**Khulna University of Engineering & Technology**

**Department of Computer Science and Engineering**

****

CSE 4239: Data Mining

Assignment: Cluster Analysis

**Submitted By**

Faisal Ahmed

Roll: 1607048

**Submitted To**

Animesh Kumar Paul

Assistant Professor

Department of CSE

KUET

**Google Colab Notebook**

1. K-Means Clustering <https://colab.research.google.com/drive/1Px8UImtuA9p0vHSu3qlAETqIWsSSZtTh>
2. K-Medoid Clustering

<https://colab.research.google.com/drive/1Svfm8HodJ4surxWD8-IuBsT2SRrh_2i3>

1. BIRCH

<https://colab.research.google.com/drive/134IcoXFDKXuvp1znXFpsPEcCwNicanGc>

1. Agglomerative Clustering <https://colab.research.google.com/drive/1EgwtPoDUqJJmcghxxvVce3WvCAxyVOWc>
2. DBSCAN

<https://colab.research.google.com/drive/1IMfJSXADNhApTkFMyo62EB0HyRCzPBy-#scrollTo=a6PjluVA-hGx>

1. CHAMELEON <https://colab.research.google.com/drive/1Vb1BOuqsx4MzxzPkaafriPSp0WHysDUi#scrollTo=1Bq-Ng8QmAoJ>

**Table of contents**

1. Introduction
2. Dataset
3. Algorithm
   1. K-Means clustering
   2. K- Medoid clustering
   3. BIRCH
   4. Agglomerative clustering
   5. DBSCAN
   6. CHAMELEON
4. Comparative analysis of different cluster
5. Conclusion
6. References

**1.Introduction**

Clustering algorithms have a powerful tool to analyze a large volume of data generated by modern application. The aim of clustering is to categorize the data according to similar pattern. Clustering is an unsupervised machine learning algorithm where the data has not any particular label. One of the challenging issues in clustering is that define the number of clusters in a particular dataset. There are many types of clustering algorithm and their approach is different. The clustering algorithm is evaluating on various parameter such as runtime, Silhouette Coefficient, ARI, NMI. Clustering algorithm divided into many categories.

1. Partition based clustering
2. Hierarchical based clustering
3. Density based clustering
4. Grid based
5. Model based

Different types of clustering algorithm analysis is an active area of research. Each algorithm is evaluating on theatrically and experimentally. In this report, the analysis and comparison of different clustering algorithm are:

1. K-Means
2. K-Medoids
3. Birch
4. Agglomerative
5. DBSCAN
6. Chameleon

**2. Dataset**

The dataset is containing different types of properties of wheat. The dataset is containing eight columns with wheat properties and target wheat labels. The target level and geometric parameters of wheat kernels measured:

1. Area, A
2. Perimeter, P
3. Compactness, C=4\*pi\*A /P^2
4. Length of kernel, LK
5. Width of kernel, WK
6. Asymmetry coefficient, A\_Coef
7. Length of kernel groove, LKG
8. Target

Link of the seed dataset: <https://archive.ics.uci.edu/ml/datasets/seeds>

**Correlation matrix**

Correlation matrix shows the correlation between each attributes of the dataset. Higher correlation values indicate that the data are highly correlation with each other and lower correlation values indicate that the data has no relation with each other. the range of correlation is -1 to 1.

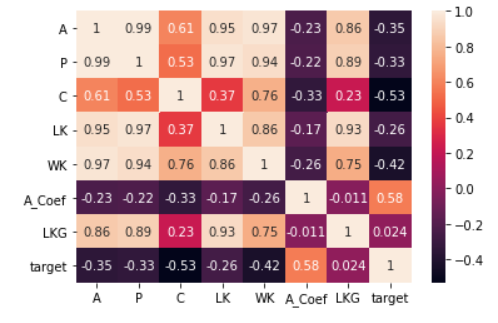


Figure 1: Correlation matrix

The dataset split into features label and the target level. Scaling of data is required to fit the clustering algorithm. The dataset is containing very much large values than others. Without scaling the data, clustering model will not perform well. Data are scaled such that the mean will be 0 and the variance will be 1. In some cases, scaling is required for an algorithm which is based on the distance like clustering and hierarchical clustering. K-Means clustering algorithm has required with scaled data.

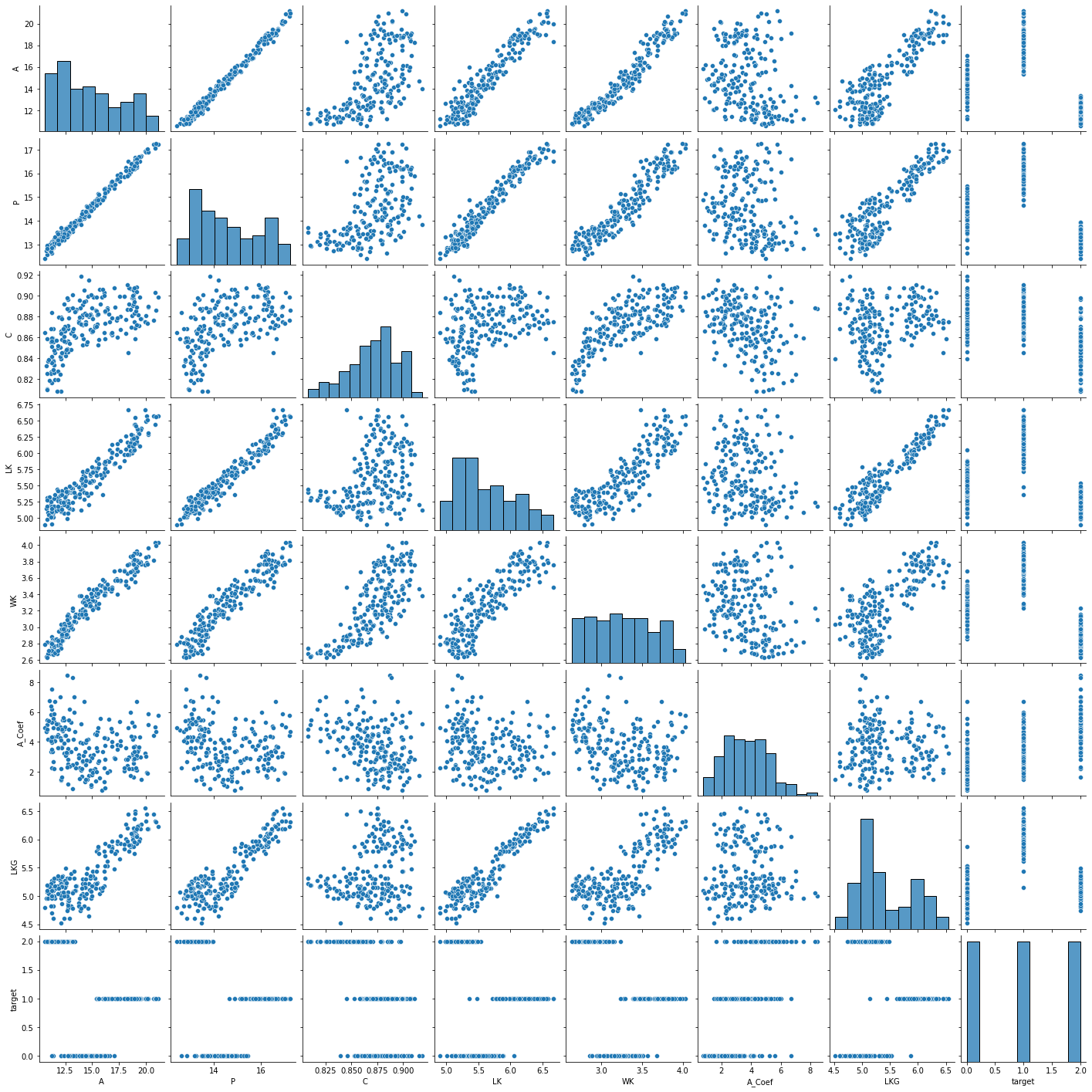


Figure 2: Pair plot of the dataset

**3. Algorithm**

There are different types of clustering algorithm. Mostly the algorithm is classified into partitioning algorithm, density based, hierarchical, grid based etc.

**3.a. K-Means Clustering**

K-means clustering is an unsupervised machine learning algorithm that solves different types of clustering problem solution. Similar data points belong to one cluster and dissimilar points are belonged to another cluster. At first, define the cluster number (k cluster) into a particular dataset. Every cluster (k) is a centroid position for each cluster. Initially k-means randomly choose a particular position for the center then calculate the distance between each other’s point. Each point belonging to a given data set with the nearest centroid. The cluster center is updated based on the distance from nearest centroid.

Two features are chosen to fit the k-means and plot the graph to visualize the cluster. For the experiment of k-means, different types of features and cluster are chosen.

**Features: A vs P**

number of clusters, k=3

silhouette score= 0.62

adjusted rand score= 0.65

normalized mutual info score=0.68

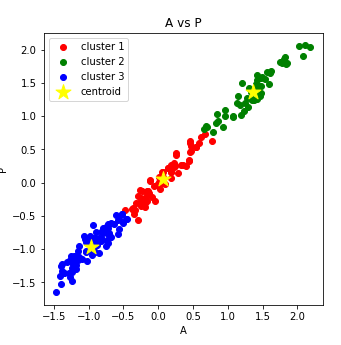


Figure 4: K-means with cluster 3 (A vs P)

**Features: WK and LK**

number of clusters, k=5

silhouette score= 0.39

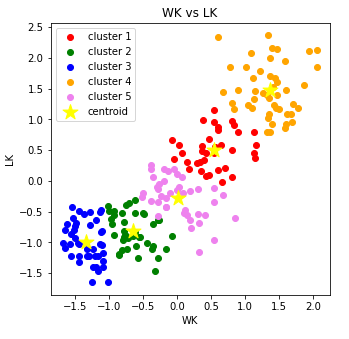


Figure 5: K-means with cluster 5 (WK vs LK)

**Features: C and LKG**

number of clusters, k=7

silhouette score= 0.36

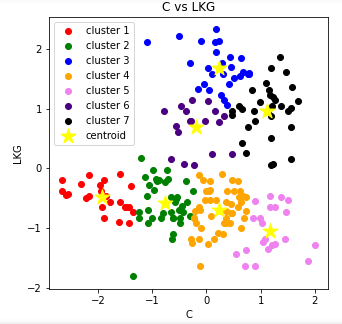


Figure 6: K-means with 7 (C vs LKG)

**3.b. K-Medoid Clustering**

K medoid is one of the important methods of partitioning clustering algorithm. K-medoid is based on calculating the medoids by minimizing the absolute distance between the points and selected centroid rather than minimizing the square distance. In some cases, k- medoids is robust to noise and outliers than the k-means. The data object is randomly split as k cluster and most similar data are group in those cluster.

**Features: A vs P**

number of clusters, k=3

silhouette score= 0.62

adjusted rand score= 0.65

normalized mutual info score=0.67

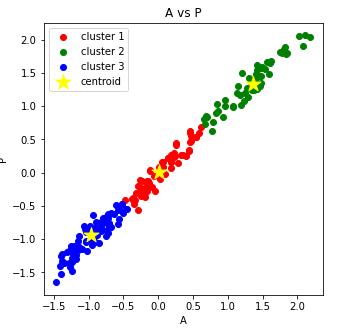


Figure 7: K-medoids with cluster 3 (A vs P)

**Features: WK vs LK**

number of clusters, k=5

silhouette score= 0.40

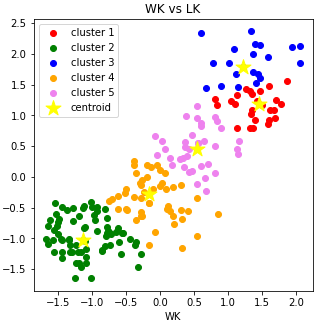


Figure 8: K-medoids with cluster 5 (WK vs LK)

**Features: C vs LKG**

number of clusters, k=7

silhouette score= 0.31

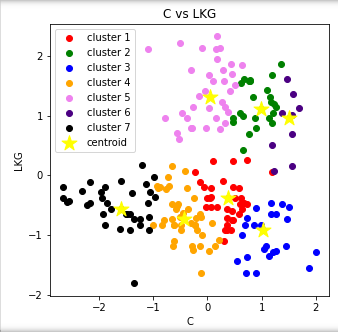


Figure 9: K-medoids with cluster 7 (C vs LKG)

**3.c. BIRCH**

BIRCH algorithm builds a dendrogram known as a clustering feature tree (CF tree). The tree is built by scanning the dataset in an incremental and dynamic way. The algorithm is mainly is separated into two phases. Firstly, build a memory tree and then the algorithm is applied to cluster the leaf nodes. The tree is a height balanced tree which is based on two parameters called branching factor B and threshold T.

**Features: A vs P**

number of clusters, k=3

silhouette score= 0.59

adjusted rand score= 0.55

normalized mutual info score=0.58

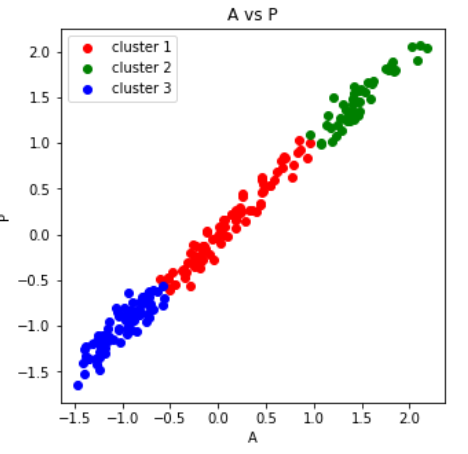


Figure 10: BIRCH with cluster 3 (A vs P)

**Features: WK vs LK**

number of clusters, k=5

silhouette score= 0.38

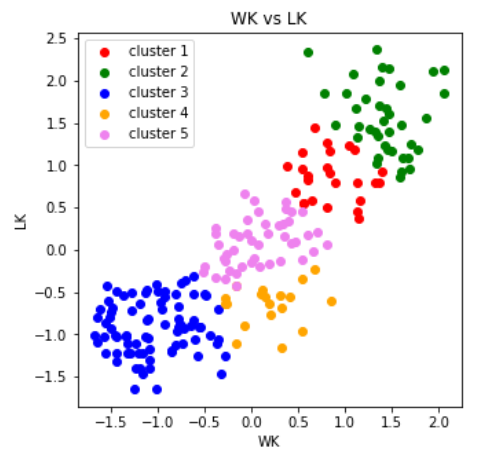


Figure 11: BIRCH with cluster 5 (A vs P)

**Features: C vs LKG**

number of clusters, k=7

silhouette score= 0.32

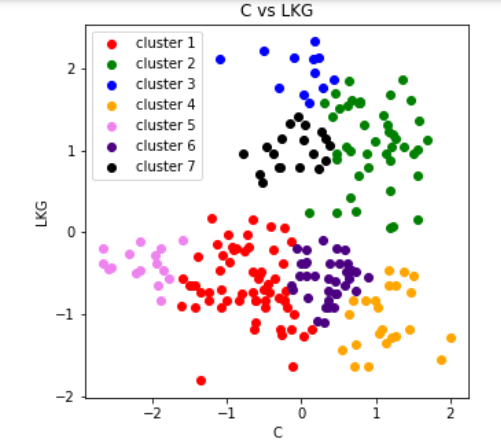


Figure 12: BIRCH with cluster 7 (C vs LKG)

**3.d. Agglomerative clustering**

The agglomerative clustering is one of the common hierarchical clustering algorithms that used to group objects in cluster with similar properties. The algorithm is also known as AGNES (AGglomerative NESting). Initially the object points are declare as a single cluster then successively merged those small cluster to make a large cluster. Dendrogram is used to visualize the tree based agglomerative cluster.

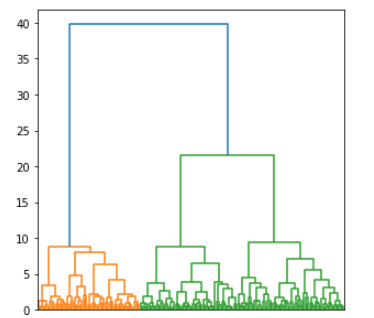


Figure 13: Dendrogram of agglomerative clustering

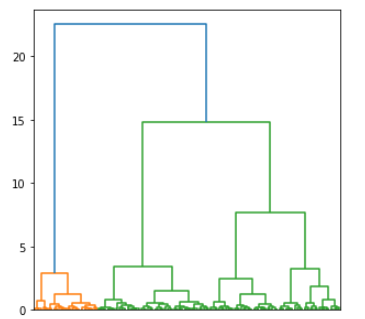
**Features: A vs P**

number of clusters, k=3

silhouette score= 0.54

adjusted rand score= 0.51

normalized mutual info score=0.56

  
Figure 14: Dendrogram of features A vs P

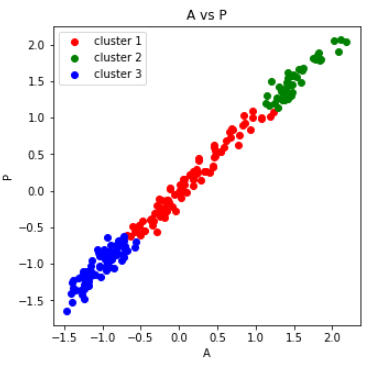


Figure 15: Agglomerative clustering with cluster 3 (A vs P)

**Features: WK vs LK**

number of clusters, k=5

silhouette score= 0.36

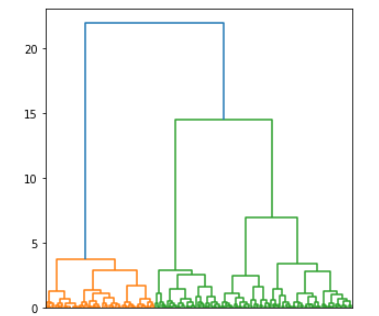


Figure 16: Dendrogram of features WK vs LK

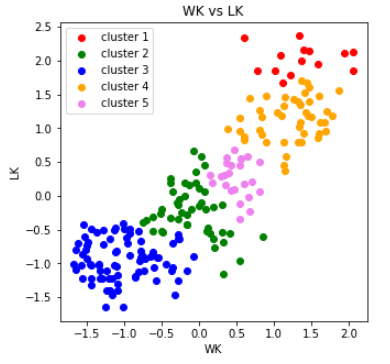


Figure 17: Agglomerative clustering with cluster 5 (WK vs LK)

**Features: C vs LKG**

number of clusters, k=7

silhouette score= 0.15

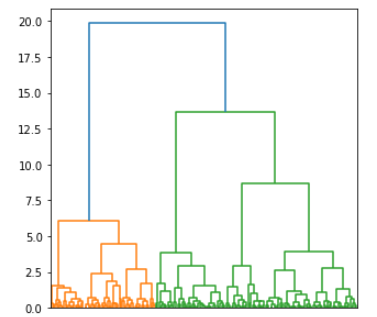


Figure 18: Dendrogram of features WK vs LK

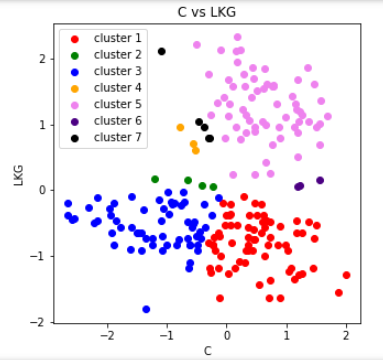


Figure 19: Agglomerative clustering with cluster 7 (C vs LKG)

**3.e. DBSCAN**

DBSCAN is a density based clustering algorithm that stands for Density Based Spatial Clustering of Applications with Noise. It is very useful for large amount of data which is containing noise and outliers.

The main two parameter of DBSCAN:

1. **Eps:** the maximum distance between two samples for one to be considered as in the neighborhood of the other.
2. **Min samples**: the number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself.

**Features: A vs P**

Eps: 0.2

Min samples: 4

silhouette score= 0.33

adjusted rand score= 0.001

normalized mutual info score=0.03

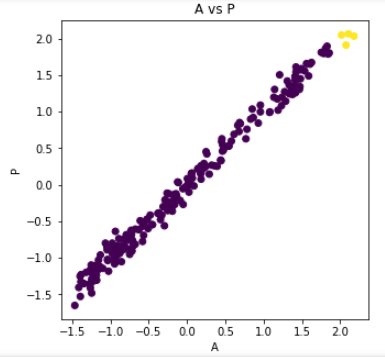


Figure 20: DBSCAN with eps=0.2 and min samples=4 (A vs P)

**Features: WK vs LK**

Eps: 0.3

Min samples: 3

silhouette score= 0.02

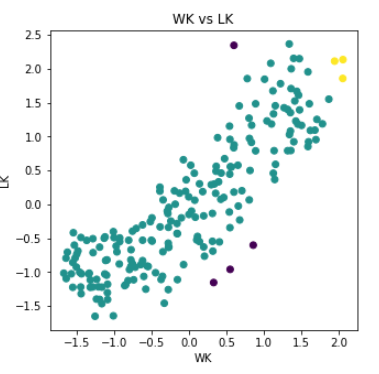


Figure 21: DBSCAN with eps=0.3 and min samples=3 (WK vs LK)

**Features: C vs LKG**

Eps: 0.3

Min samples: 3

silhouette score= 0.15

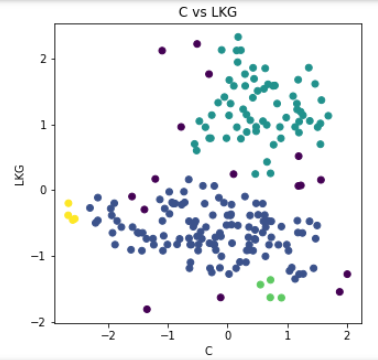


Figure 22: DBSCAN with eps=0.3 and min samples=3 (C vs LKG)

**Features: A\_Coef vs LKG**

Eps: 0.4

Min samples: 3

silhouette score= 0.22

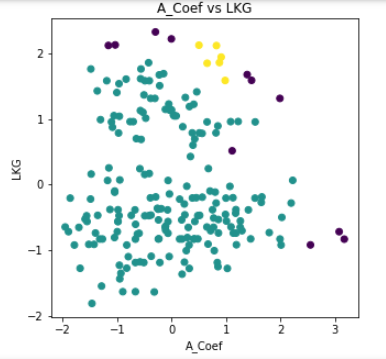


Figure 21: DBSCAN with eps=0.4 and min samples=3 (A\_Coef vs LKG)

**3.f. CHAMELEON**

CHAMELEON is a hierarchical cluster that use dynamic modeling. The algorithm is based on how well connected the object within the cluster and the proximity of cluster. The algorithm is divided into two phases.

1. A graph partitioning algorithm to divide the data set into a set of individual clusters.
2. An agglomerative hierarchical mining algorithm to merge the cluster

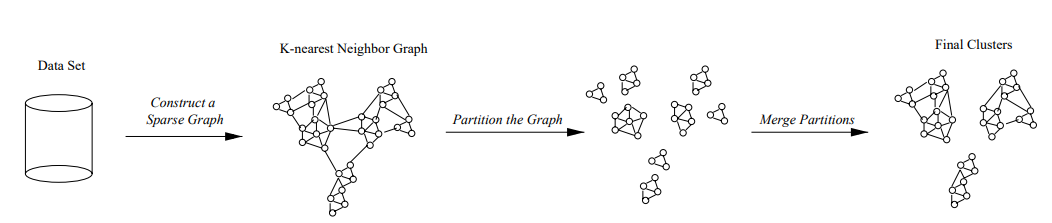


Figure 22: CHAMELEON algorithm (Image source: [1])

K Nearest Neighbor (KNN) is used to construct a sparse graph. The agglomerative clustering algorithm is used to determine the most similar sub cluster by taking the inter connectivity and closeness of the cluster.

**Features: A vs P**

number of clusters, k=3

silhouette score= 0.37

adjusted rand score= 0.48

normalized mutual info score=0.55

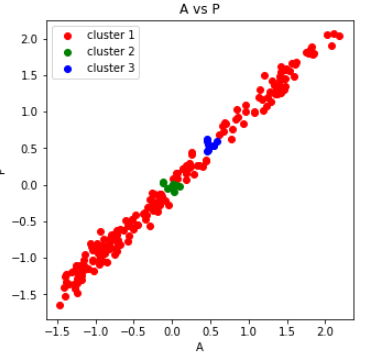


Figure 23: CHAMELEON with cluster 3 (A vs P)

**Features: WK vs LK**

number of clusters, k=5

silhouette score= 0.13

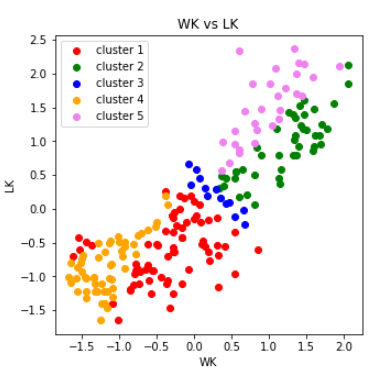


Figure 24: CHAMELEON with cluster 5 (LK vs WK)

**Features: C vs LKG**

number of clusters, k=7

silhouette score= 0.22

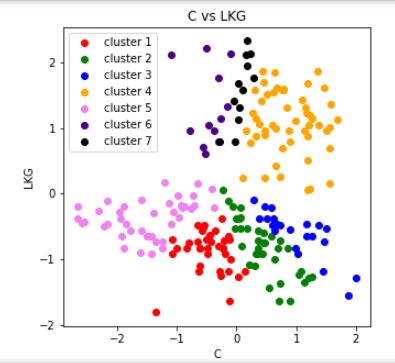


Figure 25: CHAMELEON with cluster 7 (C vs LKG)

**4. Comparative analysis of different cluster**

For the comparative analysis of different clustering algorithm, the approach of different cluster algorithm is different. The clustering algorithm is evaluating on silhouette score, adjusted rand score (ARI) and normalized mutual info score (NMI).

Table 1: Features A and P (Cluster: 3)

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Silhouette score | ARI | NMI |
| K-means | 0.62 | 0.65 | 0.68 |
| K-Medoids | 0.62 | 0.65 | 0.67 |
| BIRCH | 0.59 | 0.55 | 0.58 |
| Agglomerative | 0.54 | 0.51 | 0.56 |
| DBSCAN | 0.33 | 0.001 | 0.02 |
| CHAMELLEON | 0.37 | 0.48 | 0.55 |

Table 2: Features WK and LK (Cluster: 5)

|  |  |
| --- | --- |
| Algorithm | Silhouette score |
| K-means | 0.39 |
| K-Medoids | 0.40 |
| BIRCH | 0.38 |
| Agglomerative | 0.36 |
| DBSCAN | 0.02 |
| CHAMELLEON | 0.13 |

Table 3: Features C and LKG (Cluster: 7)

|  |  |
| --- | --- |
| Algorithm | Silhouette score |
| K-means | 0.36 |
| K-Medoids | 0.31 |
| BIRCH | 0.32 |
| Agglomerative | 0.15 |
| DBSCAN | 0.15 |
| CHAMELLEON | 0.22 |

**5. Conclusion**

Clustering is an unsupervised learning algorithm so that the output level is not predefined. As a result, it is confusing to accurate declare the number of cluster (K) into a particular problem. Evaluation metrics is shows that the score is different for each algorithm. As an example, silhouette score is different for each clustering algorithm on same features of data. If the dataset contains a variance of data, then it is truly saying that no clustering algorithm definitely perform well for all the evaluation criteria. At a time, one algorithm is performing well only a particular problem but not for all the problem.

**6. References**

1. A. Fahad et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," in IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, Sept. 2014, doi: 10.1109/TETC.2014.2330519.
2. Santosh Nirmal, ‘’ Comparative Study between K-Means and K-Medoids Clustering Algorithms,’’ in International Research Journal of Engineering and Technology,2019
3. <https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/>
4. <https://www.kaggle.com/dongeorge/seed-from-uci>
5. <https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80>