Clustering, Anomaly Detection, and Dimension Reduction Techniques for Analyzing Two Datasets

Shoham Yamin ID 319151213

1 Abstract

This study presents an analysis of two datasets, each with ground truth, utilizing clustering, anomaly detection, and dimension reduction techniques. The two datasets include a static dataset and a dynamic dataset. The objective is to cluster the remaining columns and determine the best clustering method, associate the clusters with the ground truth, and identify the external variables best associated with the clusters. We use at two clustering technique and two dimension reduction methods.

2 Introduction

Clustering is a powerful unsupervised learning method used to group similar data points based on their features or characteristics. Clustering is used in a wide range of applications, including customer segmentation, anomaly detection, and image recognition. The objective of this study is to analyze two datasets with clustering, anomaly detection, and dimension reduction techniques. The two datasets are a static dataset, and a dynamic dataset, each with ground truth. The study aims to cluster the remaining columns, associate the clusters with the ground truth, and identify the external variables best associated with the clusters.. Finally, we reduce the dimension of the data and propose a visualization scheme to highlight the clusters and variables associated with them.

3 Methods

We begin by preprocessing the datasets, which involves removing any missing or incomplete data and scaling the features to have zero means and unit variances. We then use at least two clustering techniques to cluster the data. We also use the Elbow method to determine the number of clusters when using the k-means algorithm and using HDBSM. We evaluate the quality of the clustering using

internal and external validation measures, such as the Within-cluster sum of squares (WSS) and the adjusted Rand index. We associate the clusters with the ground truth and identify the external variables best associated with the clusters. We examine the statistical significance of the association using the ANOVA test. Finally, we reduce the dimension of the data using TSNE and PCA for dimension reduction.

4 Results

For the Gas data set we use k-means algorithm with PCA to visualize the data before the clustering and after the clustering. Here are the results for this dataset:

We use the Elbow method with k-means to determine the number of clusters. We get that the elbow is 7.

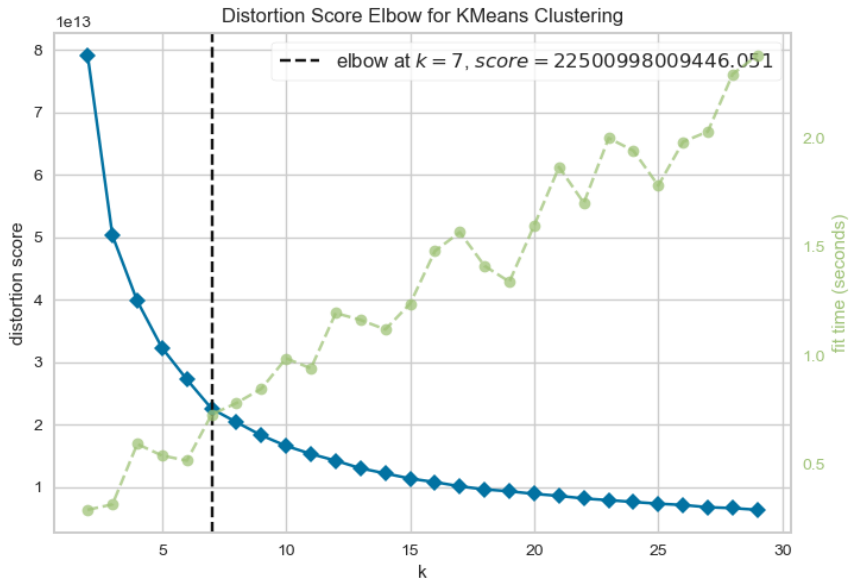


Figure 1: Gas Elbow kmeans with 7 clusters

After clustering the data with k-means we plot the clustering of the samples in 3D with PCA reduction algorithm. Here is the data with the ground truth using PCA algorithm. We can see the similarity between the clustering and the ground truth.

Now we use the ANOVA test to check which features are not significant and we get that feature 41 getting the highest p-value:

Anomaly detection:

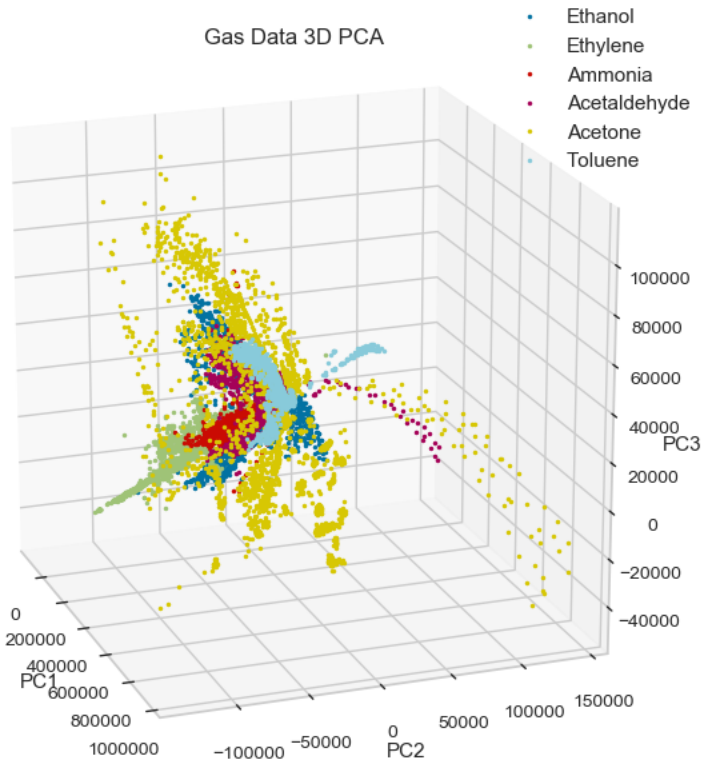


Figure 2: Gas data set with the ground truth and PCA for dimension reduction.

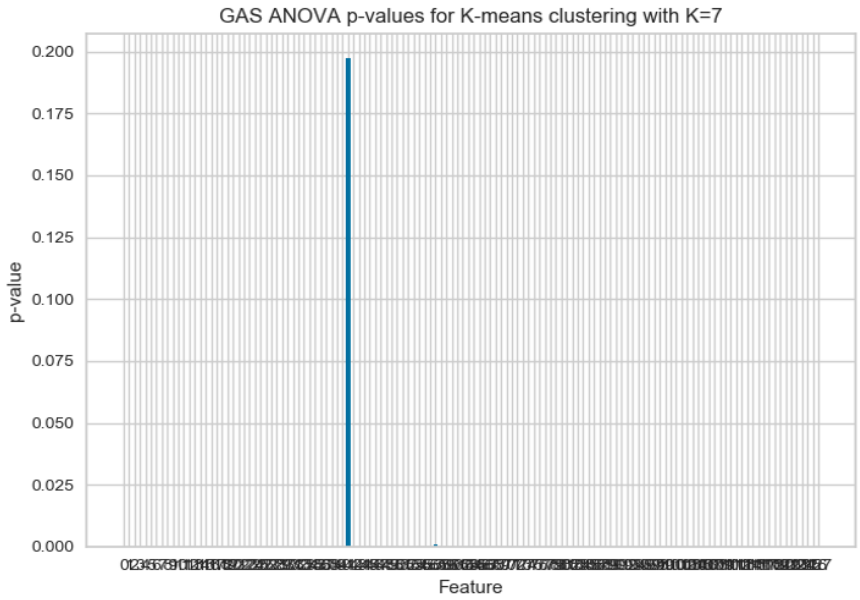
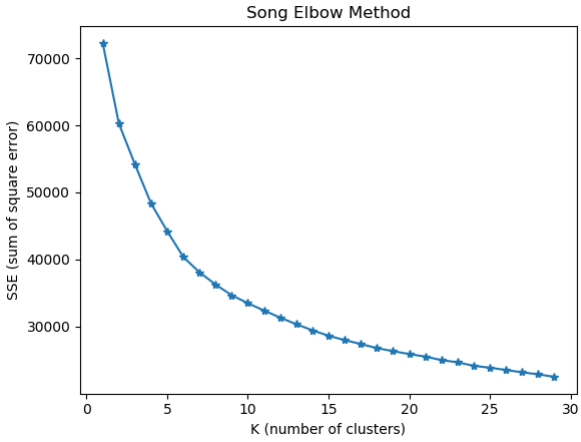


Figure 3: Gas ANOVA test 41 feature with highest p-value.

When using the k-means algorithm, it doesn't mark any sample outside of the cluster mining so we can say that we don't find any anomalies in this dataset.

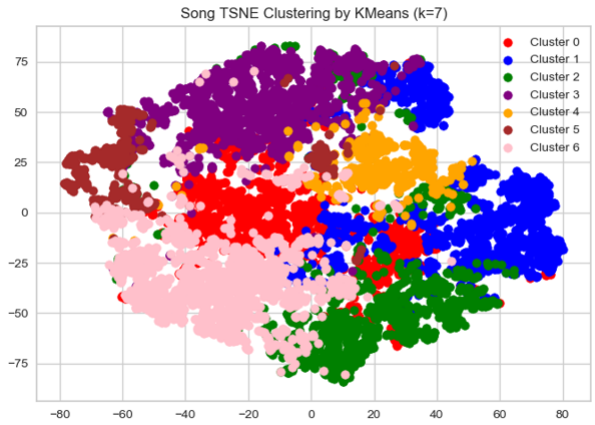
For the Songs Dataset, we use both k-means and HDBSCAN with PCA and t-SNE.

Here we also use the elbow method, and we get that 7 clusters are the best for this dataset with k-means:



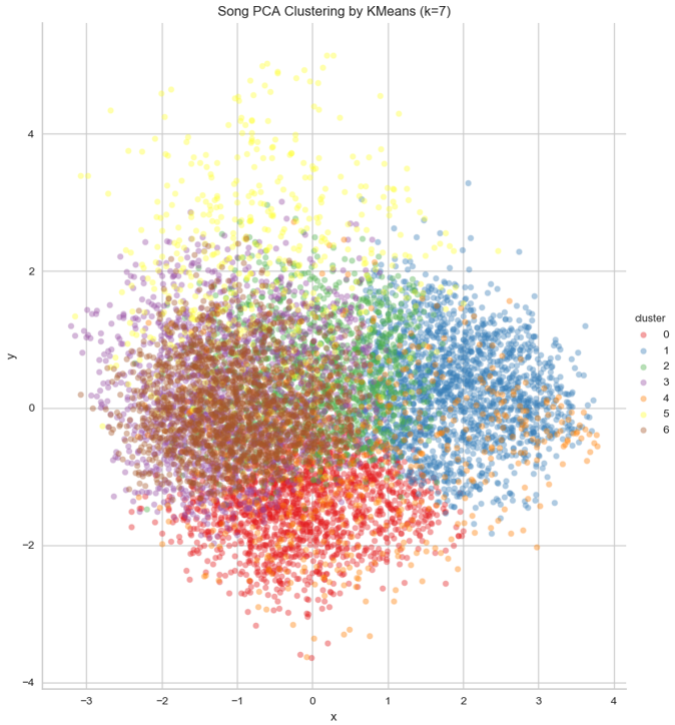
Elbow method for songs dataset with k-means algorithm, k=7 .

Here is t-SNE with the 7 clusters:



t-SNE with 7 clusters for songs dataset.

We also tried to use PCA in 2D, but the results were not as good as with t-SNE:



PCA with 7 clusters for the songs dataset.

5 Results

This study has revealed interesting findings in both the Gas dataset and the songs dataset. In the Gas dataset, the ANOVA test demonstrated that most of the features had an impact on the clustering process. This is a significant result, as it suggests that multiple variables contribute to the groupings identified by the clustering algorithms. Additionally, we observed that the data is almost separated into the exact number of labels, which is a promising indication of the quality of the clustering results.

Moving on to the songs dataset, we found that t-SNE was a more suitable visualization technique compared to PCA in some cases. Through t-SNE, we were able to identify 7 categories for each track, which is a notable finding. This result raises questions about the differences between music genres and whether they can be more effectively identified through improved preprocessing or more refined techniques.

Overall, this study provides valuable insights into the use of clustering, anomaly detection, and dimension reduction techniques in analyzing datasets with ground truth. The findings underscore the importance of careful preprocessing, appropriate visualization methods, and statistical testing to obtain meaningful results.

You can find the code for this project on my Git repository:

References

[1] `https://github.com/mdeff/fma`.

[2] `https://www.kaggle.com/code/shabanamir/unsupervised-ml-project-music-clustering/`.

[3] `https://towardsdatascience.com/understanding-one-way-anova-df44f02922fe`.

[4] Intel, "Chatgpt." `https://chat.openai.com/chat`.