single class. It provides insights into how well-defined the clusters are in terms of containing only samples from a single class.

The output presents the calculated values for each of the evaluation metrics:

- Silhouette Score 0.1472
- Calinski-Harabasz Score 54.2739
- Davies-Bouldin Score 2.2696
- Normalized Mutual Info 0.3828
- Adjusted Rand Index:** 0.4637
- Adjusted Mutual Info 0.3817
- V-Measure 0.3828
- Completeness Score 0.3854
- Homogeneity Score 0.3802

These scores collectively provide an overall assessment of the clustering algorithm's performance and the quality of the clusters formed, indicating how well the algorithm was able to identify and group similar data points while keeping distinct clusters separate. The results help in understanding the strengths and weaknesses of the clustering process, aiding in the interpretation of the algorithm's performance and the subsequent decision-making process.

IV. Evaluation of Clustering Models Using Various Metrics and Dimensionality Reduction Visualization

It performs an evaluation of various clustering models using different metrics to assess the effectiveness of each model in clustering the given dataset. The evaluation metrics used here are commonly employed in the field of machine learning and are crucial in determining the quality of the clusters formed.

It importing various necessary modules and functions from the 'sklearn.metrics' library to calculate different clustering evaluation metrics. It then defines a function 'evaluate-clustering-metrics' that takes in a clustering model, data,

and scaled data as input and computes the evaluation metrics for the given model. The function calculates metrics such as Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Normalized Mutual Info, Adjusted Rand Index, Adjusted Mutual Info, V-Measure, Completeness Score, and Homogeneity Score for the provided data and clustering labels.

Next initializes a list 'clustering models' that contains different clustering models, including KMeans, AgglomerativeClustering, and SpectralClustering, each with varying numbers of clusters. It then iterates over each model, evaluates the metrics using the 'evaluate clusteringmetrics' function, and prints the results for each model. The results provide valuable insights into the performance of each clustering model and the quality of the clusters formed based on the specified evaluation metrics.

- KMeans Clustering

The Silhouette Score is 0.1354, suggesting moderate cohesion and separation.The Calinski-Harabasz Score is 54.1903, indicating well-separated and dense clusters.The Davies-Bouldin Score is 2.2583, suggesting moderate cluster quality.The Normalized Mutual Info is 0.4084, implying a substantial level of mutual information between the clusters and the true labels.The Adjusted Rand Index is 0.5082, signifying a strong similarity between the predicted and true clusterings.The V-Measure, Completeness Score, and Homogeneity Score are all around 0.408, suggesting balanced performance.

- AgglomerativeClusterg

inThe Silhouette Score is approximately 0.1129, indicating moderate cohesion and separation.The Calinski-Harabasz Score is around 46.2812, suggesting well-separated and dense clusters.The Davies-Bouldin Score is 2.3981, implying slightly lower cluster quality compared to KMeans.The Normalized Mutual Info, Adjusted Rand Index, and other scores are also provided, indicating the performance of the AgglomerativeClustering model.

- SpectralClusterg

inThe output shows warnings related to the model not fully connecting and spectral embedding not working as expected.The Silhouette Score is relatively high at 0.5997, suggesting good cluster cohesion and separation.The Calinski-Harabasz Score is significantly lower compared to other models at 7.8738, indicating less well-separated and dense clusters.Other metrics, including Normalized Mutual Info and Adjusted Rand Index, are also provided, indicating the performance of the SpectralClustering model.

- Figures:

We generate visualizations for the clustering results using PCA and t-SNE for dimensionality reduction. These visualizations aid in understanding the distribution of data points in the reduced feature space and provide insights into the clustering performance in a lower-dimensional setting. Additionally, the plots with labeled clusters from the t-SNE and PCA results help in visualizing the separation between the clusters.

V. Clustering and Visualization using Affinity Propagation Algorithm and Dimensionality Reduction Techniques

In this we utilizes the Affinity Propagation algorithm for clustering and visualization purposes. The Affinity Propagation algorithm is a clustering algorithm that does not require the user to specify the number of clusters beforehand. It determines the number of clusters based on the data provided and the internal algorithmic mechanisms. We used the Affinity Propagation algorithm to generate cluster labels based on the provided dataset.

After obtaining the cluster labels, we employs dimensionality reduction techniques such as PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the dimensionality of the data to two dimensions. The purpose of this dimensionality reduction is to visualize the clusters in a lower-dimensional space, making it easier to interpret the relationships and patterns within the data.

The PCA and t-SNE results are then plotted in a single figure for comparison. The first subplot represents the clusters in the reduced feature space obtained using PCA, while the second subplot displays the clusters based on the t-SNE results. The clusters are differentiated by color, with blue representing one cluster and red representing another. Additionally, the legend provides clarity by labeling each cluster accordingly.

Overall, we perform clustering using the Affinity Propagation algorithm and subsequently visualizes the clusters using dimensionality reduction techniques, namely PCA and t-SNE. This visualization aids in understanding the distribution and relationships between data points in a lower-dimensional space, providing insights into the underlying patterns and structures within the dataset.