Clustering

H. Cantzler
Institute for Perception, Action and Behaviour,
Division of Informatics, University of Edinburgh,
Edinburgh, EH1 2QL, UK
helmutc@dai.ed.ac.uk

Organising observed multi-dimensional data into meaningful structures is a common task which is vital in many scientific and commercial fields. Cluster techniques are utilised to divide a large set of objects into separate classes (also called clusters, groups or partitions) of densely populated regions in the data space.

The data space is usually not uniformly occupied. Generally, clustering techniques are to be seen as tools for the exploration of the data space. They identify the sparse and the crowded places and hence discover the overall pattern of the dataset distribution. The dataset is split according to some object variables, which are frequently the result of measurements.

The two major types of classification techniques are non-hierarchical (partitioning) and hierarchical clustering (tree clustering) [1]. The best known partitioning technique is k-means partitioning. K-means partitioning based on initially specifying a number of classes. Each class has a seed point and all objects within a prescribed distance are included in that class. Objects are moved between those classes with the goal of minimising variability within classes and maximising variability between classes. The most well-known criteria is to minimise the sum of the squared distances between all elements of a class.

Hierarchical clustering results in hierarchies of nested partitions. The clusters themselves are repeatedly grouped to larger clusters. In contrast to the non hierarchical techniques, these clusters are not defined a priori but are created by the clustering algorithm. Data spaces from different clustering problems may have different mathematical properties that influence which

clustering strategy may be applied. Therefore, the distance function (similarity between two objects in the dataset) and the chosen clustering strategy greatly determine the resulting clustering. We focus in this appendix on the hierarchical clustering techniques as they were used in this research.

1 Hierarchical clustering

Hierarchical cluster techniques are used when a stratified structure of clusters at different heterogeneity levels are desired. Divisive clustering (top-down) starts from the entire data set and iteratively splits it until every class consists of one object only. Agglomerative clustering (bottom-up) goes the other way around, as follows. First, each object represents its own cluster. Then, the distance function is used to find the pair of clusters which is most similar (closest distance to each other). This pair is merged to form a new bigger cluster. So, clusters are grouped together to form larger and larger clusters. As clusters get larger and larger more distant clusters are linked together, but their elements become increasingly dissimilar. At every stage one wants the two most similar clusters to be merged. The algorithm terminates when all objects are combined to one cluster (the entire data set) or the degree of dissimilarity reaches a certain threshold.

2 Distance measurements

As a fundamental requirement, a notion of distance has to be introduced in the object space. This means we need to define the similarity or distance function between the individual objects of a data set. This distance depends on the mathematical properties of the data space. It can be based on a single dimension or multiple dimensions. The most common way of computing distances between the two objects o_1 and o_2 in a multi-dimensional space is to compute the Euclidean distance $\sqrt{(o_1 - o_2)^2}$. However, often other derived measures of distance are more suitable for applications.

The most commonly used Euclidean distance can be squared to get the Squared Euclidean distance. This distance places progressively greater weight on objects that are further apart. Many other distance measurements such as the Manhattan (City-block) and Chebychev distance can be used as a similarity measurement. Usually, all the distances between the objects are

calculated once and are then saved in a distance matrix.

3 Linkage algorithms

The measurements in the distance matrix are distances between single objects in our data set. However, once several objects have been grouped together in clusters, how do we determine the distance between those clusters? We need to define the distance between clusters as well. There are various possible linkage algorithms which differ in how they derive cluster distances from the distances of the objects.

Single linkage clustering defines the distance between two clusters as the minimal distance of any two objects belonging to different clusters (nearest neighbour method). Single linkage clustering is best suited to detect chains or elongated structures. However, it is less suitable for isolating spherical or poorly separated clusters. Complete linkage clustering is opposite to single linkage clustering as it uses the maximal distance of objects in different clusters (furthest neighbours). All entries in a cluster are linked to one another within some minimum similarity. This method usually performs well in cases when the objects actually form naturally distinct clumps in the data space. The resulting clusters have spherical shapes, where all members of a class are tightly bound together. This method is inappropriate if the data tend to form rather elongated clusters. In between the two previous methods is the average linkage clustering. It uses the average distance of the pairwise links between the two clusters based on all objects in the clusters. The resulting clusters are intermediate in tightness of single linkage and complete linkage. Many other methods are possible. Some use the centroids or medians to decide which clusters to merge.

References

[1] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, 1988.