Data Mining and Discovery

Report

Veera Raghunatha Reddy Naguru – 22028322

Clustering Products from Sales Transactions Dataset using Graph Based Clustering algorithms and K-Means Clustering algorithm.

Abstract:

This report is prepared as part of my module coursework. I chose Sales Transactions Dataset Weekly as my dataset for performing clustering analysis. The objective to perform clustering analysis is to find products that are clustered together so depending on the trends and sales of the products the stocks can be well maintained. We used different clustering algorithms to cluster products in this dataset and observed the patterns of clustered products. Also, discussed the efficiency of used clustering algorithms for this dataset.

Introduction:

Sales Transaction Dataset Weekly dataset is available in UCI ML Repository. This dataset contains 811 products as records with information as number of sales happened in a week over a year and these weekly sold quantities are spread across columns. Along with these columns of data, the dataset also has the normalized data of these weekly sales.

Clustering is a process to group items from a list/dataset with items that exhibit same features related to an item. In our case, we have 811 products that need to be grouped together with similarity of products that exhibits same performance in sales over weeks.

Data Preprocessing:

The dataset contains 105 columns with a column including product labels and 52 columns of weekly sales and 52 columns of normalised data of weekly sales. For clustering we can use the normalised data directly, but due the need for building similarity matrices and SNN matrices, I preferred to use the data of 52 columns containing weekly sales in actual values of quantity.

Further to it, I performed some statistical analysis to understand the data and looked for any missing values. The dataset doesn't contain any missing values and all columns are in float datatype apart from product labels column. I removed the product labels column which leaves me with just the weakly sales data. It is a high dimensional data, and it will be tough to visualise clustered results, so I performed Principal Component Analysis with 2 features and the resulted dataset can be viewed below.

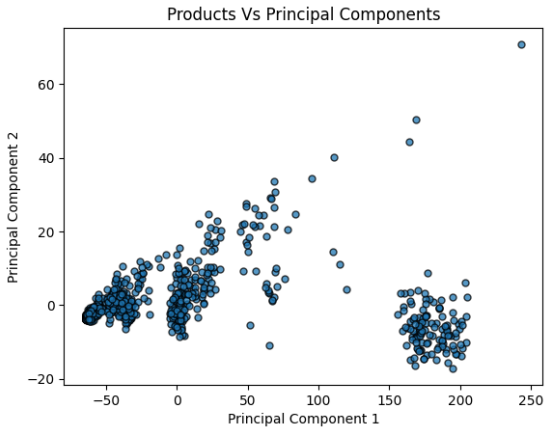

Figure 1:Products scattered over Principal Component Features

The resultant data is scaled using StandardScaler.

Clustering:

MST Clustering:

MST Clustering is a graph-based clustering which uses dissimilarity graph for creating a minimum spanning tree and creates clusters by breaking links of the large dissimilarity nodes.

This clustering showed very less silhouette score, less than 0.1 with the clusters generated. The silhouette scores are plotted below.

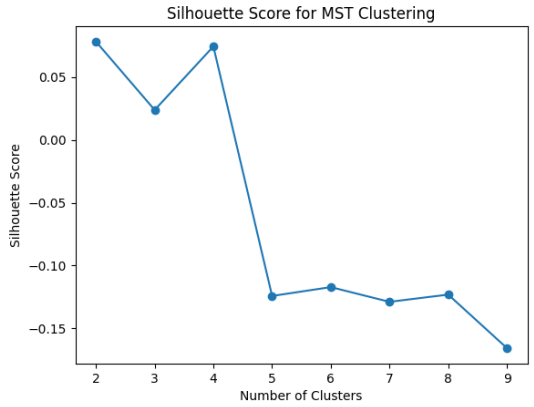

Figure 2:Silhouette Score graph for MST Clustering

Like MST Clustering, SNN Density-based Clustering didn't performed well in clustering this dataset. And these two clustering algorithms uses similarity and dissimilarity graphs, it has never mentioned they exploits the properties of these graphs while partitioning.

Spectral Clustering:

Spectral clustering is a graph partitioning approach which explicitly use the properties of similarity graph to determine the clusters.

Unlike MST and SNN Density-based algorithms, Spectral Clustering have reasonable Silhouette score of 0.72 for 2 clusters, which is enough to say the cluster is good to consider. The resultant clusters can be seen below.

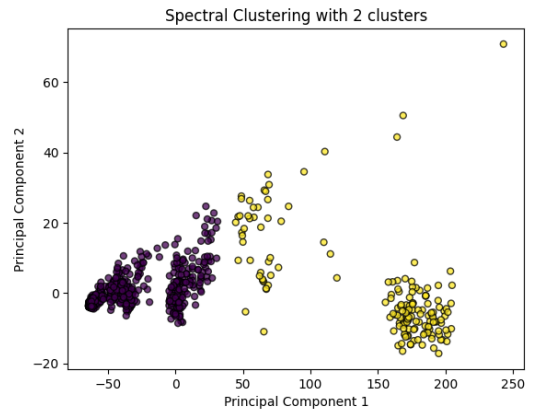

Figure 3:Resultant clusters with Spectral Clustering (N=2)

We can see the products in one of the clusters are scattered around over principal components and most of the products are grouped at a close distance and we can say that they might have similarly performed over weeks.

K-Means Clustering:

K-Means Clustering is a method of vector quantization, which clusters data points with its closest cluster centre or cluster centroid.

On plotting the elbow curve for finding optimal number of clusters K, it is observed that the inertia change is minimal beyond K=3. And the silhouette score of K-Means clustering with K=3 is 0.615 and the silhouette score of K-Means clustering with K=2 is 0.75.

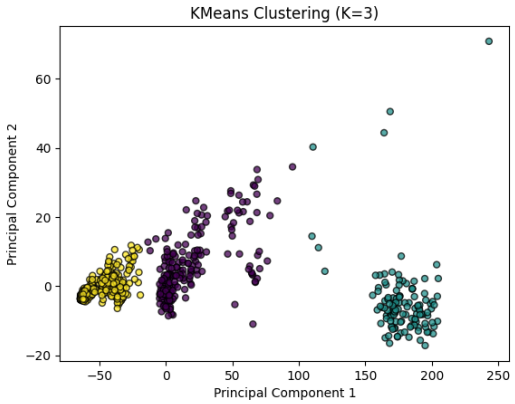

Figure 4:Resultant clusters with K-Means Clustering (K=3)

Results and Discussion:

Silhouette scores for all the clustering algorithms performed are calculated and can be seen below.

Clustering Algorithm	Silhouette Score
MST Clustering (N=2)	0.078
Spectral Clustering (N=2)	0.723
Spectral Clustering (N=5)	0.455
K-Means Clustering (K=3)	0.615
K-Means Clustering (K=2)	0.75

All the resultant clusters have been plotted to understand the grouping of products.

It is observed that Spectral and K-Means clustering algorithms performed grouping products quite like each other.