Comparing Different Clustering Algorithms

By Ayush Singh (BT16CSE098)

Introduction

Mini project on comparing the different variations of clustering algorithms. The whole point of clustering is to be given specific features of a data set, and be able to distinguish the elements into different groups or clusters based on those features. There are many different types of these algorithms, with varying run time and accuracy. Specific algorithms perform better or worse relative to the type of features inputted. In this report, we'll be feeding different algorithms to a dataset. Partitioning, hierarchical, density-based, grid-based, and model-based.

Then, we'll be analyzing the algorithms on **standard metrics like**

Metric Name	Knowledge of Groud Truth
Adjusted Rand index	Yes
Mutual Information based scores	Yes
Homogeneity, completeness and V-measure	Yes
Fowlkes-Mallows scores	Yes
Silhouette Coefficient	No
Calinski-Harabaz Index	No
Davies-Bouldin Index	No

to see where each algorithm excels.

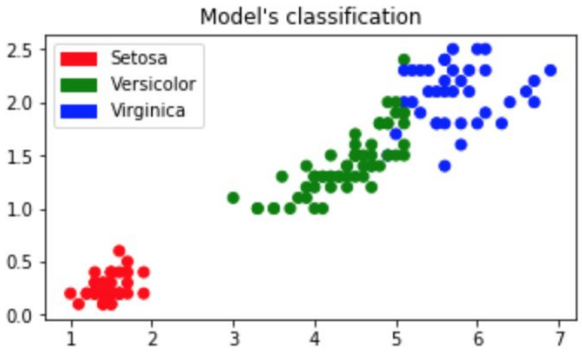
Dataset: IRIS

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

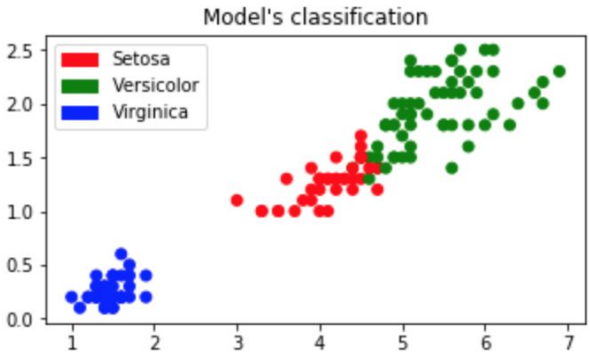
Predicted attribute: class of iris plant.

Algorithms

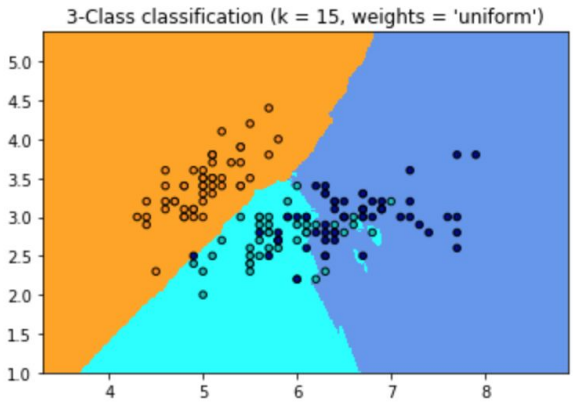
1. K - Means Clustering



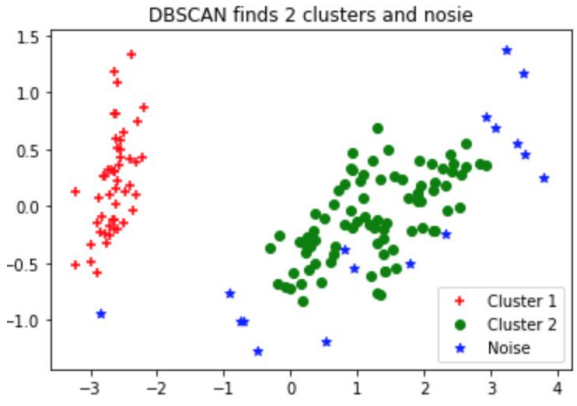
2. K - Medoids Clustering



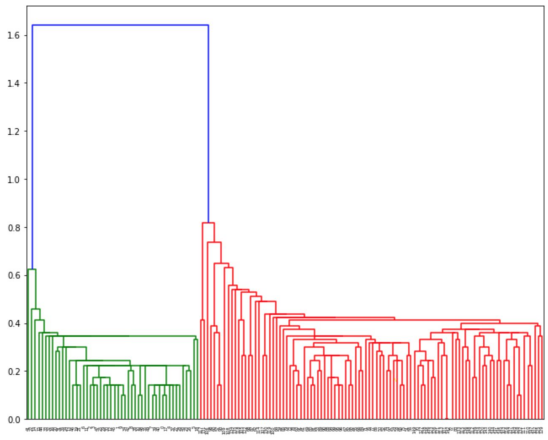
3. K - Nearest Neighbour Clustering



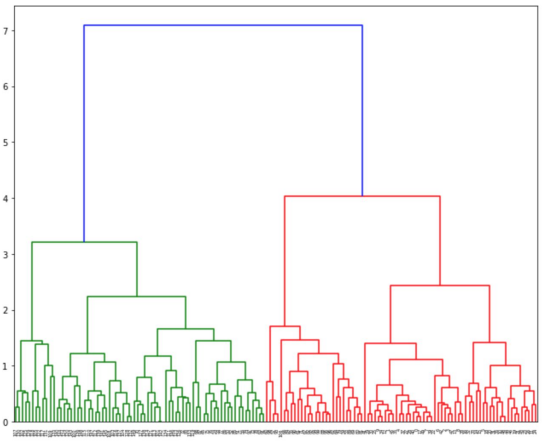
4. DBSCAN



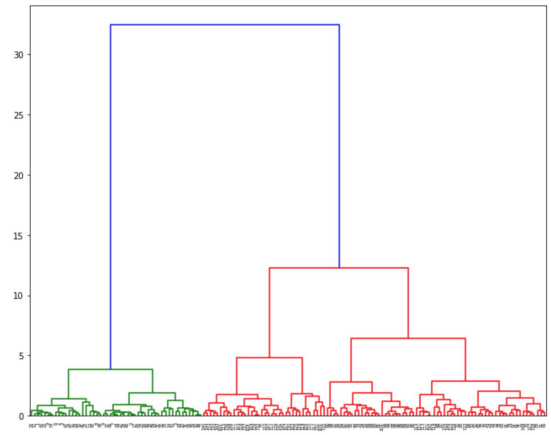
5. AGNES (single linkage)



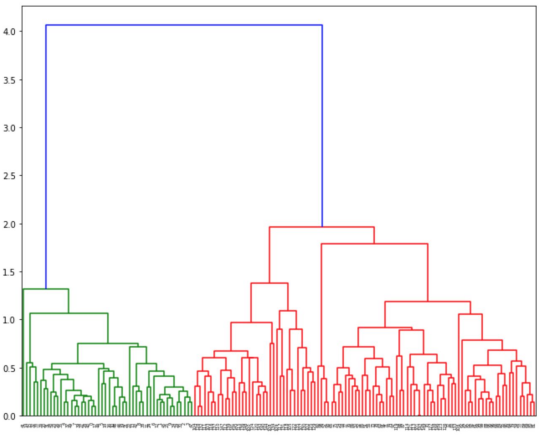
(complete linkage)



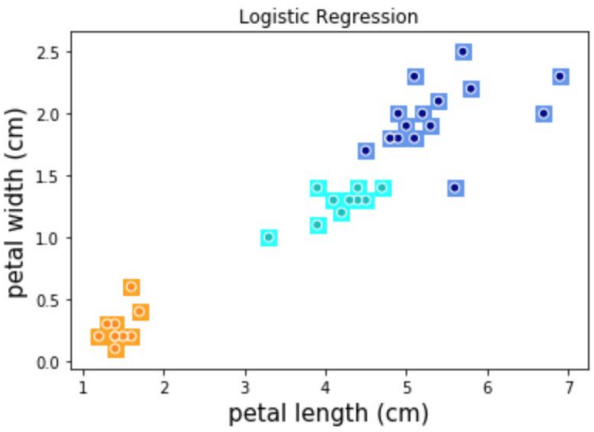
(ward linkage)



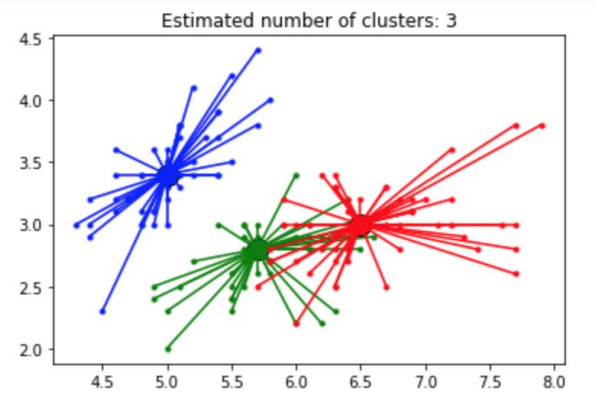
(average linkage)



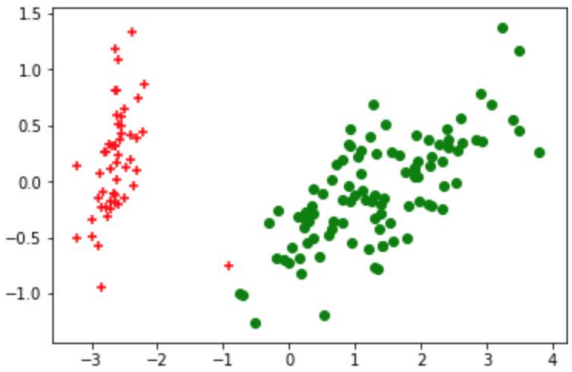
6. Logistic Regression



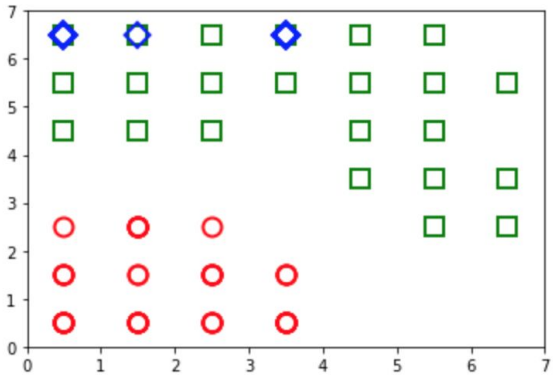
7. Affinity Propagation



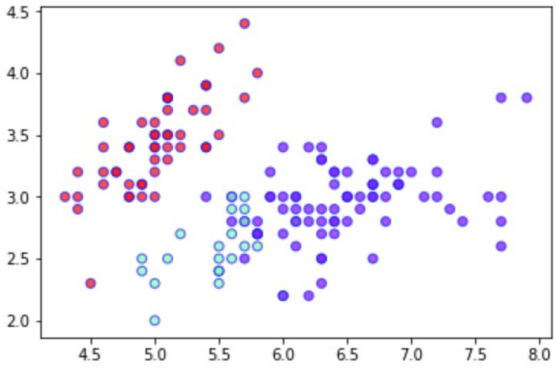
8. Mean Shift Clustering



9. SOM



10. BIRCH



Comparison

	ARI	MI	HCV	FM	SC	CH	DB
K-Means	0.730238	0.748372	0.751485	0.820808	0.552819	561.628	0.661972
K-Medoids	0.758338	0.775961	0.778732	0.839493	0.520198	521.561	0.668624
AGNES - Wand	0.731199	0.757803	0.760801	0.82217	0.554324	558.058	0.656256
AGNES - Single	0.563751	0.582093	0.587916	0.763517	0.512111	277.995	0.447154
AGNES - Average	0.759199	0.793425	0.795982	0.840729	0.554161	556.88	0.658444
AGNES - Complete	0.642251	0.696348	0.700115	0.768637	0.513595	485.905	0.633334
BIRCH	0.609625	0.670611	0.674706	0.751487	0.501952	458.473	0.625831
DBSCAN	0.520619	0.554365	0.559946	0.705385	0.486034	220.298	7.22245
Log. Regression	0.783785	0.815292	0.824974	0.852679	NaN	NaN	NaN
Affinity Propogation	0.802209	0.797547	0.80005	0.867704	0.521386	528.511	0.687832
Mean Shift	0.558371	0.550981	0.553749	0.764206	0.685788	509.703	0.388552

Where

ARI - Adjusted Rand Index
MI - Mutual Information Based Scores
HCV - Homogeneity, Completeness and V-measures
FM - Fowlkes-Mallows Scores
SC - Silhouette Coefficient
CH - Calinski Harabaz Index
DB - Davies Bouldin Index

Note: Refer **comparison.ipynb** for comparison plots of each algorithm.

Code available at https://github.com/AyushSingh098/clustering-comparison