Overview

Algorithms for hierarchical clustering: an overview

Fionn Murtagh[1,2]* and Pedro Contreras[2]

We survey agglomerative hierarchical clustering algorithms and discuss efficient implementations that are available in R and other software environments. We look at hierarchical self-organizing maps, and mixture models. We review grid-based clustering, focusing on hierarchical density-based approaches. Finally, we describe a recently developed very efficient (linear time) hierarchical clustering algorithm, which can also be viewed as a hierarchical grid-based algorithm.

How to cite this article:
WIREs Data Mining Knowl Discov 2011. doi: 10.1002/widm.53

INTRODUCTION

Agglomerative hierarchical clustering has been the dominant approach to constructing embedded classification schemes. It is our aim to direct the reader's attention to practical algorithms and methods—both efficient (from the computational and storage points of view) and effective (from the application point of view). It is often helpful to distinguish between *method*, involving a compactness criterion and the target structure of a two-way tree representing the partial order on subsets of the power set, as opposed to an *implementation*, which relates to the detail of the algorithm used.

As with many other multivariate techniques, the objects to be classified have numerical measurements on a set of variables or attributes. Hence, the analysis is carried out on the rows of an array or matrix. If we do not have a matrix of numerical values to begin with, then it may be necessary to skilfully construct such a matrix. The objects, or rows of the matrix, can be viewed as vectors in a multidimensional space (the dimensionality of this space being the number of variables or columns). A geometric framework of this type is not the only one which can be used to formulate clustering algorithms. Suitable alternative forms of storage of a rectangular array of values are not inconsistent with viewing the problem in geometric terms (and in matrix terms, e.g., expressing the adjacency relations in a graph).

Motivation for clustering in general, covering hierarchical clustering and applications, includes the following: analysis of data, interactive user interfaces, storage and retrieval, and pattern recognition.

Surveys of clustering with coverage also of hierarchical clustering include Gordon,[1] March,[2] Jain and Dubes,[3] Gordon,[4] Mirkin,[5] Jain et al.,[6] and Xu and Wunsch.[7] Lerman[8] and Janowitz[9] present overarching reviews of clustering including use of lattices that generalize trees. The case for the central role of hierarchical clustering in information retrieval was made by van Rijsbergen[10] and continued in the work of Willett and coworkers.[11] Various mathematical views of hierarchy, all expressing symmetry in one way or another, are explored by Murtagh.[12]

This paper is organized as follows. In section *Distance, Similarity, and Their Use*, we look at the issue of normalization of data, prior to inducing a hierarchy on the data. In section *Motivation*, some historical remarks and motivation are provided for hierarchical agglomerative clustering. In section *Algorithms*, we discuss the Lance–Williams formulation of a wide range of algorithms, and show how these algorithms can be expressed in graph theoretic terms and in geometric terms. In section *Efficient Hierarchical Clustering Algorithms Using Nearest Neighbor Chains*, we describe the principles of the reciprocal nearest neighbor (RNN) and nearest neighbor (NN) chain algorithm to support building a hierarchical clustering in a more efficient manner compared to the

*Correspondence to: fmurtagh@acm.org
[1]Science Foundation Ireland, Wilton Place, Dublin, Ireland
[2]Department of Computer Science, Royal Holloway, University of London, Egham, UK
DOI: 10.1002/widm.53

Lance-Williams or general geometric approaches. In section *Hierarchical Self-Organizing Maps and Hierarchical Mixture Modeling*, we overview the hierarchical Kohonen self-organizing feature map, and also hierarchical model-based clustering. We conclude this section with some reflections on divisive hierarchical clustering, in general. Section *Density- and Grid-Based Clustering Techniques* surveys developments in grid- and density-based clustering. The following section, *A New, Linear Time Grid Clustering Method: m-Adic Clustering*, presents a recent algorithm of this type, which is particularly suitable for the hierarchical clustering of massive datasets.

DISTANCE, SIMILARITY, AND THEIR USE

Before clustering comes the phase of data measurement, or measurement of the observables. Let us look at some important considerations to be taken into account. These considerations relate to the metric or other spatial embedding, comprising the first phase of the data analysis *stricto sensu*.

To group data we need a way to measure the elements and their distances relative to each other in order to decide which elements belong to a group. This can be a similarity, although on many occasions a dissimilarity measurement, or a 'stronger' distance, is used.

A distance between any pair of vectors or points i, j, k satisfies the properties of symmetry, $d(i, j) = d(j, k)$; positive definiteness, $d(i, j) > 0$ and $d(i, j) = 0$ iff $i = j$; and the triangular inequality, $d(i, j) \leq d(i, k) + d(k, j)$. If the triangular inequality is not taken into account, we have a dissimilarity. Finally, a similarity is given by $s(i, j) = \max_{i,j}\{d(i, j)\} - d(i, j)$.

When working in a vector space, a traditional way to measure distances is a Minkowski distance, which is a family of metrics defined as follows:

$$L_p(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{i=1}^{n} |\mathbf{x}_{i,a} - \mathbf{x}_{i,b}|^p \right)^{1/p} ; \ \forall p \geq 1, \ p \in \mathbb{Z}^+, \tag{1}$$

where $\mathbb{Z}^+$ is the set of positive integers.

The Manhattan, Euclidean, and Chebyshev distances (the latter is also called maximum distance) are special cases of the Minkowski distance when $p = 1$, $p = 2$, and $p \to \infty$.

As an example of similarity, we have the *cosine* similarity, which gives the angle between two vectors. This is widely used in text retrieval to match vector queries to the dataset. The smaller the angle between a query vector and a document vector, the closer a query is to a document. The normalized cosine similarity is defined as follows:

$$s(\mathbf{x}_a, \mathbf{x}_b) = \cos(\theta) = \frac{\mathbf{x}_a \cdot \mathbf{x}_b}{\|\mathbf{x}_a\| \|\mathbf{x}_b\|}, \tag{2}$$

where $\mathbf{x}_a \cdot \mathbf{x}_b$ is the dot product and $\| \cdot \|$ is the norm.

Other relevant distances are the Hellinger, variational, Mahalanobis, and Hamming distances. Anderberg[13] gives a good review of measurement and metrics, where their interrelationships are also discussed. Also, Deza and Deza[14] have produced a comprehensive list of distances in their *Encyclopedia of Distances*.

By mapping our input data into a Euclidean space, where each object is equiweighted, we can use a Euclidean distance for the clustering that follows. Correspondence analysis is very versatile in determining a Euclidean, factor space from a wide range of input data types, including frequency counts, mixed qualitative and quantitative data values, ranks or scores, and others. Further reading on this is to be found in Benzécri[15] and Le Roux and Rouanet,[16] and Murtagh.[17]

AGGLOMERATIVE HIERARCHICAL CLUSTERING

Motivation

Agglomerative hierarchical clustering algorithms can be characterized as *greedy*, in the algorithmic sense. 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 nonterminal 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.

An ultrametric, or tree metric, defines a stronger topology compared to, e.g., a Euclidean metric geometry. For three points, i, j, k, metric and ultrametric respect the properties of symmetry (d, $d(i, j) = d(j, i)$) and positive definiteness ($d(i, j) > 0$ and if $d(i, j) = 0$ then $i = j$). A metric though (as noted in section *Distance, Similarity, and Their Use*) satisfies the triangular inequality, $d(i, j) \leq d(i, k) + d(k, j)$ while

an ultrametric satisfies the strong triangular or ultrametric (or non-Archimedean), inequality, $d(i, j) \leq \max\{d(i, k), d(k, j)\}$. In section *Distance, Similarity, and Their Use*, there was further discussion on metrics.

The single linkage hierarchical clustering approach outputs a set of clusters (to use graph theoretic terminology, a set of maximal connected subgraphs) at each level—or for each threshold value which produces a new partition. The single linkage method with which we begin is one of the oldest methods, its origins being traced to Polish researchers in the 1950s.[18] The name *single linkage* arises as the interconnecting dissimilarity between two clusters or components is defined as the least interconnecting dissimilarity between a member of one and a member of the other. Other hierarchical clustering methods are characterized by other functions of the interconnecting linkage dissimilarities.

As early as the 1970s, it was held that about 75% of all published work on clustering employed hierarchical algorithms.[19] Interpretation of the information contained in a dendrogram is often of one or more of the following kinds: set inclusion relationships, partition of the object sets, and significant clusters.

Much early work on hierarchical clustering was in the field of biological taxonomy, from the 1950s and more so from the 1960s onward. The central reference in this area, the first edition of which dates from the early 1960s, is Ref 20. One major interpretation of hierarchies has been the evolution relationships between the organisms under study. It is hoped, in this context, that a dendrogram provides a sufficiently accurate model of underlying evolutionary progression.

A common interpretation made of hierarchical clustering is to derive a partition. A further type of interpretation is instead to detect maximal (i.e., disjoint) clusters of interest at varying levels of the hierarchy. Such an approach is used by Rapoport and Fillenbaum[21] in a clustering of colors based on semantic attributes. Lerman[8] developed an approach for finding significant clusters at varying levels of a hierarchy, which has been widely applied. By developing a wavelet transform *on* a dendrogram,[22] which amounts to a wavelet transform in the associated ultrametric topological space, the most important—in the sense of best approximating—clusters can be determined. Such an approach is a topological one (i.e., based on sets and their properties) as contrasted with more widely used optimization or statistical approaches.

In summary, a dendrogram collects together many of the proximity and classificatory relationships in a body of data. It is a convenient representation which answers such questions as: 'How many useful groups are in this data?' and 'What are the salient interrelationships present?' But it can be noted that differing answers can feasibly be provided by a dendrogram for most of these questions, depending on the application.

Algorithms

A wide range of agglomerative hierarchical clustering algorithms have been proposed at one time or another. Such hierarchical algorithms may be conveniently broken down into two groups of methods. The first group is that of linkage methods—the single, complete, weighted, and unweighted average linkage methods. These are methods for which a graph representation can be used. Sneath and Sokal[20] may be consulted for many other graph representations of the stages in the construction of hierarchical clusterings.

The second group of hierarchical clustering methods are methods which allow the cluster centers to be specified (as an average or a weighted average of the member vectors of the cluster). These methods include the centroid, median, and minimum variance methods.

The latter may be specified either in terms of dissimilarities, alone, or alternatively in terms of cluster center coordinates and dissimilarities. A very convenient formulation, in dissimilarity terms, which embraces all the hierarchical methods mentioned so far, is the *Lance–Williams dissimilarity update formula*. If points (objects) i and j are agglomerated into cluster $i \cup j$, then we must simply specify the new dissimilarity between the cluster and all other points (objects or clusters). The formula is

$$d(i \cup j, k) = \alpha_i d(i, k) + \alpha_j d(j, k) + \beta d(i, j) + \gamma \mid d(i, k) - d(j, k) \mid,$$

where α_i, α_j, β, and γ define the agglomerative criterion. Values of these are listed in the second column of Table 1. In the case of the single link method, using $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$, and $\gamma = -\frac{1}{2}$ gives us

$$d(i \cup j, k) = \frac{1}{2} d(i, k) + \frac{1}{2} d(j, k) - \frac{1}{2} \mid d(i, k) - d(j, k) \mid,$$

which, it may be verified, can be rewritten as

$$d(i \cup j, k) = \min\{d(i, k), d(j, k)\}.$$

TABLE 1 | Specifications of Seven Hierarchical Clustering Methods

Hierarchical Clustering Methods (and Aliases)	Lance–Williams Dissimilarity Update Formula	Coordinates of Center of Cluster, which Agglomerates Clusters i and j	Dissimilarity between Cluster Centers g_i and g_j
Single link (nearest neighbor)	$\alpha_i = 0.5$ $\beta = 0$ $\gamma = -0.5$ (More simply: $\min\{d_{ik}, d_{jk}\}$)		
Complete link (diameter)	$\alpha_i = 0.5$ $\beta = 0$ $\gamma = 0.5$ (More simply: $\max\{d_{ik}, d_{jk}\}$)		
Group average (average link, UPGMA)	$\alpha_i = \frac{\lvert i \rvert}{\lvert i \rvert + \lvert j \rvert}$ $\beta = 0$ $\gamma = 0$		
McQuitty's method (WPGMA)	$\alpha_i = 0.5$ $\beta = 0$ $\gamma = 0$		
Median method (Gower's, WPGMC)	$\alpha_i = 0.5$ $\beta = -0.25$ $\gamma = 0$	$\mathbf{g} = \frac{\mathbf{g}_i + \mathbf{g}_j}{2}$	$\lVert \mathbf{g}_i - \mathbf{g}_j \rVert^2$
Centroid (UPGMC)	$\alpha_i = \frac{\lvert i \rvert}{\lvert i \rvert + \lvert j \rvert}$ $\beta = -\frac{\lvert i \rvert \lvert j \rvert}{(\lvert i \rvert + \lvert j \rvert)^2}$ $\gamma = 0$	$\mathbf{g} = \frac{\lvert i \rvert \mathbf{g}_i + \lvert j \rvert \mathbf{g}_j}{\lvert i \rvert + \lvert j \rvert}$	$\lVert \mathbf{g}_i - \mathbf{g}_j \rVert^2$
Ward's method (minimum variance, error sum of squares)	$\alpha_i = \frac{\lvert i \rvert + \lvert k \rvert}{\lvert i \rvert + \lvert j \rvert + \lvert k \rvert}$ $\beta = -\frac{\lvert k \rvert}{\lvert i \rvert + \lvert j \rvert + \lvert k \rvert}$ $\gamma = 0$	$\mathbf{g} = \frac{\lvert i \rvert \mathbf{g}_i + \lvert j \rvert \mathbf{g}_j}{\lvert i \rvert + \lvert j \rvert}$	$\frac{\lvert i \rvert \lvert j \rvert}{\lvert i \rvert + \lvert j \rvert} \lVert \mathbf{g}_i - \mathbf{g}_j \rVert^2$

$|i|$ is the number of objects in cluster i; $\mathbf{g}_i$ is a vector in m-space (m is the set of attributes), either an initial point or a cluster center; $\|.\|$ is the norm in the Euclidean metric. The names UPGMA, etc. are because of Sneath and Sokal.[20] Coefficient α_j, with index j, is defined identically to coefficient α_i with index i. Finally, the Lance and Williams recurrence formula is (with $|.|$ expressing absolute value)

$$d_{i \cup j,k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma \mid d_{ik} - d_{jk} \mid .$$

Using other update formulas, as given in column 2 of Table 1, allows the other agglomerative methods to be implemented in a very similar way to the implementation of the single link method.

In the case of the methods which use cluster centers, we have the center coordinates (in column 3 of Table 1) and dissimilarities as defined between cluster centers (column 4 of Table 1). The Euclidean distance must be used for equivalence between the two approaches. In the case of the *median method*, for instance, we have the following (cf. Table 1).

Let $\mathbf{a}$ and $\mathbf{b}$ be two points (i.e., m-dimensional vectors: these are objects or cluster centers) which have been agglomerated, and let $\mathbf{c}$ be another point. From the Lance–Williams dissimilarity update formula, using squared Euclidean distances, we have

$$\begin{aligned} d^2(a \cup b, c) &= \frac{d^2(a,c)}{2} + \frac{d^2(b,c)}{2} - \frac{d^2(a,b)}{4} \\ &= \frac{\|\mathbf{a} - \mathbf{c}\|^2}{2} + \frac{\|\mathbf{b} - \mathbf{c}\|^2}{2} - \frac{\|\mathbf{a} - \mathbf{b}\|^2}{4}. \end{aligned} \quad (3)$$

The new cluster center is $(\mathbf{a} + \mathbf{b})/2$, so that its distance to point $\mathbf{c}$ is

$$\left\| \mathbf{c} - \frac{\mathbf{a} + \mathbf{b}}{2} \right\|^2. \quad (4)$$

That these two expressions are identical is readily verified. The correspondence between these two perspectives on the one agglomerative criterion is

