View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Elsevier - Publisher Connector

Discrete Applied Mathematics 7 (1984) 191–199
North-Holland

COUNTING DENDROGRAMS: A SURVEY

Fionn MURTAGH
Department of Computer Science, University College Dublin, Dublin 4, Ireland

Received 7 December 1981
Revised 17 November 1982

A dendrogram is a tree representation of data, used in hierarchical cluster analysis. The enumeration of non-isomorphic dendrograms, with specified numbers of terminal or leaf nodes, is the problem addressed here. A number of sub-classes of this problem are distinguished, arising out of whether or not a dendrogram is considered to be binary, labelled and ranked, and results are reviewed for each.

1. Introduction

A *hierarchic classification* is a sequence of partitions of a set of n objects, starting with the partition into n one-object clusters, and successively merging two or more clusters until the final partition consists of one n-object cluster. A *dendrogram* may be defined as a rooted tree, where each of the n terminal nodes represents an object, where each non-terminal node represents a non-singleton cluster, and finally where the root node represents the entire object-set. If there are precisely $n-1$ merges in the hierarchic clustering, the corresponding dendrogram is said to be binary: each non-terminal node has precisely two offspring nodes. Examples of binary dendrograms to be discussed are shown in Figs. 1, 2, and 3.

Some introductions to cluster analysis justify hierarchic clustering on the grounds that the largeness of the number of partitionings of n objects into m groups rules out exhaustive search of all possible partitionings ([1, p. 3] or [4, p. 334]). It is of interest to pursue this argument to the combinatorial study of dendrograms. Other introductions to cluster analysis (referenced in later sections) give enumeration results, but for some particular definition of dendrogram only. The present paper gathers together a number of scattered studies in this area. In all, results are given for enumerating 7 different definitions of dendrogram, and a previously unremarked link with the problem of alternating permutations is noted.

2. Principal types of dendrogram

A first major characteristic of dendrograms, as has already been noted, is whether they are binary or not. There are precisely $n-1$ non-terminal nodes in a binary

0166-218X/84/$3.00 © 1984, Elsevier Science Publishers B.V. (North-Holland)

dendrogram. Many clustering programs output dendrograms in this form (e.g. the widely used CLUSTAN package), and most of the dendrograms to be discussed fall into the binary category.

A second characteristic of dendrograms is whether or not the terminal nodes are labelled.

A third, and final, characteristic is whether or not ranks (or level values) are associated with the nodes of the dendrogram. Either the rank values of clusters, or alternatively only the embedded or nested structure of clusters, are taken into account. This breakdown of dendrograms between ranked and non-ranked is the same as that used by Sibson [16] in characterising dendrograms as locally or globally order invariant.

For binary dendrograms we will review the following cases:

- labelled, ranked (L-R),
- labelled, non-ranked (L-NR),
- unlabelled, non-ranked (NL-NR),
- unlabelled, ranked (NL-R),

and for non-binary dendrograms, results will be given for:

- labelled, ranked,
- labelled, non-ranked.

Finally a type of binary dendrogram which will be called quasi-labelled, non-ranked will be looked at.

To illustrate these definitions of dendrograms, Fig. 1 shows 5 binary dendrograms. Dendrograms (i) and (ii) are identical when considered as NL-R dendrograms; but considered as L-R dendrograms, they are non-isomorphic due to the relative positionings of labels *a*, *b*, and *c*. In the context of NL-NR dendrograms, all the dendrograms shown in Fig. 1 are isomorphic.

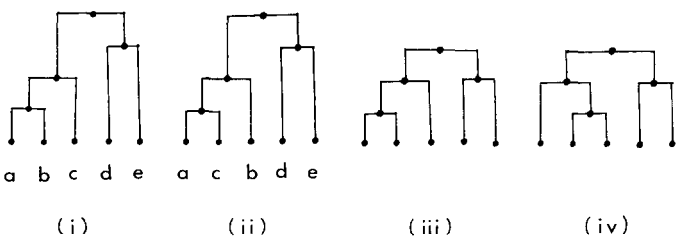

Fig. 1. Four dendrograms used to describe different types (see text).

3. Labelled, ranked, binary dendrograms

If $a(n)$ is the number of L-R dendrograms on n terminal nodes, the following

formula is given in, amongst others, [7], [11, p. 24], and [13, p. 342]:

$$a(n) = n!(n-1)!/2^{n-1}.$$

This is obtained from the product of the $\binom{n}{2}$ ways to choose the first cluster, $\binom{n-1}{2}$ ways to choose the second, and so on until the final agglomeration, for which there are $\binom{2}{2}$ possibilities. It follows that a recurrence relation is given by

$$a(n) = \binom{n}{2} a(n-1), \qquad a(1) = 1.$$

As an example, in the case of $n = 5$ when $a(n) = 180$, Fig. 2 shows a set of 5 dendrograms. There are 60 possible labellings for dendrogram (i), and 30 (i.e. number of successive choices of 2 labels, 1 label, and 2 labels) for each of the remainder.

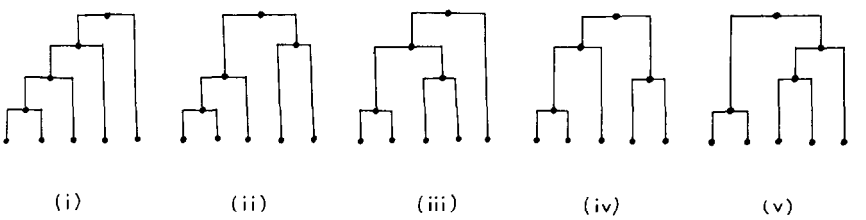

Fig. 2. Five dendrograms on $n = 5$.

4. Labelled, non-ranked, binary dendrograms

Non-ranked dendrograms may be viewed in terms of bracketing problems (see [3, p. 52ff]). The hierarchic clustering of 6 objects

$$\{(a, b),\ (d, e),\ (a, b, c),\ (a, b, c, d, e),\ (a, b, c, d, e, f)\}$$

may be represented as

$$((((ab)c)(de))f)$$

if no distinction is made between the ranks of (a, b) and (d, e), for example.

Let $b(n)$ be the number of non-isomorphic L-NR dendrograms on n terminal nodes. A recurrence relation given by [8], and varying slightly from the form given in [2, pp. 200–205], is

$$b(n) = \tfrac{1}{2} \sum \left\{ \binom{n}{k} b(k) b(n-k) \colon k = 1, \ldots, n-1 \right\},$$

$$b(1) = 1.$$

Here, the first k labelled objects can be chosen in $\binom{n}{k}$ ways and clustered, or bracketed, in $b(k)$ ways; the $(n-k)$ remaining objects can be bracketed in $b(n-k)$ different ways. When all possibilities as k ranges over 1 to $n-1$ are collected, the resulting L-NR dendrograms are doubly counted which gives rise to the factor of $\frac{1}{2}$.

In Fig. 2, as an example, dendrograms (ii), (iv), and (v) are isomorphic if they are taken as non-ranked. As in the case of L-R dendrograms, there are 60 possible labellings for dendrogram (i), and 30 for dendrogram (ii). In the case of dendrogram (iii) the choice in labelling either pair of objects which are first clustered leads to half the result obtained previously (i.e. 30). Totalling, then, gives $b(5) = 105$.

A more simple recurrence relation may also be obtained for enumerating L-NR dendrograms (a number of references are given in [12]):

$$b(n) = (2n-3)b(n-1).$$

This is arrived at as follows. Altogether there are $2(n-1)-1$ nodes in a dendrogram on $n-1$ terminal nodes. Consider the nodes numbered in any fashion from 1 to $2n-3$. The addition of an nth node can be made to the dendrogram at a point just above any of these nodes. Hence given a dendrogram on $n-1$ terminal nodes, there are $2n-3$ possible dendrograms on n terminal nodes. It follows that

$$b(n) = \prod \{2k-3 \colon k = 2, \ldots, n\}.$$

Another formula is given in [8]:

$$b(n) = (2n-2)!/(2^{n-1}(n-1)!).$$

Correcting the proof in [8], we use the generating function

$$\phi(x) = \sum \{b(n)x^n/n! \colon n = 1, 2, \ldots\}$$

which, using the first recurrence relation for $b(n)$ above, multiplying both sides by $x^n/n!$, and summing from 2 to infinity, gives

$$2(\phi - x) = \phi^2.$$

Solving for ϕ, and using the binomial expansion of the square root, gives as coefficient of $x^n/n!$ the above $b(n)$.

5. Unlabelled, non-ranked, binary dendrograms

As in the previous case, this may be viewed as a bracketing problem. The following recurrence relation is given in [3, pp. 54–55], [8], and [9]:

$$c(n) = \sum \{c(k)c(n-k) \colon k = 1, \ldots, \tfrac{1}{2}(n-1)\} \quad \text{for } n \text{ odd},$$

$$c(n) = \sum \{c(k)c(n-k) \colon k = 1, \ldots, \tfrac{1}{2}n - 1\} + \tfrac{1}{2}c(\tfrac{1}{2}n)(c(\tfrac{1}{2}n)+1) \quad \text{for } n \text{ even},$$

$$c(1) = c(2) = 1.$$

The proof of this is as follows. If n is odd, then k unlabelled objects are selected and clustered in $c(k)$ ways, and there are $c(n-k)$ ways to cluster the remaining $n-k$ objects. Index k ranges over 1 to $\frac{1}{2}(n-1)$, only, so as not to doubly count the dendrograms. In the case of n even, the term corresponding to $k=\frac{1}{2}n$ arises when the dendrogram on n terminal nodes connects together two subdendrograms on $\frac{1}{2}n$ terminal nodes. These two subdendrograms must be chosen from $c(\frac{1}{2}n)$ possible non-isomorphic types, repetition being allowed. Since the number of choices of two things out of x, with repetition, is $\binom{x+1}{2}$ we arrive at the required term.

As an example for $n=5$, Fig. 2 shows the three non-isomorphic NL-NR dendrograms: (ii), (iv), and (v) are isomorphic.

6. Unlabelled, ranked, binary dendrograms

NL-R dendrograms do not appear to have been examined before, but they present some interesting connections with permutation problems. A recurrence relation for this type of dendrogram can be obtained as follows. Let $d(n,m)$ be the number of NL-R dendrograms on n terminal nodes, such that there are exactly m nodes with two offspring terminal nodes. Thus in Fig. 2, dendrogram (i) is enumerated by $d(5,1)$ and all the others by $d(5,2)$. We then have:

$$d(n,m) = md(n-1,m)+(n-2m+1)d(n-1,m-1),$$

$$d(n,1) = 1, \qquad m \leq \tfrac{1}{2}n.$$

This relation is arrived at by noting that a dendrogram on n terminal nodes is constructed from a dendrogram on $n-1$ terminal nodes by changing a terminal node into a non-terminal node (i.e. by 'appending' 2 new terminal nodes to a former terminal node), and by re-sequencing the ranks. There are m distinct possibilities for changing, in this way, a terminal node which was formerly 'paired'; and to create such a pair of terminal nodes, there are $(n-1)-2(m-1)$ such possibilities.

The number of distinct NL-R dendrograms, $d(n)$, is then given by summing $d(n,m)$ over $m=1,2,\ldots,\frac{1}{2}n$. The numbers $d(n,m)$ are discussed in [6].

The link between NL-R dendrograms and permutation problems – the subject of the last reference – is provided by the 'packed representation' [15] of a hierarhy: for any terminal node indexed by i, with the exception of the rightmost, define $p(i)$ as the rank at which the terminal node is first united with some terminal node to its right. Thus dendrogram (ii) in Fig. 2 may be written as $p=(1243)$, and dendrogram (iii) as $p=(1324)$. Alternatively this permutation may be arrived at by means of the oriented, binary tree whose nodes are given by, and labelled by, the ranks of the dendrogram. In Fig. 3, an inorder traversal of this tree gives the packed representation as 12534687. For all equivalent representations of the one dendrogram, we will use as a representative the hierarchy where the sequence of agglomerations is from left-to-right wherever this is possible (i.e. the non-terminal node of least rank is always in the left subtree of a given node). All the dendrograms shown in

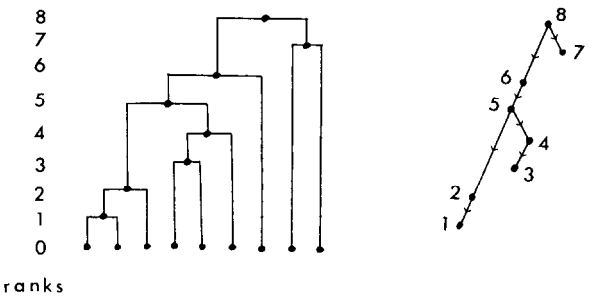

Fig. 3. Dendrogram ($n = 9$) and associated oriented binary tree.

Figs. 2 and 3 are in this form, which will be called standard form.

Using the unique permutation representation of NL-R dendrograms in standard form allows the enumeration of the latter by $d(n)$ as follows:

$$d(n+1) = e(n) = \tfrac{1}{2} \sum \left\{ \binom{n-1}{k} e(k)e(n-k-1) \colon k = 0, 1, \ldots, n-1 \right\},$$

$$e(0) = e(1) = 1.$$

This is proved as follows. Consider element n as fixed. Then choose k elements to the left of this element in $\binom{n-1}{k}$ ways, and arrange them in $e(k)$ ways. This leaves the remaining $(n-k+1)$ elements to be arranged to the right of the fixed element in $e(n-k+1)$ ways. When index k ranges over 0 to $n-1$ it can be seen that isomorphic dendrograms correspond to the cases where the elements to the left and right of the fixed element are interchanged (e.g. for $n = 5$: 13524 and 24513; or 12435 and 51243; etc.): hence the factor of $\frac{1}{2}$. Finally, enumerating permutations on $1, 2, \ldots, n$ – i.e. $e(n)$ – also enumerates dendrograms on $n+1$ terminal nodes – i.e. $d(n+1)$.

Such a recurrence relation has been used in the enumeration of alternating permutations (see [5]), i.e.

$$p(i) < p(i+1) > p(i+2) \qquad (1 \leq i \leq n-2)$$

(a down-up permutation if true when $i = 1$, otherwise an up-down permutation).

The number of complementary up-down or down-up permutations is counted by the André numbers (see [3]). A constructive proof of the following result is provided by the procedure described in [5]: There is a bijection between

- NL-R dendrograms in standard form (on n terminals),
- down-up permutations, and
- up-down permutations (both on $n-1$ elements).

7. Labelled, ranked, non-binary dendrograms

The previous sections have dealt with all major cases of binary dendrograms. We now turn attention to results for non-binary dendrograms.

The number of labelled, ranked, but non-binary dendrograms is given in [10], [11, p. 24], and [18, p. 259], as

$$f(n) = \sum \{S(n,k)f(k): k = 1, \ldots, n-1\},$$
$$f(1) = 1$$

where $S(n,k)$ is the Stirling number of the second kind. This recurrence relation is arrived at by noting that there are $S(n,k)$ ways to construct a partition with k classes; k ranges over 1 to $n-1$; and there are $f(k)$ ways to further construct partitions from the k classes formed, which can be considered as k objects. Examples of this type of dendrogram are given in [10], and a variant on the above formula is given in [14].

8. Labelled, non-ranked, non-binary dendrograms

The generalization of L-NR dendrograms to the non-binary case is given by [9, p. 17] in a form slightly different from the following:

$$g(n,k) = kg(n-1,k) + (n+k-2)g(n-1,k-1)$$

where k is the number of levels in the dendrogram on n terminal nodes, and ranges over 1 to $n-1$.

Generalizing the proof of

$$b(n) = (2n-3)b(n-1)$$

in the case of L-NR binary dendrograms (see Section 4 above) which the recurrence $g(n,k)$ reduces to when $k = n-1$, we can distinguish two cases: a dendrogram of k levels is constructed from a dendrogram of k levels by appending a new terminal node to an already existing node at any one of these k levels; or alternatively a new level is created. In the latter case, a new terminal node is appended just above one of the $n-1$ terminal nodes or the $k-1$ non-terminal nodes, leading to $(n+k-2)$ ways.

Reference [9] also discusses the enumeration of unlabelled, non-ranked dendrograms in the non-binary case, but does not give any simple formula. We have not found any simple results for the generalization of NL-R dendrograms to the non-binary case, either.

9. Quasi-labelled, non-ranked, binary dendrograms

Finally it is of interest to note that if the set of objects on which the dendrogram is built is given a prescribed order, then the number of dendrograms is given by the Catalan numbers. The second last level of the dendrogram breaks the ordered set into two (ordered) subsets of k and $n-k$ objects, where k can range over 1 to $n-1$.

Each of the ordered subsets can in turn be broken. This gives the recurrence relation [11, p. 27]:

$$h(n) = \sum \{h(k)h(n-k): k = 1, \ldots, n-1\}$$
$$h(1) = 1.$$

The last-mentioned reference also gives the result

$$h(n) = \frac{1}{n}\binom{2n-2}{n-1} = \frac{(2n-2)!}{n!(n-1)!}$$

which is proved using the ordinary generating function, giving the equation

$$\phi(x) = -x + \phi^2(x).$$

The binomial expansion of the square root in the solution leads to the desired term as the coefficient of x^n.

10. Conclusion

The numbers of non-isomorphic dendrograms for each of the definitions of dendrogram considered are tabulated for small n in Tables 1 and 2. The combinatorial study of dendrograms, apart from its inherent interest, can be of use in stochastic classification. An early paper in this area, [17], proposed a model of random dendrograms, and this was pursued in [8]. Among other enumeration problems relating to dendrograms and which have not been dealt with here, mention may be made of enumerating non-ranked dendrograms by height (see [8]); counting

Table 1
Numbers of non-isomorphic dendrograms for four types of binary dendrogram

n	L-R $a(n)$	L-NR $b(n)$	NL-NR $c(n)$	NL-R $d(n)$
1	1	1	1	1
2	1	1	1	1
3	3	3	1	1
4	18	15	2	2
5	180	105	3	5
6	2700	945	6	16
7	56700	10395	11	61
8	1587600	135135	23	272
9	57153600	2027025	46	1385
10	2571912000	34459425	98	7936

Notes: n = number of terminal nodes, L = labelled, NL = unlabelled, R = ranked, NR = non-ranked.

Table 2
Numbers of two types of non-binary dendrograms, and a particular binary dendrogram

n	$f(n)$	$g(n)$	$h(n)$
1	1	1	1
2	1	1	1
3	4	4	2
4	32	26	5
5	436	236	14
6	9012	2752	42
7	262760	39208	132
8	10270696	660032	429

Notes: $f(n)$ = labelled, ranked, non-binary; $g(n)$ = labelled, non-ranked, non-binary; $h(n)$ = quasi-labelled, non-ranked, binary.

the nodes in non-binary trees; and determining asymptotic results for numbers of dendrograms (see [8] and [10] for some results in this area).

References

[1] M.R. Anderberg, Cluster Analysis for Applications (Academic Press, New York, 1973).
[2] J.-P. Benzécri et coll., L'Analyse des Données, Tome I, La Taxinomie (Dunod, Paris, 1976).
[3] L. Comtet, Advanced Combinatorics (D. Reidel, Dordrecht, 1974).
[4] R.M. Cormack, A review of classification, J. Royal Statist. Soc. (A) 134 (1971) 321–367.
[5] R. Donaghey, Alternating permutations and binary increasing trees, J. Combin. Theory (A) 18 (1975) 141–148.
[6] D. Foata and M.P. Schützenberger, Nombres d'Euler et permutations alternantes, in: J.N. Srivastrava et al., ed., A Survey of Combinatorial Theory (North-Holland, Amsterdam, 1973) 173–187.
[7] O. Frank and K. Svensson, On probability distributions of single-linkage dendrograms, J. Statist. Comput. Simul. 12 (1981) 121–131.
[8] E.F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. Appl. Prob. 3 (1971) 44–77.
[9] B. Leclerc, Description, Evaluation et Comparaison des Hiérarchies de Parties (Centre d'Analyse et de Mathématique Sociale, Paris 6, 1982).
[10] T. Lengyel, On the numbers of all agglomerative clustering hierarchies, in: COMPSTAT 1982, Part II (Physica-Verlag, Wien, 1982) 177–178.
[11] I.C. Lerman, Classification Automatique et Analyse Ordinale des Données (Dunod, Paris, 1981).
[12] F.J. Rohlf, Numbering binary trees with labelled terminal vertices, IBM Research Report RC 8942 (1981).
[13] G. Saporta, Théories et Méthodes de la Statistique (Editions Technip, Paris, 1978).
[14] M. Schader, Hierarchical analysis: classification with ordinal object dissimilarities, Metrika 27 (1980) 127–132.
[15] R. Sibson, SLINK: an optimally efficient algorithm for the single-link cluster method, Comput. J. 16 (1973) 30–34.
[16] R. Sibson, Order invariant methods for data analysis, J. Royal Statist. Soc. (B) 34 (1972) 311–349.
[17] P.H.A. Sneath, Some statistical problems in numerical taxonomy, The Statistician 17 (1967) 1–12.
[18] M. Volle, Analyse des Données (Economica, Paris, 1981).