Assigning Confidence to Clustering Algorithms

Aamal Hussain

March 9, 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.

The major problem here concerns the metric used to judge the distance between two paths. This is particularly a problem in route clustering as we have to take into account issues of varying traffic conditions etc.

The modalities for distance/similarity measures which are proposed are

- Euclidean distance (Obviously if time is not taken into account this is shite)
- Cross correlation coefficients (such as Pearson's rank etc.). Can also be a function of time (Can also use these as the metric which considers the confidence of attaching a route to a cluster.)

- Short time series distance (STS): In which each time series is considered as a piecewise linear function
- Dynamic time warping distance. First align the time series and then use Euclidean distance
- KL divergence (I quite like this one if the data sets are somewhat aligned first)
- J divergence/Chernoff divergence

The paper then suggests methods for clustering time series data. I focus on the ones which use the raw time series data since that's what we have. (The others are feature based approaches and model based approaches).

2 Adding Confidence to Gene Expression Clustering

In this study, the authors consider how to cluster gene expression data which contains large amounts of random variation. I anticipate that this would be the same case in the GPS data which will be subject to noise. Importantly, they are able to assign a statistical significance metric when clustering to improve the reliability of the cluster. Previous work relies on doing this by counting how many times a gene is assigned to the same cluster (seems long). Other recent work which has used GWN to perturb data assigns a perturbed cluster to the original iff it contains a majority of elements in common with the original cluster. (If there is a way to extract features from the GPS data, we could consider this as a method of evaluation, e.g. if routes passed under the same 10 bridges).

Their dissimilarity metric, known as the 'Munneke metric' is able to take into account dissimilarity based on the magnitude and the pattern of gene expression (Perhaps such a combination of metrics would also be useful in route clustering). Importantly, however, this metric provides a strong confidence measure of the clustering algorithm as it provides a non-negative real value to measure the dissimilarity between the data and a particular cluster.