Assigning Confidence to Clustering Algorithms

Aamal Hussain

February 26, 2020

- Clustering algorithms in general, with a focus on route clustering
- Adding confidence bounds to clustering algorithms

1 Clustering of time series data: A survey

In this case, the data features changes over time. These are divided into five categories

1. Partitioning: Constructs
$$k \leq n$$
partitions of the n data tuples. Each partition represents a cluster containing at least one object. This allows for objects to be in clusters with different degrees Allows for the use of fuzzy methods to calculate routes off fuzzy sets.
    - Fuzzy c-means
    - Fuzzy c-medioids
2. Hierarchical: Groups data into trees of clusters.
3. Density Based: Grow a cluster as long as the density in the neighbourhood exceeds some threshold
4. Grid based: Quantise the space into cells and perform clustering on these cells
5. Model Based: Assumes a model for the data Would allow for Bayesian clustering, which would allow for the 'no class' property to be considered.

Clustering for time series data the following methods are mentioned

1. Relocation clustering: Uses a particular criterion function and works by comparing the resulting function with members attached in different clusters. This would also allow for the 'no class' option to be considered. This method works only when the time series data are of equal length.
2. Agglomerative Hierarchical: Places each object into its own cluster and then merges these into larger clusters by trying to minimise the sum of squares variance. This method falters once clusters are chosen since it is unable to adjust
3. Fuzzy k-means considers using a fuzzy membership scheme for partitions. By warping it is possible to make these more appropriate for time series data of unequal length, but then the distance metric needs to account for this.