**DATASET 1 : SEEDS**

References :

[seeds] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

[sil] <http://finzi.psych.upenn.edu/library/bios2mds/html/sil.score.html>

What it[seeds] says? : They have used Complete Gradient Clustering Algorithm and K-means on this dataset. It correctly classifies 67/70 Rosa, 59/70 Kama and 67/70 Canadian wheats.[seeds]

What I’ve done? : Clustered the geometric parameters data on various other clustering algorithms and compared the result.

K-means : 60/70 Rosa, 60/70 Kama and 68/70 Canadian wheats, 78.4% betweenss/totss

PAM : 57/70 Rosa, 60/70 Kama and 70/70 Canadian wheats

Clara : (Estimated 2 clusters on the basis of avg silhouette width among clusters) 132/140 (Rosa+Canadian) and 66/70 Kama

Hierarchical : 52/70 Rosa, 47/70 Kama and 70/70 Canadian wheats

DBSCAN : 69/70 Rosa, 63/70 Kama and 68/70 Canadian wheats

NbClust using method=ward.D2 : 54/70 Rosa, 63/70 Kama and 70/70 Canadian wheats.

MClust – clusters into 5

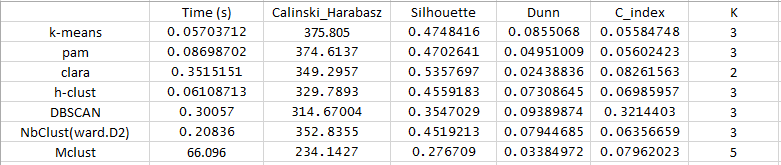
**Conclusion :** The proposed algorithm CGCA clusters the Rosa type of wheat better than other algorithms but fails to distinguish between Kama and Canadian types at several cases. Whereas other clustering algorithms except Clara gives best clusters for Canadian wheat and comparable results for Rosa and Kama

Clara estimated appropriate clusters to be 2 on the basis of average silhouette width. This might be so because of the geometric data used and Canadian samples are lie close to the area of a high density of Rosa wheat variety samples.[seeds]

Silhouette criterion is better used to compare the output of the same clustering algorithm applied to the same data but for different number of clusters.[sil]

Comparison between clustering algorithms

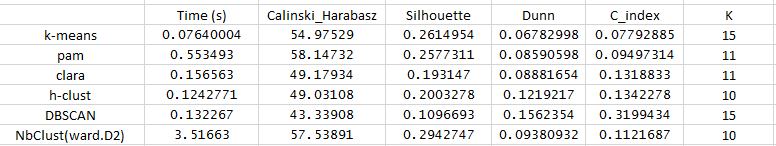
For Seeds Dataset



On the basis of time taken to run the algorithms on the given dataset, the classical k-means is the fastest followed up by h-clust and pam. Other algorithms due to their complexity and accounting of various others internal criterions to estimate optimum number of clusters take up more time such as DBScan and NbClust

The Calinsky-Harabasz and Dunn index shows that k-means, pam and h-clust are very much equally performing and gives similar results. NbClust which used ward.D2 method and ch index proves to be a good algorithm with its performance close to the classical k-means algorithm. The cluster estimation feature of NbClust is best in class as it uses 20+ indexes to estimate optimum clusters medoids. The silhouette index adds more weight to the above observations. Clara due to its lower number of clusters estimated result shows a larger value for silhouette index but the comparison table and Dunn index prove its incapability of estimating correct cluster means. MClust estimates more clusters than expected and takes a lot of time too. This is also shown by the low values of CH and Dunn index.

For Libras Dataset



As the dataset with large number of clusters is input, algorithms like pam and clara show good cluster estimation result when using silhouette width was used as a criterion. But pam takes a slightly higher time of execution.

K-means as usual shows the best result. Nbclust came close to the k-means algorithm with predicting 10 as the optimum number of clusters. But since it uses many indices in computation, it takes greater amount of time of execution.

The Calinski Harbasz value is maximum for pam and nbclust. On the other hand c index is minimum for kmeans and pam. These observations prove that pam clustering is favourable for large values of clusters.

DBSCAN and hclust also work good comparatively. The highest value of Dunn index is for DBSCAN which might be because both work upon the density-based spatial distribution of data.

On the basis of silhouette and C index metrics, nbclust proves to be a very good algorithm apart from k-means and pam in this case.

Model based clustering (mclust) is unable to execute on the full dataset on Intel Core i3 processor. Besides it run quite successfully taking a good amount of time on a sample of the dataset(100-200 observations) but the internal clustering criterions shows that it doesn’t work well with too many cluster centers.