G-02 : Inquiry into Strengths & Limitations of Clustering Methods

Ghanim Alsuwaidi - 22010350
MBZUAI

Introduction

Clustering is a grouping technique employed on a set of observations in a given data-set in which more similar observations are classified as belonging in the same ‘cluster’ based on shared characteristics that sufficiently differentiate an entity from those in different clusters, as per the criterion of similarity or dissimilarity that a given clustering method employs.
Studying the mechanisms of clustering methods is a viable way to gain insights as to which clustering method is best employed upon a given data-set or towards a desired goal.

Challenges

A particular challenge in facilitating this study will be to propose parameter selection methods to best represent the performance of the clustering methods. Furthermore, the ’accuracy’ of a method cannot be endorsed if it is applied on unlabelled data, as it will cluster observations based on properties native to the data, irrespective of any nuance that could be present in real data.

Code Flowchart

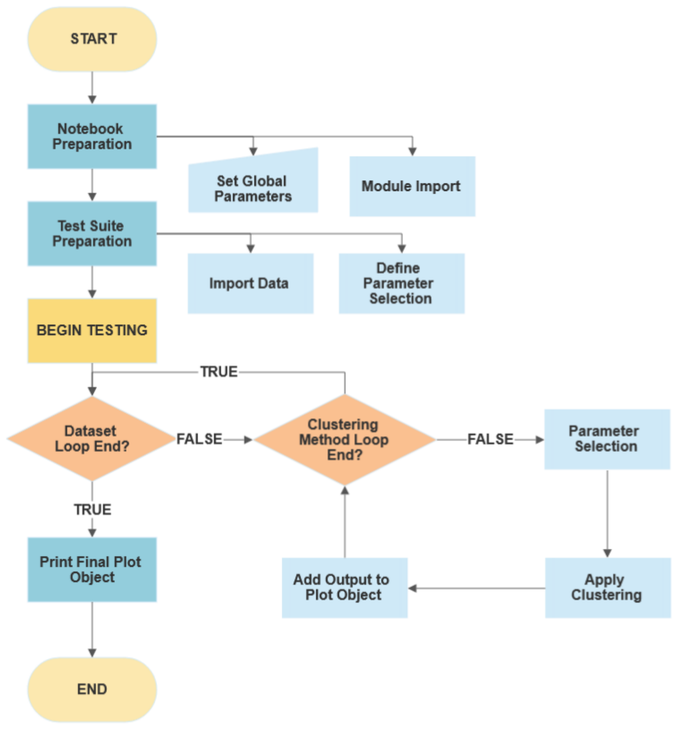

Fig. 1: Nested ’FOR’ loop over distributions and methods.

Test Distributions

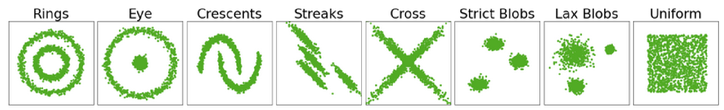

Fig. 2: Distributions used in testing the methods.

Results

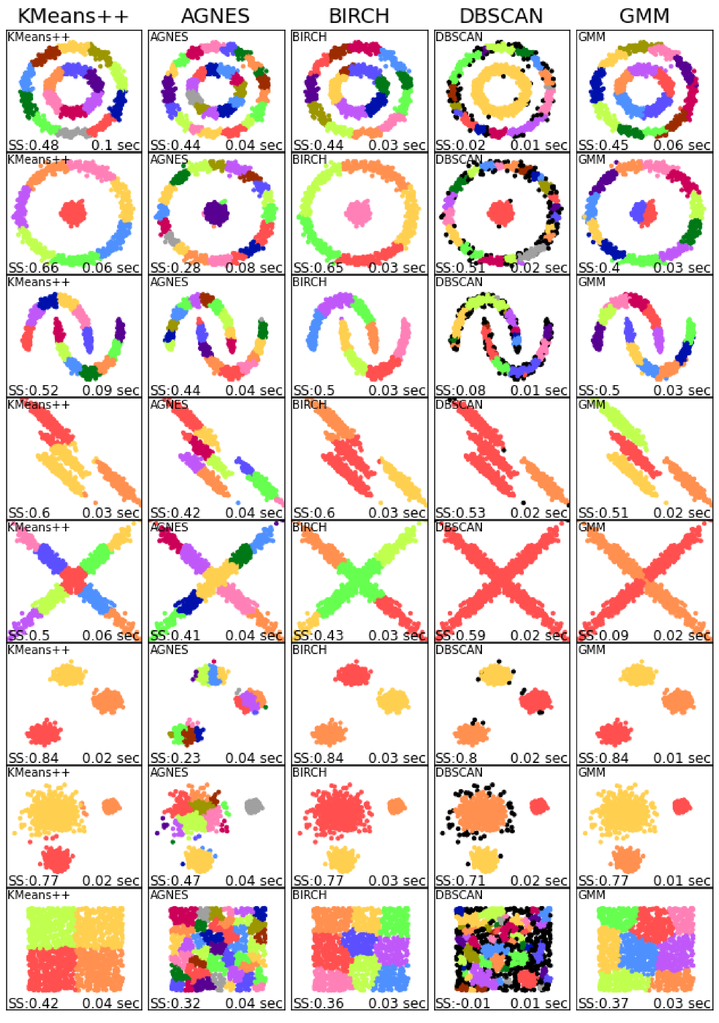

Fig. 3: Clustering output across distributions.

Remarks

The values by the bottom-left corner represents the Silhouette score of the clustering outcome on the data-set, while the value by the lower-right represents the time taken for the clustering method to produce its output on the given distribution.
The black observations are outliers that do not belong to any cluster. Outliers are currently only observed in the **DB-SCAN** clustering outputs.

- **K-Means:** Forms equally sized clusters for all distributions, for better or worse.
- **AGNES:** Highly customizable. Cannot include new data-point without re-clustering.
- **BIRCH:** Highly efficient. Concludes clustering in one read of the data-set. Designed for use with large data. Clustering of new data is performed on the output tree rather than the entire data-set.
- **DB-SCAN:** Cluster output immune to outliers. Is concerned only with the density of data-points, for better or worse.
- **GMM:** Able to identify overlapping clusters. Utilizes ’soft’ designations, based on probabilities.

Parameter Estimation

Automatic parameter estimation is implemented in several ways depending on the mechanism of a clustering method. This is especially the case when considering that the **DB-SCAN** method is not concerned with producing an output that sufficiently separates the clusters, as would be the case with other methods.
As such, it is necessary to implement parameter estimation methods tailored to a given method.

- **K-Means & BIRCH:** Parameter estimation applied for cluster quantity. Implemented by looping the method on iterating values of cluster quantity, and computing the [sec:SS]Silhouette Score of the output. The cluster quantity yielding the highest score is selected.
- **AGNES & DB-SCAN:** Parameter estimation applied for dissimilarity and epsilon threshold respectively. Implemented via the use of k-nearest neighbours to identify a cut-off point where any further increase of the threshold requires the inclusion of outliers to the clusters. When plotting the threshold versus samples, a discernible ’knee’ or ’elbow’ is usually visible, which identifies this cut-off. The selection is implemented by selecting the point on the line with the most prominent curve.
- **GMM:** Parameter estimation applied for cluster quantity. Implemented by looping the method on iterating values of cluster quantity, and computing the **BIC** (Bayesian Information Criterion) of the output. The cluster quantity yielding the lowest score is selected.