LABWORK 2

PRINCIPAL COMPONENT ANALYSIS

Mai Hải Đăng - BI12-076

01.12.2023

Machine Learning and Data Mining II

Contents

1. Introduction

- This report is dedicated to the understanding of data clustering, more specifically hierarchical clustering and K-means clustering. There will only be a brief mention about PCA (principal components analysis) and SVD (single value decomposition) as we will use it in some experiments.
- Hierarchical clustering
    - Definition of hierarchical clustering.
    - What is agglomerative type clustering?
    - What are the evaluation methods?
    - Dendrogram analysis.
- K-means clustering
    - Experiment with different k-means clustering protocols.
    - Different methods to initialize the centroids.
    - Analyze and compare the results with different values of k.
    - Evaluate our clustering quality.
    - Apply PCA (principal components analysis) or SVD (single value decomposition) to see the performance of k-means before and after the dimensionality reduction.
- We will have a brief mention about the concept of data types, statistical measurements, PCA, and SVD.
- The original dataset are Gas Turbine CO and NOx Emission [1] and Dry Bean [2]
- Disclaimer: I'm definitely neither an expert in fields of Energy Technology nor Biology, some of the explanation might just be from intuitive assessment, and general research online.

2. Concepts

2.1. Basic concepts

2.1.1. Data

- Discrete data: Take on distinct, separate values. Countable (can count the number of different values it can take).
- Continuous data: Take on an infinite number of values within a specified range. Measurable.
- Qualitative data: Descriptive in nature and deals with qualities or characteristics. It is non-numerical and is often expressed in words.
- Quantitative data: Measurable quantities and is expressed in numerical terms. Can be continuous (measured on a scale) or discrete (countable)

2.1.2. Principal components analysis

- The PCA objective is to project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.
- Through the steps: standardize the data; compute the covariance matrix, the eigenvectors, and eigenvalues; sort in descending order of the eigenvectors based on their eigenvalues; select the top k eigenvectors to form the basis of the new subspace; then project the data onto the new subspace.

2.1.3. Statistical Measurements

- The mean is a measure of central tendency and represents the average value of a set of numerical values.
- Variance measures the spread or dispersion of a set of values from the mean. It gives an indication of how much individual data points differ from the average.
- Covariance measures how two variables change together. A positive covariance indicates a positive relationship, while a negative covariance indicates a negative relationship.
- Correlation is a standardized measure of the strength and direction of a linear relationship between two variables. It is a unitless measure that ranges from -1 to 1.

2.1.4. Single value decomposition

- The SVD objective is to decomposes a matrix into three simpler matrices, facilitating various operation and providing insights into the structure of the original matrix
- It is represented as: $A = U\Sigma V^T$
    - U is an orthogonal matrix containing the left singular vectors, or the eigenvectors of the matrix.
    - Σ is a diagonal matrix containing the singular values, these values are non-negative and are arranged in descending order, or the eigenvalues of the matrix.
    - The columns of V (or rows of V^T) represent the right singular vector. Like U, they form an orthogonal basis but for the row space of A.

2.2. Clustering

2.2.1. Definition

- Data clustering is a procedure in which we make a cluster of entities based on their similar features.
- To create a similarity clusters distance measured is used, which are as follows: Euclidean distance, Manhattan, or taxicab distance, Mahalanobis distance, Inner product space, etc. … All distances are used to find similarity between different points, but mostly Euclidean distance is used to measure objects for similarity attributes.

2.2.2. Evaluation

a) Davies-Bouldin index:

- DBI can show clustering quality with intra-cluster similarity and cluster-like similarity.
- DBI works by calculating the average value of each item in the data set. The value of each point is calculated as the sum of the compactness values divided by the distance between the two center points of the group as separation. The smaller DBI value indicates the best number of clusters.[4]
- Formula:

$$DB = \frac{1}{N}\sum_{i=1}^{K} \max_{i \neq j}\left(\frac{\mu_i + \mu_j}{d(w_i, w_j)}\right)$$

where: $\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} d(w_i, x_j)$: Average distance of x_j the point j-th of the cluster i-th from the centroid w_i. n_i is the number of points belonging to cluster i.

b) Dunn index

- The aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. [5]
- There are variances of calculating the Dunn Index
    1. Centroid-based: Finding the centroid of each cluster, then calculating the distance from each point to its cluster's centroid (d1). It then calculates the average distance for each cluster and for all clusters (d2). Then the DI calculated as the min(d1) divided by the max(d2).
    2. Pairwise: Calculating the mean inter-cluster distance (the average distance between all points in two different clusters) and the mean intra-cluster distance (the average distance between all points within a single cluster). Then the DI calculated as the min of intra-cluster distance divided by the maximum inter-cluster distance.
- In this report I suggest using the centroid-based variance because during the experiments, since the datasets are rather big, it would be too computationally exhaustive to use pairwise calculation.
- Formula:

$$DI_m = \frac{\underset{1 \leq i < j \leq m}{min} \delta(C_i, C_j)}{\underset{1 \leq k \leq m}{max} \Delta_k}$$

Where:

m is the number of clusters.

Δ_k intra-cluster distance of cluster k.

$\delta(C_i, C_j)$ inter-cluster distance between two cluster C_i, C_j.

c) Silhouette index

- Measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)
- The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. [6]
- Formula:

For data point $i \in C_I$, let

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(x_i, x_j)$$

be the mean distance between i and all other data points in the same cluster where $|C_I|$ is the number of points belong to cluster C_I, and $d(x_i, x_j)$ is the distance between data point i and j in the cluster.

And,

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(x_i, x_j)$$

to be the smallest mean distance of i to all points in any other cluster.

Silhouette (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, if |C_I| > 1$$

2.3. Hierarchical clustering

2.3.1. Definition

- Is a method of clustering analysis that seeks to build a hierarchy of clusters.
- Agglomerative or additive hierarchical clustering (AHC) is a type of hierarchical clustering where the method is a “bottom-up” approach. Each observation starts in its own cluster, and pairs of clusters which have the minimize distance are merged and move up the hierarchy.

2.3.2. Algorithm

1. Treat each data point as a single cluster, forming n initial clusters, where n is the number of data points.
2. Calculate the distance (or dissimilarity) between each pair of clusters. The choice of distance metric depends on the application.
3. Find the closest (most similar) pair of clusters based on the computed distances and merge them into a single cluster. Update the distances between this new cluster and the remaining clusters.
4. Repeat step 2 and 3.
5. Stopping criterion: Only a single cluster remains.

➔ The time complexity: $O(n^2)$

2.3.3. Single linkage vs. Complete linkage

Single linkage	**Complete linkage**
Single linkage (also known as nearest-neighbor linkage) measures the distance between two clusters based on the shortest distance between any two points in the two clusters.	Complete linkage (also known as farthest-neighbor linkage) measures the distance based on the maximum distance between any two points in the two clusters.
$D(X, Y) = \min_{x \in X, y \in Y} d(x, y)$	$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$
□ Form clusters where the members are close to each other. □ The clusters can be elongated. □ It is sensitive to outliers and noise	□ Form compact, spherical clusters. □ Less sensitivity to outliers.

2.3.4. Dendrogram

- Is a diagram representing a tree, it illustrates the arrangement of clusters produced by the corresponding analyses.

2.4. K-means clustering

2.4.1. Definition

- Is a method of cluster analysis.
- Create a cluster n number of objects into k clusters. [3]
- It uses vector quantization and aims to assign each observation to the cluster with the nearest mean or centroid.
- The goal is to minimize the *Sum of Squared Distances* between the data points and their corresponding cluster centroids, resulting in clusters that are internally homogeneous and distinct from each other.

2.4.2. Basic algorithm

1. Choose the number of clusters k.
2. Select k random points from the data as centroids.
3. Assign all points to the closest cluster centroid.
4. Recompute the centroids of the newly formed clusters.
5. Repeat steps 3 and 4.
6. Stopping criterion
   a. Centroids of newly formed clusters do not change.
   b. Point remains in the same cluster.
   c. Maximum number of iterations is reached.

➔ The time complexity: $O(n \cdot k \cdot d)$ where n, k, d are the number of data points, clusters, and feature respectively.

2.4.3. K-mean++

- An algorithm for choosing the initial values (or "seeds") for the k-means clustering algorithm (step 1 in 2.3.2). [7]
- The first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the *remaining* data points with probability proportional to its squared distance from the point's closest existing cluster center.
- Algorithm
  1. Choose one center uniformly at random among the data points.
  2. For each data point x not chosen yet compute$D(x)$, the distance between x and the nearest center that has already been chosen.
  3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$.
  4. Repeat Steps 2 and 3 until k centers have been chosen.
  5. Now that the initial centers have been chosen (from step 2 onward in 2.3.2).

2.4.4. Elbow Method

- The **elbow method is a graphical representation of finding the optimal 'K'** in a K-means clustering. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.
- Algorithm
  1. Select the number of clusters for the dataset (K)
  2. Select the K number of centroids randomly from the dataset.
  3. Now we will use Euclidean distance or Manhattan distance as the metric to calculate the distance of the points from the nearest centroid and assign the points to that nearest cluster centroid, thus creating K clusters.
  4. Now we find the new centroid of the clusters thus formed.
  5. Again, reassign the whole data point based on this new centroid, then repeat step 4. We will continue this for a given number of iterations until the position of the centroid doesn't change, i.e., there is no more convergence.

3. Data Analysis

3.1. X-dataset

- Extremely simple, just an array of data {1, 2, 9, 12, 20} in 1-D space. We can see a linear increase in the data from left to right.

3.2. Gas Turbine CO and NOx Emission

3.2.1. General information

- The dataset contains 36733 instances of 11 sensor measures aggregated over one hour collecting data from year 2011-2015 from a gas turbine located in Turkey for the purpose of studying flue gas emissions, namely CO and NOx exhausted for assessing the performance and environmental impact of gas turbine.
- This dataset will be used to experiment with unsupervised learning models. Using internal validation to determine the quality of the clusters.

Table 1. General characteristics of Gas Turbine Emission dataset.

Data	Role	Data Type	Non-null	Description
AT	Feature	Continuous, Quantitative	True	Ambient temperature (ºC)

AP	Feature	Continuous, Quantitative	True	Ambient pressure (mbar)
AH	Feature	Continuous, Quantitative	True	Ambient humidity (%)
AFDP	Feature	Continuous, Quantitative	True	Air filter difference pressure (mbar)
GTEP	Feature	Continuous, Quantitative	True	Gas turbine exhause pressure (mbar)
TIT	Feature	Continuous, Quantitative	True	Turbine inlet temperature (ºC)
TAT	Feature	Discrete, Quantitative	True	Turbine after temperature (ºC)
CDP	Feature	Continuous, Quantitative	True	Compressor discharge pressure (mbar)
TEY	Feature	Continuous, Quantitative	True	Turbine energy yield (MWH)
CO	Feature	Continuous, Quantitative	True	Carbon monoxide (mg/m3)
NOx	Feature	Continuous, Quantitative	True	Nitrogen oxides (mg/m3)

Table 2. Statistical Measurements of Gas Turbine Emission dataset.

Data	Mean	Standard Deviation	Variance
AT	17.712726	7.447451	55.464530
AP	1013.070165	6.463346	41.774841
AH	77.867015	14.461355	209.130787
AFDP	3.925518	0.773936	0.598976
GTEP	25.563801	4.195957	17.606059
TIT	1081.428084	17.536373	307.524376
TAT	546.158517	6.842360	46.817896
CDP	12.060525	1.088795	1.185475
TEY	133.506404	15.618634	243.941740
CO	2.372468	2.262672	5.119683
NOx	65.293067	11.678357	136.384028

Table 3. Explained variance by the principal components of Gas Turbine Emission dataset.

Principal Components	Explained Variance
PC1	0.54285619
PC2	0.21720478
PC3	0.13215557
PC4	0.05557157
Sum	0.9477881004626526

3.2.2. In-dept analysis

- Features with high correlation (see Figure 1):
  1. Waste assessment: During combustion process two key pollutant in the exhaust gases are CO and NOx. Indicating that if the combustion efficiency (the ratio of heat released by the fuel) is low it will ultimately lead to extra waste of fuel and creating CO and NOx. This is confirmed with the positive correlation between the two gases.
  2. Performance assessment: We can see a strong correlation between the total energy yielded (TEY) and the following features AFDP, GTEP, TIT, CDP. Air filter difference pressure indicates the resistance of the air encounters as it passes through the filter before entering the combustion chamber. With better AFDP help reduce the TIT which is the temperature of the air entering the gas turbine combustion chamber, allow for higher compressor discharge pressure (CDP) to compress the air leaving the compressor (heightened the GTEP) and entering the combustion chamber.
- Features with low correlation:
  1. Higher temperature (AT) will have lower pressure (AP) and humidity level (AH) due to increase evaporation.
  2. Turbine After Temperature (TAT) is indicating that the general ambient temperature (AT) are elevated. With the exhausted gases having higher temperature meaning the combustion process is rather inefficient, which mean the AFDP, GTEP, TIT, TEY, CDP which are the crucial factors influencing the efficiency and power output of the turbine are all very low.

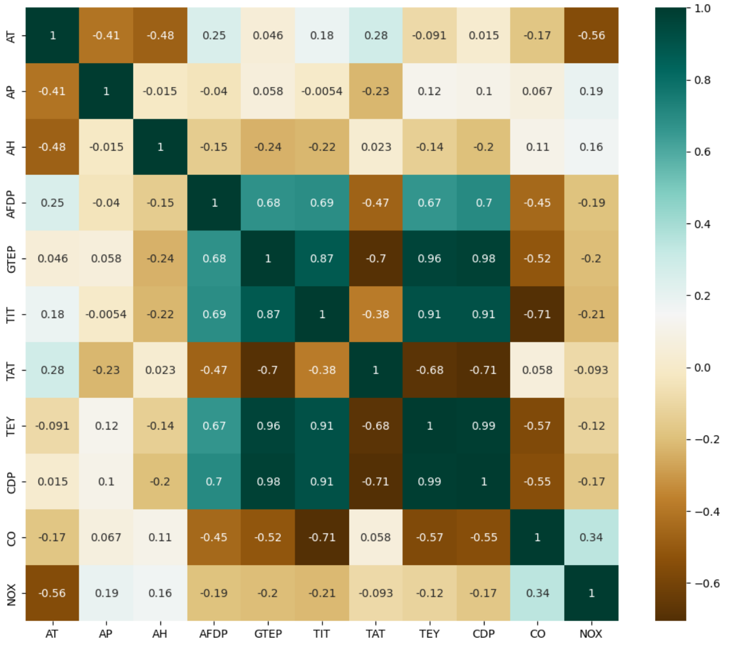

Figure 1. Gas Turbine Emission dataset correlation matrix

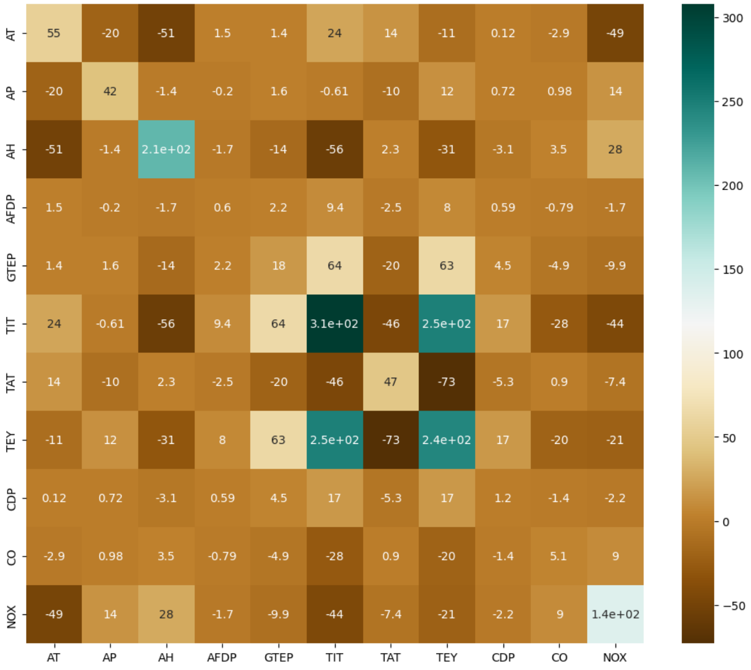

Figure 2. Gas Turbine Emission dataset covariance matrix

3.3. Dry Bean

3.3.1. General information

- Images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. A computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. A total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains.
- There exists the target class (what type of bean) allowing use to do use both internal and external validation to determine the quality of the clusters.

Table 4. General characteristics of Dry Bean Dataset.

zData	Role	Data Type	Non-null	Descriptions
Area	Feature	Discrete, Quantitative	True	The area of a bean zone and the number of pixels within its boundaries
Perimeter	Feature	Continuous, Quantitative	True	Bean circumference is defined as the length of its border.
MajorAxisLength	Feature	Continuous, Quantitative	True	The distance between the ends of the longest line that can be drawn from a bean
MinorAxisLength	Feature	Continuous, Quantitative	True	The longest line that can be drawn from the bean while standing perpendicular to the main axis
AspectRatio	Feature	Continuous, Quantitative	True	Defines the relationship between MajorAxisLength and MinorAxisLength
Eccentricity	Feature	Continuous, Quantitative	True	Eccentricity of the ellipse having the same moments as the region
ConvexArea	Feature	Discrete, Quantitative	True	Number of pixels in the smallest convex polygon that can contain the area of a bean seed
EquivDiameter	Feature	Continuous, Quantitative	True	Equivalent diameter: The diameter of a circle having the same area as a bean seed area
Extent	Feature	Continuous, Quantitative	True	The ratio of the pixels in the bounding box to the bean area
Solidity	Feature	Continuous, Quantitative	True	Also known as convexity. The ratio of the pixels in the convex shell to those found

				in beans.
Roundness	Feature	Continuous, Quantitative	True	Calculated with the following formula: (4piA)/(P^2)
Compactness	Feature	Continuous, Quantitative	True	Measures the roundness of an object
ShapeFactor1	Feature	Continuous, Quantitative	True	Distinctive aspects or other geometric attributes that help describe or differentiating one type of dry bean to another
ShapeFactor2	Feature	Continuous, Quantitative	True	
ShapeFactor3	Feature	Continuous, Quantitative	True	
ShapeFactor4	Feature	Continuous, Quantitative	True	
Class	Target	Categorical, Qualitative	True	

Table 5. Statistical Measurements of Dry Bean Dataset.

Data	Mean	Standard Deviation	Variance
Area	53048.284549	29324.095717	8.599026e+08
Perimeter	855.283459	214.289696	4.592007e+04
MajorAxisLength	320.141867	85.694186	7.343494e+03
MinorAxisLength	202.270714	44.970091	2.022309e+03
AspectRatio	1.583242	0.246678	6.085026e-02
Eccentricity	0.750895	0.092002	8.464324e-03
ConvexArea	53768.200206	29774.915817	8.865456e+08
EquivDiameter	253.064220	59.177120	3.501932e+03
Extent	0.749733	0.049086	2.409471e-03
Solidity	0.987143	0.004660	2.171913e-05
Roundness	0.873282	0.059520	3.542617e-03

Compactness	0.799864	0.061713	3.808552e-03
ShapeFactor1	0.006564	0.001128	1.272380e-06
ShapeFactor2	0.001716	0.000596	3.550668e-07
ShapeFactor3	0.643590	0.098996	9.800238e-03
ShapeFactor4	0.995063	0.004366	1.906595e-05

Table 6. Explained variance by the principal components of Dry Bean dataset.

Principal Components	Explained Variance
PC1	9.99967207e-01
PC2	3.06176794e-05
Sum	0.9999978243949357

3.3.2. In-dept analysis

- Features with high correlation (see Figure 1):
  1. The area, perimeter, major and minor axis length, convex area, equivalent diameter are all sets of parameters that describe a general shape of an object, all the factors have a dependence relationship with each other hence why they're all closely correlated.
  2. The aspect ratio defines the relationship between MajorAxisLength and MinorAxisLength. The eccentricity of an eclipse (our bean) the ratio of the distances from the center of the ellipse to one of the foci and to one of the vertices of the ellipse. Both features describe the shape of the bean. (see Figure 3)
  3. Solidity or convexity goes hand in hand with roundness. ShapeFactor4 is also highly correlated with solidity, this could be describing the symmetry.
  4. Roundness with compactness, ShapeFactor2 and ShapFactor3. Compactness measures the roundness, that's a given. We can interpret SF2 and SF3 could be describing the curvature of the bean.
  5. SF2 and SF3 is again positively correlated, which give use confident on the theory propose above (4)
- Features with low correlation (see Figure 1):
  1. The higher the roundness and compactness (describing a circle) lead to smaller aspect ratio and eccentricity (describing an eclipse).

2. ShapeFactor1, ShapeFactor2 with area, perimeter, axis lengths, convex area and equivalent diameter. This is a strange one, I can't interpret what it's trying to reflect.

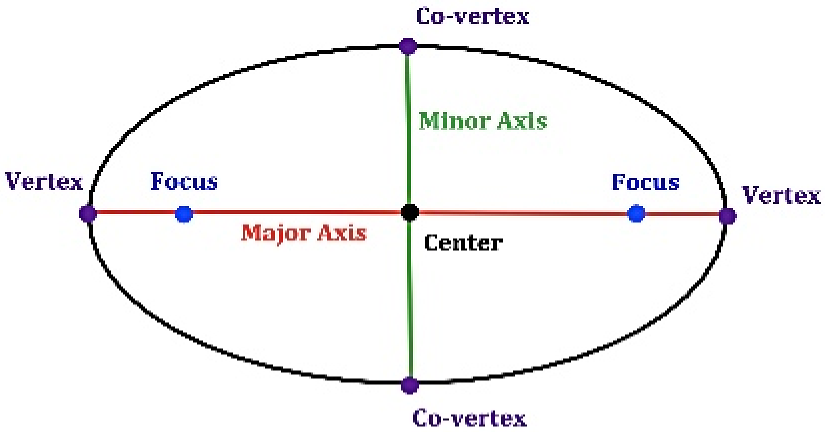

Figure 3. Attributes of an eclipse

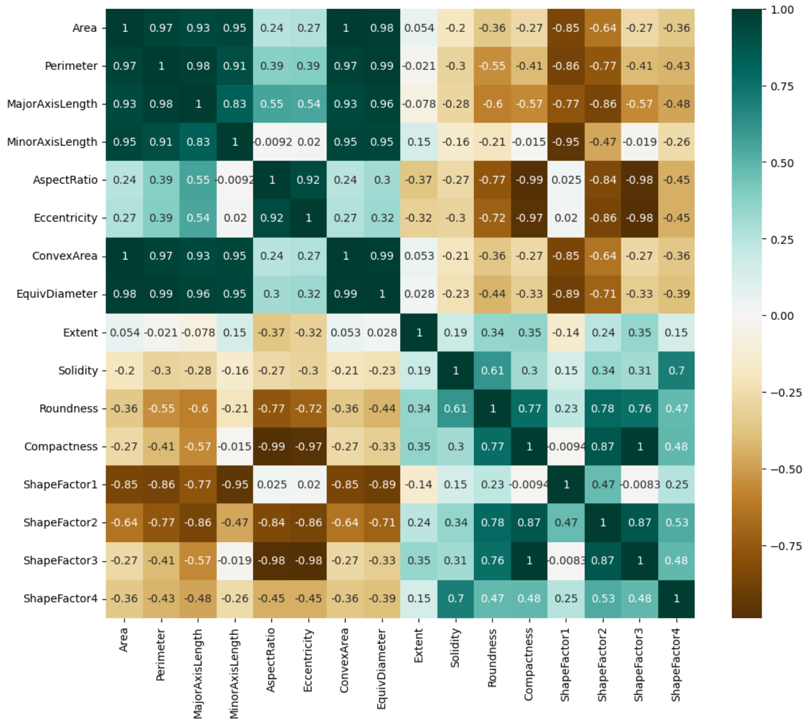

Figure 4. Dry Bean dataset correlation matrix

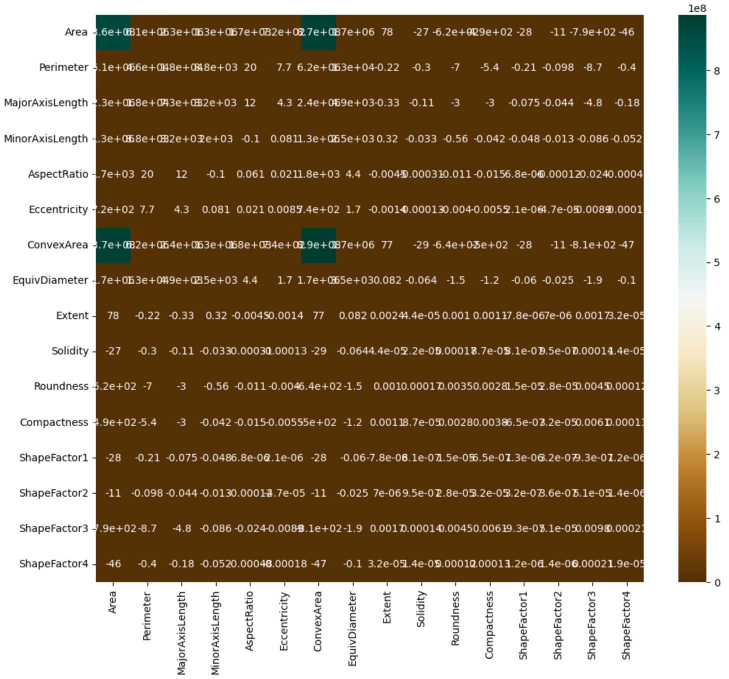

Figure 5. Dry Bean dataset covariance matrix

4. Experiments

4.1. Agglomerative Hierarchical Clustering

4.1.1. Protocol

- Apply the elbow method to decide the number of cluster we should choose.
- Draw the dendrogram with single or complete.
- Draw the decision line that cut the dendrogram into clusters.
- Draw the clusters after clustered.

4.1.2. X-dataset

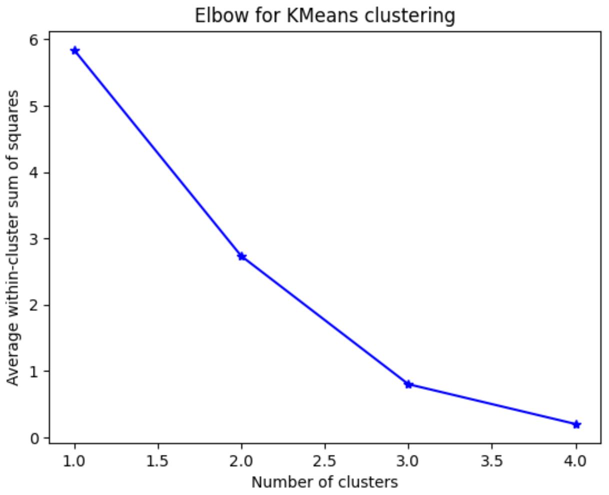

Figure 6. X-dataset. From the elbow graph, we can see three cluster is the best.

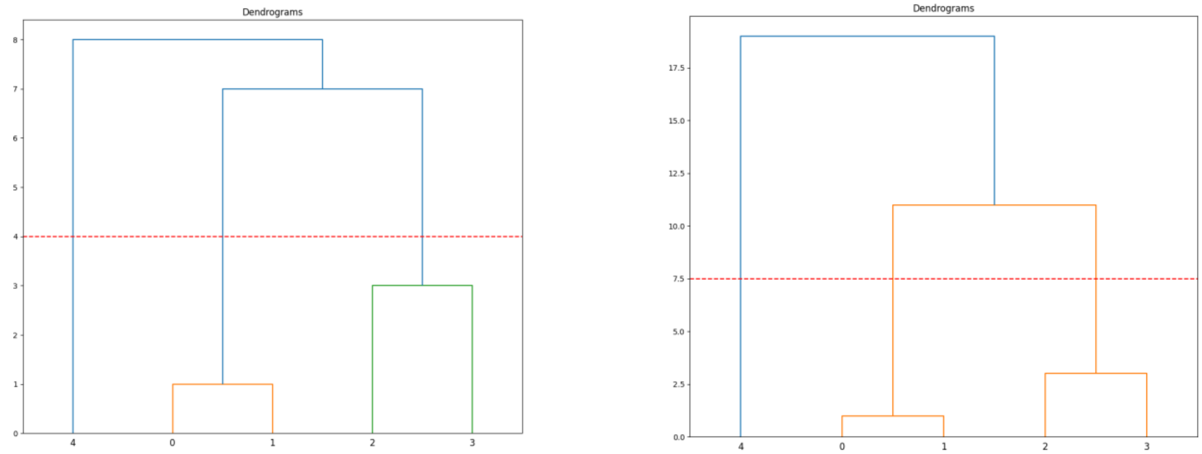

Figure 7. X-dataset. The left graph is the dendrogram drawn using single linkage and the right graph was drawn using complete linkage.

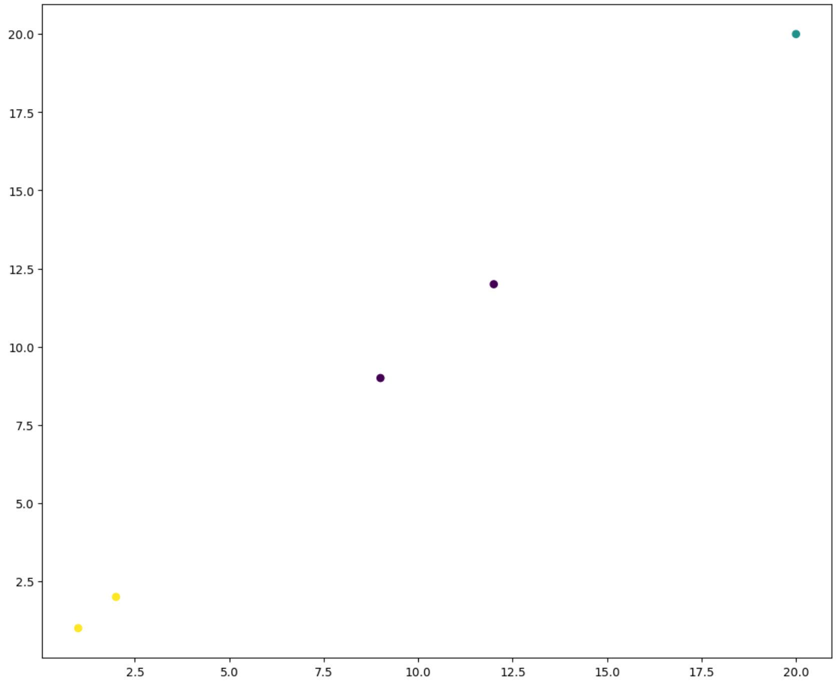

Figure 8. X-dataset after clustering.

4.1.3. Gas Turbine CO and NOx Emission Dataset

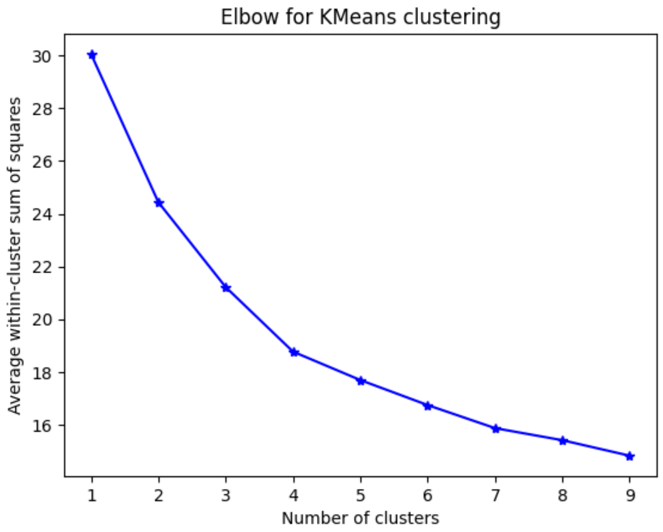

Figure 9. Gas Turbine Emission Dataset. Elbow graph.

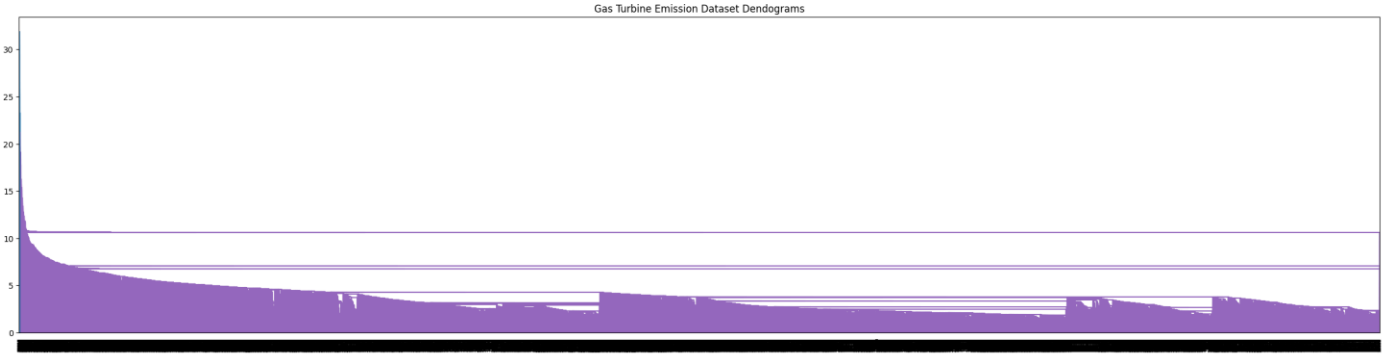

Figure 10. Gas Turbine Emission Dataset. Dendrogram single linkage.

4.1.4. Dry Bean Dataset

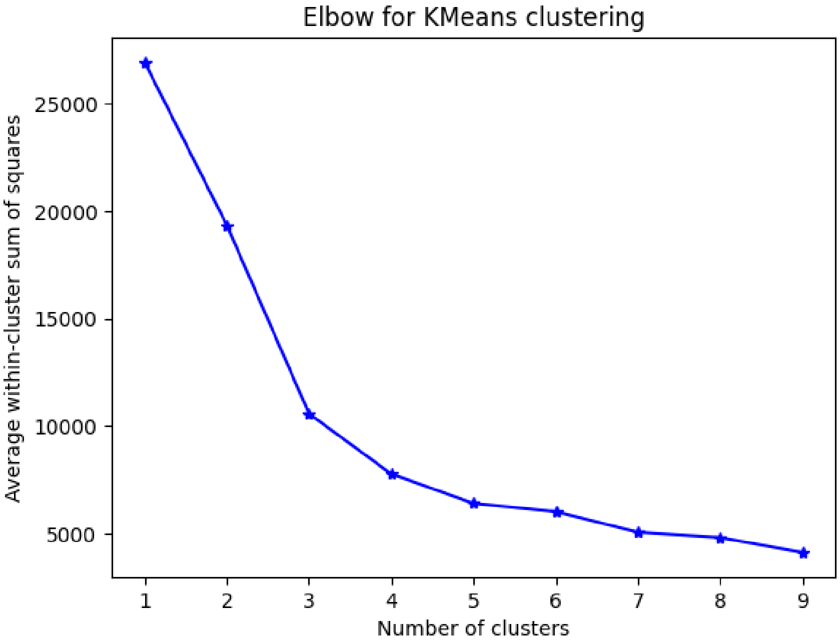

Figure 11. Dry Bean Dataset. Elbow graph.

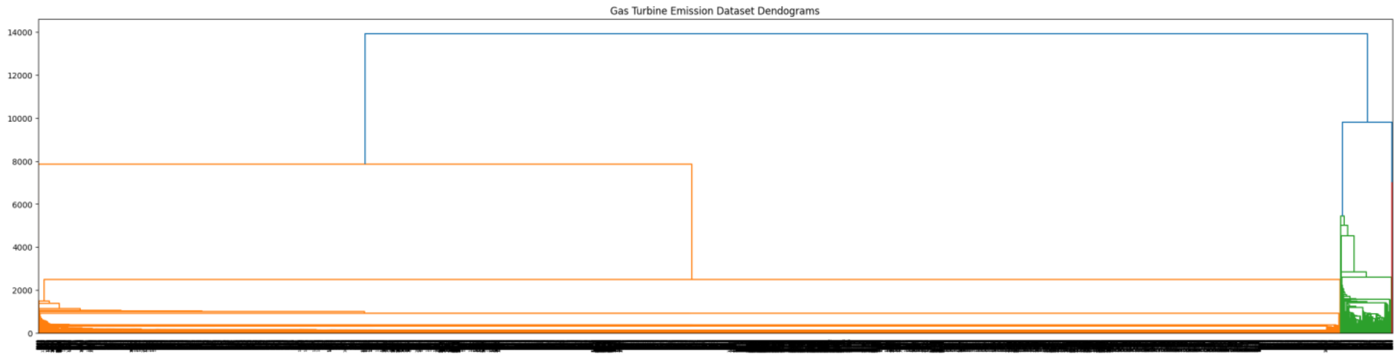

Figure 12. Dry Bean Dataset. Dendrogram single linkage.

4.1.5. Advantages and drawbacks of AHC

- Advantages:
  1. Provides a visual representation of the relationships between clusters at different levels.
  2. Not assume any specific shape for clusters, making it applicable to data with irregular or non-convex cluster shapes.
  3. No need for a specified number of clusters.
  4. Allows the use of various distance metrics.
  5. Easy to understand and implement. Especially for small to moderately sized datasets.
- Drawbacks:
  1. Computationally expensive for large datasets
  2. Sensitive to outliers since it relies on distance metrics.
  3. Fixed merging strategy.
  4. Struggle with noisy data.

4.2. K-means Clustering

4.2.1. Protocol

- Using k-means++ to initialize the centroids.
- Continue with step 2 of the algorithm from section 2.4.2.
- We will be testing with different values of k (2-10).
- Before PCA: Choose two features to visualize the cluster.
- After PCA: Choose two principal components to visualize the cluster.

4.2.2. Gas Turbine CO and NOx Emission Dataset

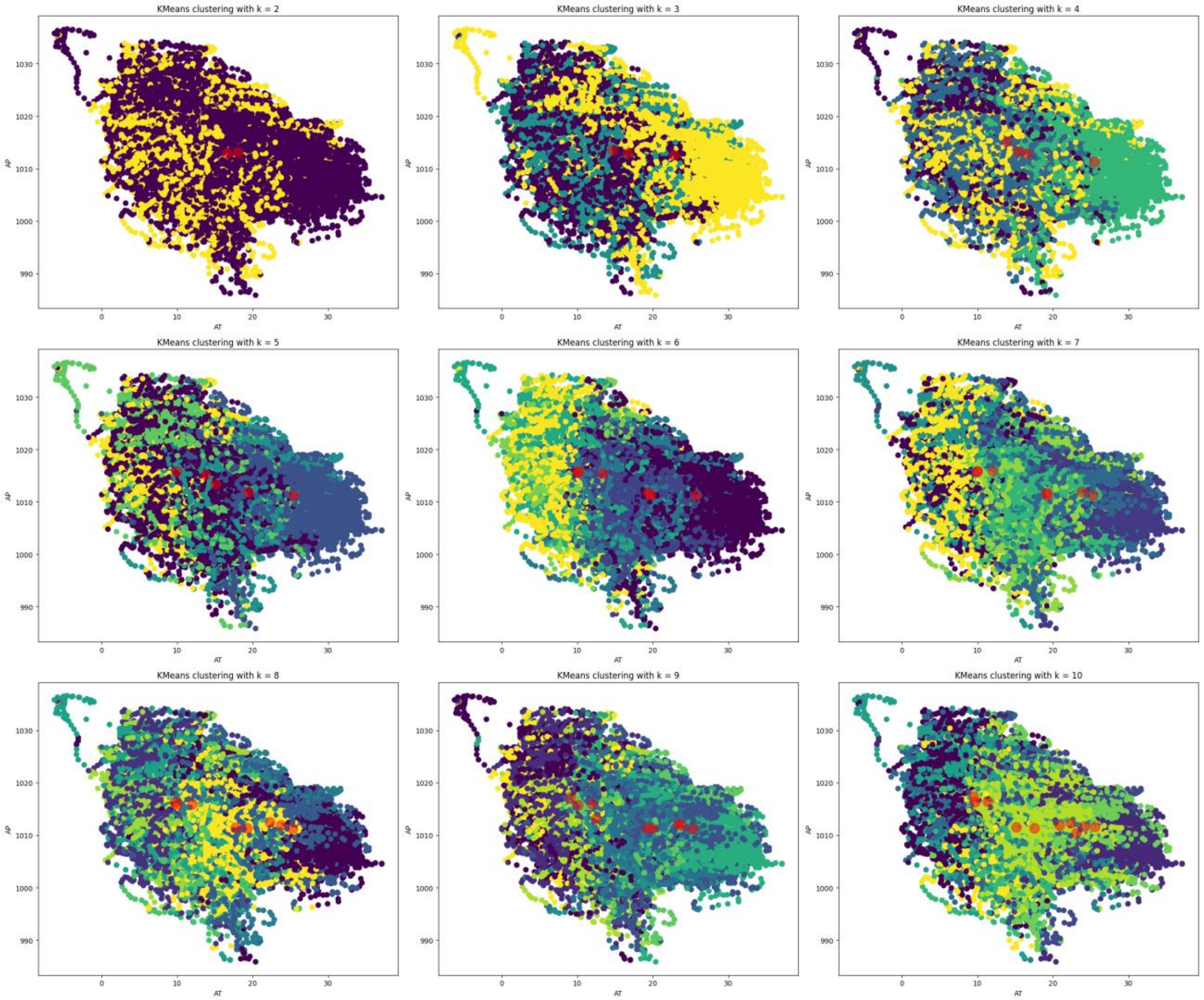

Figure 13. Gas Turbine Emission Dataset. The two-feature ambient temperature and pressure after clustering with 10 different k values before PCA.

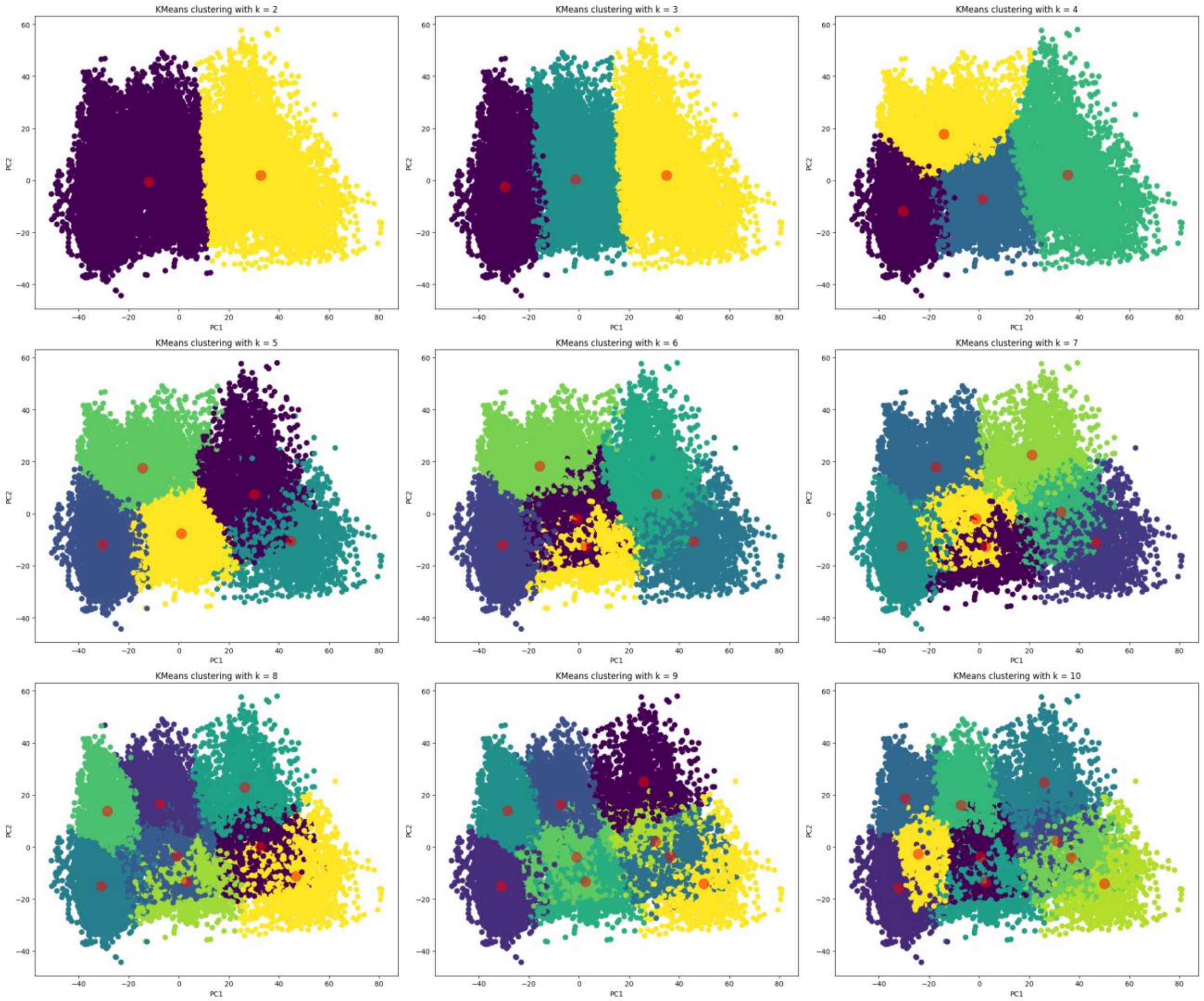

Figure 14. Gas Turbine Emission Dataset. The two-feature ambient temperature and pressure after clustering with 10 different k values after PCA.

4.2.3. Dry Bean Dataset

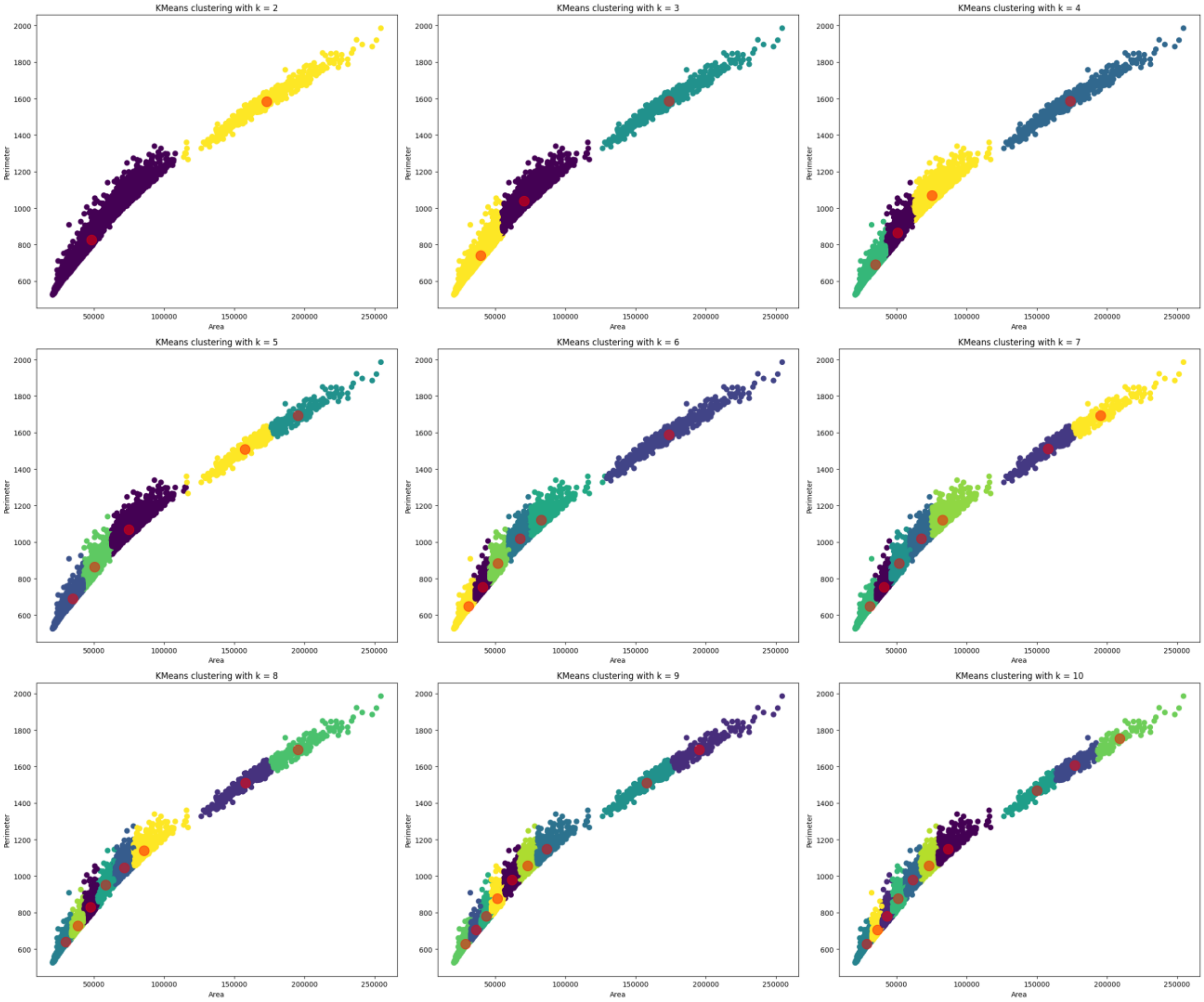

Figure 15. Dry Bean Dataset. The two-feature ambient temperature and pressure after clustering with 10 different k values before PCA.

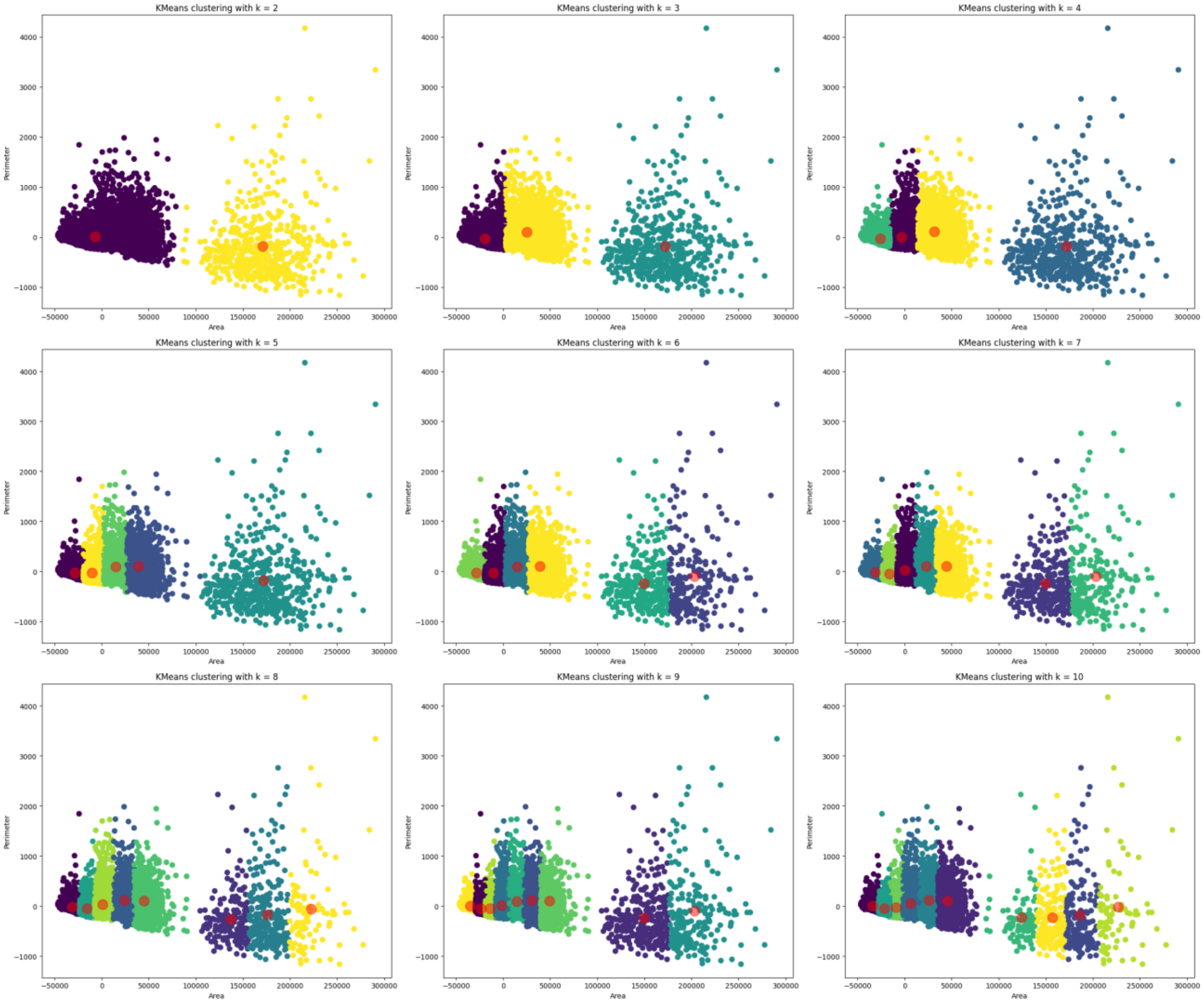

Figure 16. Dry Bean Dataset. The two-feature ambient temperature and pressure after clustering with 10 different k values before PCA.

5. Evaluation

5.1. Gas Turbine CO and NOx Emission Dataset

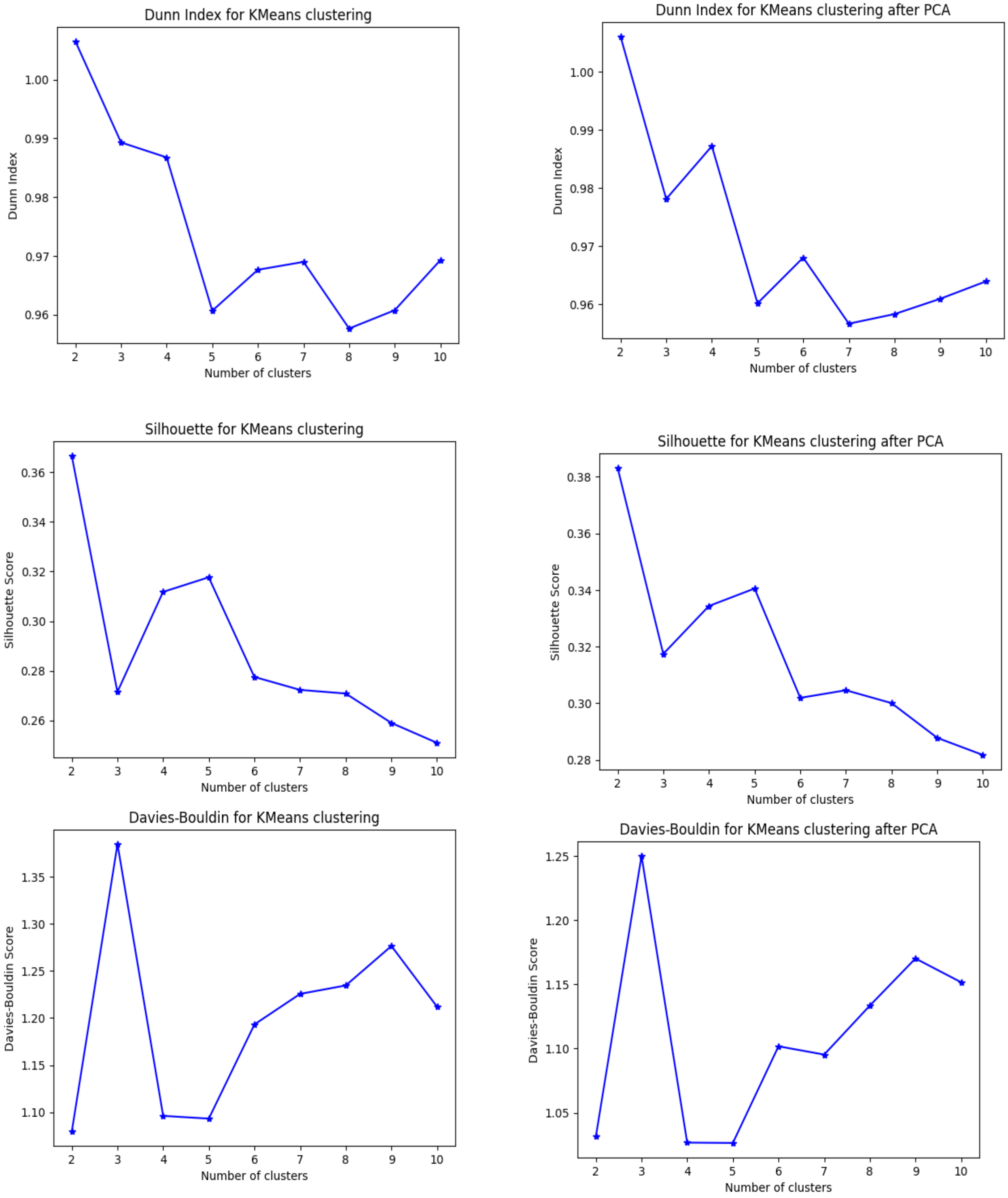

Figure 17. Gas Turbine Emission Dataset. Comparison of evaluation indexes of Dunn Index, Silhouette Index, Davies-Bouldin index before and after PCA.

5.2. Dry Bean Dataset

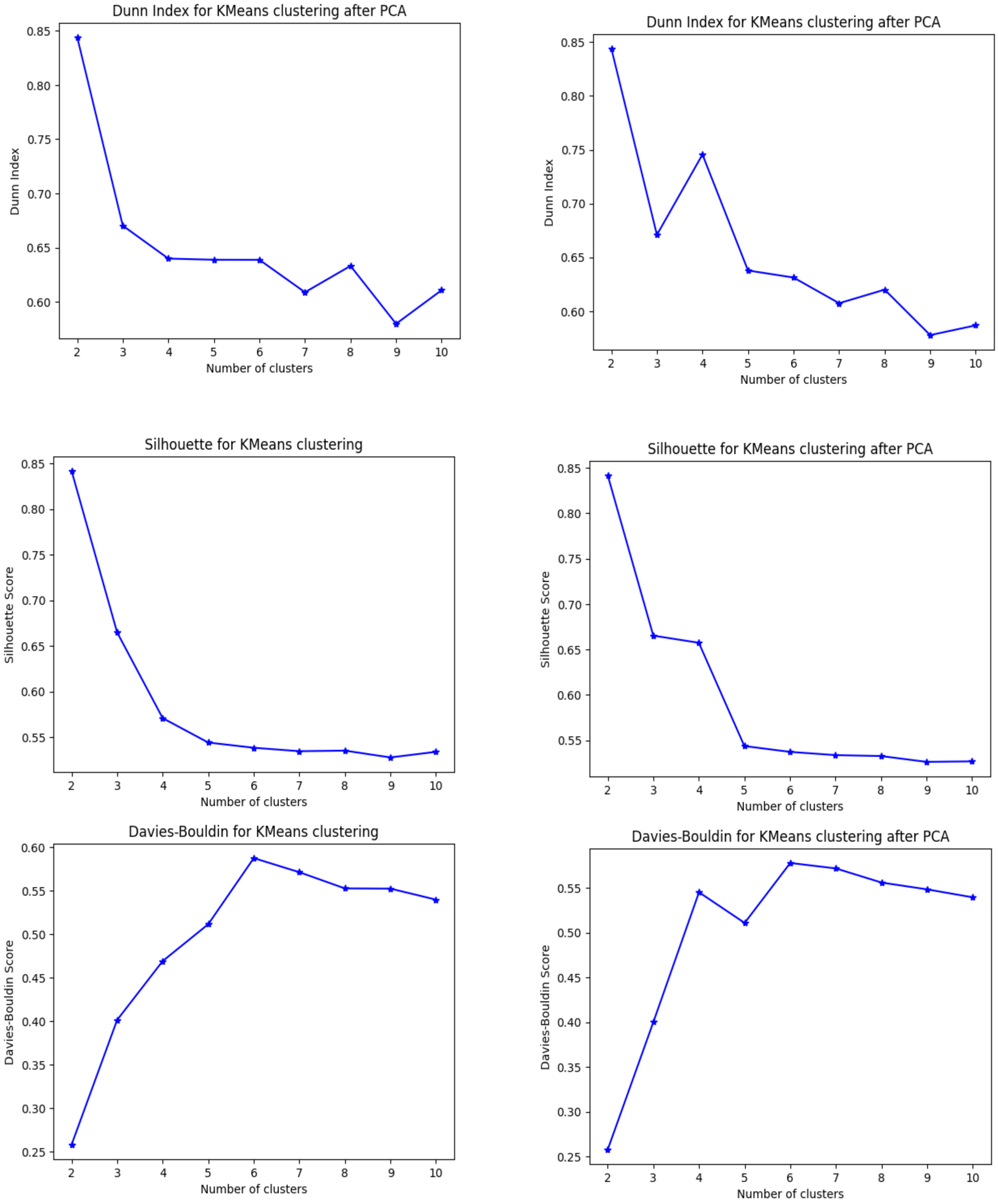

Figure 18. Dry Bean Dataset. Comparison of evaluation indexes of Dunn Index, Silhouette Index, Davies-Bouldin index before and after PCA.

References

[1] Gas Turbine CO and NOx Emission Data Set. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5WC95.

[2] Dry Bean Dataset. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C50S4B. Cortez, Paulo & Teixeira, Juliana & Cerdeira, António & Almeida, Fernando & Matos, Telmo & Reis, José. (2009). Using Data Mining for Wine Quality Assessment. 66-79. 10.1007/978-3-642-04747-3_8.

[3] Bano, Saima & Khan, Naeem. (2018). A Survey of Data Clustering Methods. International Journal of Advanced Science and Technology. 113. 10.14257/ijast.2018.113.14.

[4] Davies Bouldin index algorithm for optimizing clustering case studies ... (n.d.). https://www.temjournal.com/content/103/TEMJournalAugust2021_1099_1103.pdf

[5] J. C. Dunn† (1974) Well-Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetics, 4:1, 95-104, DOI: 10.1080/01969727408546059

[6] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[7] David Arthur and Sergei Vassilvitskii. 2007. K-means++: the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '07). Society for Industrial and Applied Mathematics, USA, 1027–1035.

[8] Nielsen, Frank (2016). "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.