Hierarchical Clustering

Fionn Murtagh
Department of Computer Science, Royal Holloway, University of London
Egham TW20 0EX, England

June 4, 2016

Hierarchical clustering algorithms can be characterized as *greedy* (Horowitz and Sahni, 1979). A sequence of irreversible algorithm steps is used to construct the desired data structure. Assume that a pair of clusters, including possibly singletons, is merged or agglomerated at each step of the algorithm. Then the following are equivalent views of the same output structure constructed on n objects: a set of $n - 1$ partitions, starting with the fine partition consisting of n classes and ending with the trivial partition consisting of just one class, the entire object set; a binary tree (one or two child nodes at each non-terminal node) commonly referred to as a dendrogram; a partially ordered set (poset) which is a subset of the power set of the n objects; and an ultrametric topology on the n objects. For background, the reader is referred to Benzécri (1979), Lerman (1981), Murtagh and Heck (1987), Jain and Dubes (1988), Arabie et al. (1996), Mirkin (1996), Gordon (1999), Jain, Murty and Flynn (1999), and Xu and Wunsch (2005).

One could say with justice that Sibson (1973), Rohlf (1982) and Defays (1977) are part of the prehistory of clustering. Their $O(n^2)$ implementations of the single link method and of a (non-unique) complete link method have been widely cited.

In the early 1980s a range of significant improvements were made to the Lance-Williams, or related, dissimilarity update schema (de Rham, 1980; Juan, 1982), which had been in wide use since the mid-1960s. Murtagh (1983, 1985) presents a survey of these algorithmic improvements. The algorithms, which have the potential for *exactly* replicating results found in the classical but more computationally expensive way, are based on the construction of *nearest neighbor chains* and *reciprocal* or mutual NNs (NN-chains and RNNs).

A NN-chain consists of an arbitrary point (a in Fig. **??**); followed by its NN (b in Fig. **??**); followed by the NN from among the remaining points (c, d, and e in Fig. **??**) of this second point; and so on until we necessarily have some pair of points which can be termed reciprocal or mutual NNs. (Such a pair of RNNs may be the first two points in the chain; and we have assumed that no two dissimilarities are equal.)

In constructing a NN-chain, irrespective of the starting point, we may agglomerate a pair of RNNs as soon as they are found. What guarantees that

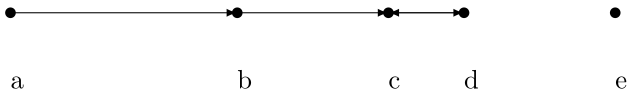

Figure 1: Fig 1. Five points, showing NNs and RNNs.

we can arrive at the same hierarchy as if we used traditional "stored dissimilarities" or "stored data" algorithms (Anderberg, 1973)? Essentially this is the same condition as that under which no inversions or reversals are produced by the clustering method. This would be where s is agglomerated at a lower criterion value (i.e. dissimilarity) than was the case at the previous agglomeration between q and r. Our ambient space has thus contracted because of the agglomeration. This is due to the algorithm used – in particular the agglomeration criterion – and it is something we would normally wish to avoid.

This is formulated as:

$$\text{Inversion impossible if: } d(i,j) < d(i,k) \text{ or } d(j,k) \Rightarrow d(i,j) < d(i \cup j, k)$$

This is Bruynooghe's *reducibility property* (Bruynooghe, 1977; see also Murtagh, 1985, 1992). Using the Lance-Williams dissimilarity update formula, it can be shown that the minimum variance method does not give rise to inversions; neither do the (single, complete, average) linkage methods; but the median and centroid methods cannot be guaranteed not to have inversions.

To return to Fig. **??**, if we are dealing with a clustering criterion which precludes inversions, then c and d can justifiably be agglomerated, since no other point (for example, b or e) could have been agglomerated to either of these.

The processing required, following an agglomeration, is to update the NNs of points such as b in Fig. **??** (and on account of such points, this algorithm was dubbed *algorithme des célibataires* in de Rham, 1980). The following is a summary of the algorithm:

NN-chain algorithm

Step 1: Select a point (i.e. an object in the input data set) arbitrarily.

Step 2: Grow the NN-chain from this point until a pair of RNNs are obtained.

Step 3: Agglomerate these points (replacing with a cluster point, or updating the dissimilarity matrix).

Step 4: From the point which preceded the RNNs (or from any other arbitrary point if the first two points chosen in Steps 1 and 2 constituted a pair of RNNs), return to Step 2 until only one point remains.

In Murtagh (1983, 1985) and Day and Edelsbrunner (1984), one finds discussions of $O(n^2)$ time and $O(n)$ space implementations of Ward's minimum variance (or error sum of squares) method and of the centroid and median methods. The latter two methods are termed the UPGMC and WPGMC criteria (respectively, unweighted and weighted pair-group method using centroids) by Sneath and Sokal (1973). Now, a problem with the cluster criteria used by these latter two methods is that the reducibility property is not satisfied by them. This means that the hierarchy constructed may not be unique as a result of inversions or reversals (non-monotonic variation) in the clustering criterion value determined in the sequence of agglomerations.

Murtagh (1984) describes $O(n^2)$ time and $O(n^2)$ space implementations for the single link method, the complete link method and for the weighted and unweighted group average methods (WPGMA and UPGMA). This approach is quite general vis à vis the dissimilarity used and can also be used for hierarchical clustering methods other than those mentioned.

Day and Edelsbrunner (1984) prove the exact $O(n^2)$ time complexity of the centroid and median methods using an argument related to the combinatorial problem of optimally packing hyperspheres into an m-dimensional volume. They also address the question of metrics: results are valid in a wide class of distances including those associated with the Minkowski metrics.

The construction and maintenance of the nearest neighbor chain as well as the carrying out of agglomerations whenever reciprocal nearest neighbors meet, both offer possibilities for parallelization, and implementation in a distributed fashion. Work in chemoinformatics and information retrieval can be found in Willett (1989), Gillet et al. (1998) and Griffiths et al. (1984). Ward's minimum variance criterion is favored.

For in depth discussion of data encoding and normalization as a preliminary stage of hierarchical clustering, see Murtagh (2005). Finally, as an entry point into the ultrametric view of clustering, and how hierarchical clustering can support constant time, or $O(1)$, proximity search in spaces of arbitrarily high ambient dimensionality, thereby setting aside Bellman's famous curse of dimensionality, see Murtagh (2004).

References

[1] Anderberg, M.R. (1973), *Cluster Analysis for Applications.* Academic Press, New York.

[2] Arabie, P., Hubert, L.J. and De Soete, G. (1996), Eds., *Clustering and Classification*, World Scientific, Singapore.

[3] Benzécri J.P. (1979), *L'Analyse des Données. I. La Taxinomie*, Dunod, Paris (3rd ed.).

[4] Bruynooghe, M. (1977), "Méthodes nouvelles en classification automatique des données taxinomiques nombreuses", Statistique et Analyse des Données, no. 3, 24–42.

[5] Day, W.H.E. and Edelsbrunner, H. (1984), "Efficient algorithms for agglomerative hierarchical clustering methods", Journal of Classification, 1, 7–24.

[6] Defays, D. (1977), "An efficient algorithm for a complete link method", Computer Journal, 20, 364–366.

[7] Gillet, V.J., Wild, D.J., Willett, P. and Bradshaw, J. (1998), "Similarity and dissimilarity methods for processing chemical structure databases", Computer Journal, 41, 547–558.

[8] Gordon, A.D. (1999), Classification, 2nd ed., Champman and Hall.

[9] Griffiths, A., Robinson, L.A. and Willett, P. (1984), "Hierarchic agglomerative clustering methods for automatic document classification", Journal of Documentation, 40, 175–205.

[10] Horowitz, E. and Sahni, S. (1979), *Fundamentals of Computer Algorithms*, Chapter 4 The Greedy Method, Pitman, London.

[11] Jain, A.K. and Dubes, R.C. (1988), *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs.

[12] Jain A.K., Murty, M.N. and Flynn P.J. (1999), "Data clustering: a review", ACM Computing Surveys, 31, 264–323.

[13] Juan, J. (1982), "Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques", Les Cahiers de l'Analyse des Données, VII, 219–225.

[14] Lerman I.C. (1981), *Classification et Analyse Ordinale des Données* Dunod, Paris.

[15] Mirkin B. (1996), *Mathematical Classification and Clustering* Kluwer, Dordrecht.

[16] Murtagh, F. (1983), "A survey of recent advances in hierarchical clustering algorithms", Computer Journal, 26, 354–359.

[17] Murtagh, F. (1985), *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg.

[18] Murtagh, F. and Heck, A. (1987), *Multivariate Data Analysis*, Kluwer Academic, Dordrecht.

[19] Murtagh, F. (1992), "Comments on 'Parallel algorithms for hierarchical clustering and cluster validity"', IEEE Transactions on Pattern Analysis and Machine Intelligence, 14, 1056–1057.

[20] Murtagh F. (2004), “On ultrametricity, data coding, and computation”, Journal of Classification, 21, 167–184.

[21] Murtagh F. (2005), *Correspondence Analysis and Data Coding with Java and R*, Chapman and Hall, Boca Raton.

[22] de Rham, C. (1980), “La classification hiérarchique ascendante selon la méthode des voisins réciproques”, Les Cahiers de l’Analyse des Données, V, 135–144.

[23] Rohlf, F.J. (1982), “Single link clustering algorithms”, in P.R. Krishnaiah and L.N. Kanal, Eds., *Handbook of Statistics*, Vol. 2, North-Holland, Amsterdam, 267–284.

[24] Sibson, R. (1973), “SLINK: an optimally efficient algorithm for the single link cluster method”, The Computer Journal, 16, 30–34.

[25] Sneath, P.H.A. and Sokal, R.R. (1973), *Numerical Taxonomy*, W.H. Freeman, San Francisco.

[26] Willett, P. (1989), “Efficiency of hierarchic agglomerative clustering using the ICL distributed array processor”, Journal of Documentation, 45, 1–45.

[27] Rui Xu and Wunsch D. (2005), “Survey of clustering algorithms”, IEEE Transactions on Neural Networks, 16, 645–678.