Ordered Dendrogram

Dimitar L.Vandev[*], Yanka G.Tsvetanova [†]

Abstract

The standard output of the hierarhical cluster methods is a tree-like diagram (dendrogram) . The idea of the presented algorithm is to rearrange dendrograms obtained by the usual clustering methods in order to improve the visual representation of the relationships between observed objects.

Each cluster is considered as an ordered sequence of objects (chain) with its left and right ends. At each step the two nearest chains are connected in their nearest ends.

There are presented some examples and obtained dendrograms are compared.

1 Introduction

There exist two well-known methods in multivariate data analysis for presenting an observed set of objects in a low dimentional space. These are multidimentional scaling and cluster analysis. Both types of techniques can be formalized as presenting distance-like data in a particular metric space by minimizing some kind of criteria.

Cluster analysis attempts to group the objects of an observed set, on the basis of similarity or distance between them, into mutually exclusive subsets (clusters) which consist of close objects. This clusters may be grouped into larger sets and so on, until all objects are eventually united in one cluster. The higher the level of aggregation is, the less similar are the objects in the respective cluster. These methods for cluster analysis are called hierarchical. The result of hierarchical classification can be represented graphically by a dendrogram.

The aim of Multidimentional scaling (MDS) is to arrange the investigated objects on a line or on a plane, or in a space of higher dimention, so that their mutual location to reflect , as far as possible, the degree of likeness or unlikeness between them. MDS is an alternative method of cluster analysis in the sence that from the resultant final configuration of points in two- or thre-dimentional space one could receive information about the structure of corresponding set of objects .

Having in mind mentioned above common points in these two methods we have an idea to rearrange the points (nodes) in the dendrogram obtained by usual clustering methods in such a way that the more similar the objects are , the closer together their corresponding nodes in the dendrogram will be and vice versa.

[1]Postal address: Laboratory of Computer Stochastics, Institute of Mathematics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria. Supported by National Fondation of Scientific Investigations, grant I-19/91.

[2]Postal address: Dep.of Math.and Phys., University of Zootechnics and Vet.Medicine, 6000 Stara Zagora, Stud.Grad, Bulgaria.

2 Notations

Let X is a set of n objects $\{o_1, o_2, \ldots, o_n\}$. and $\delta(.,.)$ is a positive real-valued symmetric function on XxX, whose values reflect the relative closeness (similarity) or distance (dissimilarity) of objects to one another. Obviously, the smaller the similarity value or larger the dissimilarity one is, the further apart the associated objects are. Frequently it is more natural to consider dissimilarities. Of course, we can always convert a similarity matrix to a dissimilarity one or vice versa by some nonincreasing transformation. For convenience we will use the dissimilarity coefficients $\delta_{ij} = \delta(o_i, o_j)$, $o_i, o_j \in X$. S denotes the set of all subsets of X and consists of 2^n elements.

Definition A subset H of S is a *hierarchy* on X if it satisfies the following conditions:

1 $X \in H$

2 $\{o_i\} \in H, 1 \leq i \leq n$

3 $C_1 \bigcap C_2 = \{ C_1, C_2, \oslash \}, \forall\, C_1, C_2 \in H$

The aim of the hierarchical clustering techniques is to construct a partition hierarchy by use of dissimilarity coefficients , satisfying the properties (1)-(3). The weakest partitioning contains n classes (clusters) and each one of them consists of one object.The strongest clustering contains all objects united into a single class.

Associated with any cluster C_k, formed at a given step of the hierarchical process is a real number Δ_k - *clustering level* and $\Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_{n-1}$. Usually Δ_k is equal to the dissimilarity value between the two clusters joined at that level.

There is a wide variety of methods for hierarchical cluster analysis. Most popular of them are the Single Linkage, Complete Linkage, Group Average, Median, Centroid, Ward's methods [1, 2, 3, 4, 5].
Different clustering methods imply different definitions of the dissimilarities (distances) between clusters and it is reasonable to give different result for the same data sets. In the most of these methods the arrangement of the objects in the clusters is arbitrary.

This hierarchical classification can be represented graphically by tree-like diagram called *dendrogram*. The observed objects in it are terminal nodes of the tree. The sequence of joining the clusters is visualized by fusion of two or more nodes into a parent node and so on until all nodes (clusters) are united into a single node at the top of the diagram. Usually a scale is incorporated into the dendrogram to indicate the dissimilarity level (aggregation distance) at which the two nearest clusters are supposed to join.

The *cardinality* of a cluster denotes the number of the objects in it.

Some of the mentioned above hierarchical methods are included as procedures in the well - known statistical packages BMDP, STATISTICA and others. These programs produce dendrograms in which a way of joining the clusters at a given level has not strict defined sence. We have an idea to rearrange such dendrograms in order to improve the visual representation of the hierarchical structure and to make them more realistic.

3 Ordering dendrograms

The dendrogram has 2^{n-1} binaty degrees of freedom. In Fig. 1 below there are four dendrograms presenting one and the same hierarchical classification. Eeach one of them may be obtained from another by rotating some vertical branches. They are indistinguishable with regard to nesting and amalgamation levels , and differ only in arrangement of the objects (terminal nodes).

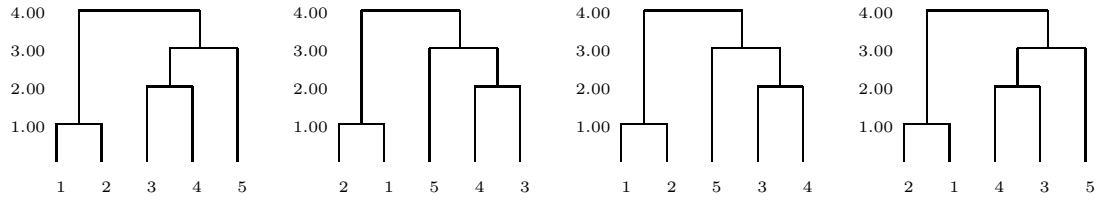

Fig.1

Our idea is to obtain an unique dendrogram using a specific way for linking the clusters , which to be "optimal" in a sence that the arrangement of nodes reflects the similarity of their respective clusters as well as it is possible.
The standard methods for hierarchical classification use as input data the dissimilarity coefficients (distances) δ between all pairs of objects. Distance δ can be defined in various ways and they satisfy or not metric properties, for example Euclidean, Manhattan (city-block), Chebychev, power distances, percent disagreement, correlation coefficients (Pearson r).

Other important point of the clustering methods is the determination of between cluster distance d. At the first step, when each object is a single cluster the distances between those clusters are defined by the chosen distance measure δ. When clusters contain more than one object it is necessary to define the distance d between two clusters. There are various ways to determine d as:

a) the distance between the two closest objects (nearest neighbours) in the different clusters (Single Link method).

b) the distance between the two furthest away objects accross the clusters (Complete Link method).

c) the average distance between all pairs of objects in the different clusters (Group Average method).

d) the distance between medians in the different clusters (Median method).

There exist also other methods , everyone of them uses a specific definition of between cluster distance. In all standard methods clusters are considered as sets of objects and to join together two clusters means to unite two sets. In this sence the obtained cluster is unique as a union of two sets. In our algorithm we consider every cluster as a partially ordered set in a sence of the relative position of the objects each to another (within the cluster) and in this case the union of two clusters is not uniquely determined.

In order to support the ordering of the objects every cluster is defined as the chain of objects.

Chain is an ordered sequence $\{o_1, o_2, \ldots, o_r\}$, $o_i \in X, 1 \leq i \leq r \leq n$ of objects .

At each step of the hierarchical clustering process there exists a set of chains, candidates for connecting and every one of the observed objects of X is included in one of the existing chains at this level. The pair of the currently nearest chains are connected.

Let C_1 and C_2 are two chains with their respective left and right ends o_{l1}, o_{r1} and o_{l2}, o_{r2}. There are four possibilities for linking these chains without changing the mutual arrangement of the objects in them. We introduce the following rules for connecting of two chains:

Rule 1 If $\delta_{ij} = min(\delta_{l1l2}, \delta_{l1r2}, \delta_{r1l2}, \delta_{r1r2})$, $(o_i \in \{o_{l1}, o_{r1}\}, o_j \in \{o_{l2}, o_{r2})$ the chains C_1 and C_2 are linked in such a way that their ends o_i and o_j to be neighbour (successive) elements in the new chain.

Rule 2 If $\delta_{ij} = max(\delta_{l1l2}, \delta_{l1r2}, \delta_{r1l2}, \delta_{r1r2})$, $(o_i \in \{o_{l1}, o_{r1}\}, o_j \in \{o_{l2}, o_{r2})$ the chains C_1 and C_2 are linked in such a way that their ends o_i and o_j become opposite ends of the new chain.

In the present paper we use the first rule , the second rule is applied by analogy. The two joined clusters are called subclusters and their union cluster - parent. It is obvious that the rules 1 and 2 are independent of the distances between objects δ and between clusters d.

As a result of applying one of the above rules we obtain an *ordered dendrogram* which has the property that the neighbour terminal nodes (objects) and nonterminal nodes (clusters) in it are really similar and vice versa.

In our original algorithm the distance between two clusters is defined as a distance between the nearest ends of the different chains. It uses Rule 1 for linking two chains. This algorithm uses only between object distances δ and can be applied successfully also to a dissimilarity matrix with missing data. Its output is a unique dendrogram. The arrangement of the terminal nodes has some interesting mathematical and topological properties. This algorithm works fast and effectively and is described in another paper which will be published. In this paper we present only a part of this algorithm , which is used for rearranging of dendrograms obtained by any clustering method.

4 Algorithm

As input data for ordering algorithm we use :

> $n(n-1)/2$ dissimilarity coefficients $\delta_{ij}, i \leq j, 1 \leq i, j \leq n(\delta_{ij} = \delta_{ji})$ of the upper half portion of the dissimilarity matrix;
>
> $n-1$ values of clustering levels $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$;
>
> set P of $n-1$ pairs of object numbers as each one of them belongs to one of the pair of subclusters which have been joined at the corresponding levels.

The values of $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$ are used for the vertical scale of the ordering dendrogram and δ_{ij} are used in finding the nearest ends of the two different chains which have to be united at a given step.

To make the presented algorithm clear we want to describe in details the data structure used to generate the ordered dendrogram.
A binary tree data structure is developed which is used in building the dendrogram and to represent the topological structure of the set of objects. This tree structure well fits to the input data set so that nearest objects(clusters) tend to be on nearby nodes of the tree. An array $TREE$ of $2n-1$ elements is used to store the information for all nodes of the tree. The n objects are assigned to the elements with indices from 1 to n. The elements with indices from $n+1$ to $2n-1$ are assigned to the sequential generated clusters $C_1, C_2, ..., C_{n-1}$. Each element of the array corresponds to a node of the tree. Merging of two nearest clusters at a given step of the iterational process means sticking the nearest ends of their corresponding chains. To support all relevant information every element of the array $TREE$ is a record, consisting of the following information:

node	=record	
	right_end,left_end	(pointers to the ends of the chain)
	left_sub,right_sub	(pointers to the subclusters C_{ls}, C_{rs} which are merged at given step(level))
	parent	(pointer to the upper cluster,it is important for the graphic representation)
	n1,n2	(pointers to the nearest ends of the chains C_{ls}, C_{lr})
	lavel	(the clustering level of this node, its value is equal to the distance d between C_{ls} and C_{rs}

end;

The algorithm can be described by the following paradigm:

START

$Initialize(TREE)$

$for\ i := 1\ to\ n-1\ do$
begin

$P1 := P[i,1]$; (*the number of the object of the first cluster $C(P1)$*)

$P2 := P[i,2]$; (*the number of the object of the second cluster $C(P2)$*)

$Search\_ends(P1, P1L, P1R)$; (*search the left and rihgt ends

$Search\_ends(P2, P2L, P2R)$; of the chains cotaining $P1$ and $P2$*)

$Nearest\_ends(P1L, P1R, P2L, P2R, F, S)$; (*search the nearest F and S*)

if (F and S are both left or right ends of their corresponding chains) $then\ Reverse$;

(*when it is necessary to reverse one of the chains this change reflects on the corresponding branch of the tree in the array $TREE$*)

(*Fulfill the record $n+i$ of currently generated node*)

$with\ TREE[n+i]\ do$

begin

$left\_end := OF; right\_end := OS$; (*opposite ends of F and S*)

$left\_sub := C(P1); right\_sub := C(P2)$;

$parent := n + i;$

$n_1 := F; n_2 := S;$

$distance := \Delta_i;$

end;

end;

STOP

After building the hierarchical tree structure printing of the dendrogram is easy because all relevant data for its graphical representation are available in the elements of the array $TREE$ described above.

5 Examples

Here we demonstrate some examples and compare usual dendrograms with ordered ones. In the first example data are derived from BMDP 81 reference manual ([6]) (Harbison et al.(1970),Appendix 1,p.8). The data are health indicators for 11 countries:

The health indicators measured are the relative number of doctors and dentists, of pharmacists, of nurses and hospital beds, the percent of animal fat and starch in the diet, and life expectancy .

The distance measure between cases is the Euclidean distance using standartized data and the Single linkage method is used . Table 1 contains the upper half of the distance matrix:

A dissimilarity matrix for 11 objects										
	2	3	4	5	6	7	8	9	10	11
1	3.97	3.84	3.82	6.83	4.07	3.52	4.40	2.99	4.30	4.75
2		1.39	1.57	5.33	3.49	2.07	2.45	2.59	2.32	2.83
3			1.21	5.10	3.11	1.73	3.01	1.85	2.78	3.54
4				5.08	3.31	2.00	3.30	2.19	3.29	3.88
5					4.44	6.48	7.10	5.82	6.09	5.33
6						3.36	4.65	2.77	3.51	3.59
7							2.37	1.50	2.50	3.82
8								2.91	1.62	2.90
9									2.68	3.62
10										1.88

Table 1

The tree diagrams of the clusters according to the classical and proposed algorithms are given respectively in Fig.2 and Fig.3. The case numbers are printed below the diagram. Each horisontal line segment in the tree corresponds to a cluster formed in the hierarchical clustering process. The vertical axis provides a scale by which to measure the dissimilarity between two merged clusters (amalgamation distances), each value of it is equal to the dissimilarity coefficient between the closest objects, one from each of the two joined subclusters, e.g. $\delta_{37} = 1.726$ is the amalgamation distance at which the clusters $\{2, 3, 4\}$ and $\{7, 9\}$ are linked together.

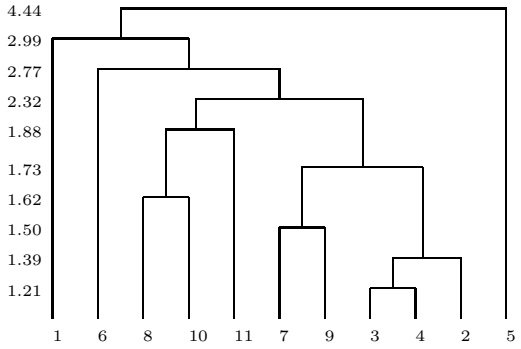

Fig.2

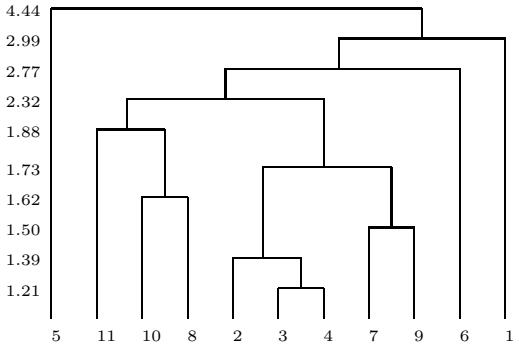

Fig.3

Let us compare the two dendrograms above. The first is obtained by use of simple algorithm (implemented in BMDP) without taking into account the possibility of ordering the objects. The second is the result of the rearranging algorithm.
Although the clustering structure in both dendrograms is identical it is easy to see from Table 1 that the order of the objects in Fig.3 is "better" than in Fig.2. For instance at the second level $\Delta_2 = 1.39$ two chains $\{3,4\}$ and $\{2\}$ are joined together. As $\delta_{23} < \delta_{24}$ it should be better to connect them in this way $\{\{2\},\{3,4\}\}$, thus the new chain to be $\{2,3,4\}$ as in Fig. 3 . Then the chains $\{2,3,4\}$ and $\{7,9\}$ are united at level $\Delta_3 = 1.73$. Comparing the distances between their ends it is appropriately to connect these chains in such a way that o_4 and o_7 to be neighbours ($\delta_{47} = min(\delta_{27}, \delta_{47}, \delta_{29}, \delta_{49})$). The chains $\{8,10,11\{$ and $\{2,3,4,7,9,\}$ are amalgamated at level $\Delta_7 = 2.32$ and as o_8 and o_2 are the closest ends the new chain is $\{11,10,8,2,3,4,7,9\}$. The object o_6 is more similar to o_9 than o_{11}. By analogy the objects o_1 and o_5 are arranged in Fig 3. It is clear that the ordered dendrogram in fig 3. gives more accurate visual presentation of the relationships between the observed objects.

The second example is from STATISTICA reference manual ([7]). For a set of different automobiles the following data are recorded: the approximate price of the car, the acceleration of the car, the braking performance of the car, an index of road holding capability, and the gass-mileage of the car. The data are standardized and the dissimilarity matrix is computed, as for the dissimilarity measure is used the Euclidean distance. As it turns out for this data, the Single linkage method produces rather "stringy" and undistinguished clusters. So we have chosen the Complete linkage method .

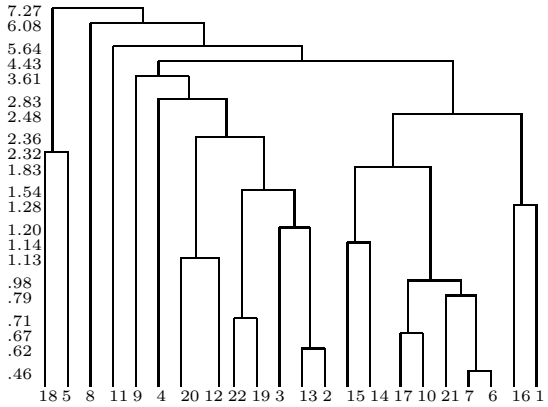

Fig.4

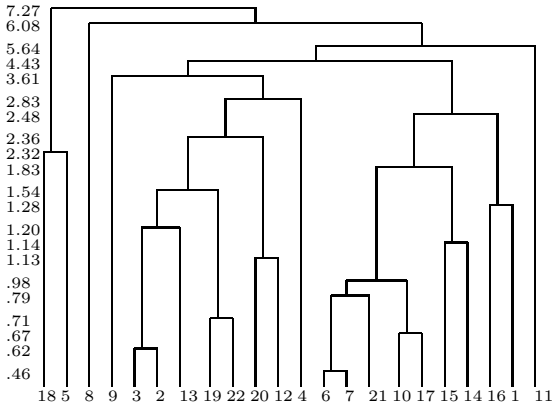

Fig.5

The Fig.4 and 5 above illustrate the hierarchical clustering scheme obtained respectively by STATISTICA package and the rearranging algorithm.

6 Discussion

The ordering algorithm produces an arrangement of the objects on a line which has undoubtedly positive merits but it is not perfect. In the final configuration of terminal nodes there exist some discordances between actual dissimilarities between objects and the mutual location of the corresponding nodes . For example , in Fig 3. o_7 is closer to o_4 than to o_3, but $\delta_{37} < \delta_{47}$. The object o_1 actually is closer to o_9 than to o_6. The reason for these "failures" is, that when two chains are linked the arrangement of objects within them is unchanged and only the distances between their ends are taken into consideration . In the clusters of higher level such discordances are more than in small clusters. The relative number of such discordances depend on the conrete data and may be estimated by our algorithm. Various ordered dendrograms can be obtained from the same dissimilarity matrix by applying different methods of cluster analysis and linking rules and to be compared.

The proposed algorithm may be improved further by generalizing the linking rules 1 and 2. As it was mentioned each nonterminal node (cluster) is an amalgamation of two clusters of lower level (subclusters). In this sence we define generalized rules for linking two clusters .

Let C_{1l} and C_{1r} , C_{2l} and C_{2r} are respectively left and right subclusters of two clusters C_1 and C_2, i.e. $C_1 = C_{1l} \bigcup C_{1r}, C_2 = C_{2l} \bigcup C_{2l}$. The new cluster (chain) $C = C_1 \bigcup C_2$ we may obtain by linking C_1 and C_2 in four ways : $\{C_{1l}, C_{1r}, C_{2l}, C2r\}$, $\{C_{1l}, C_{1r}, C_{2r}, C2l\}$, $\{C_{1r}, C_{1l}, C_{2r}, C2l\}$, $\{C_{1r}, C_{1l}, C_{2l}, C2r\}$. We introduce the following rules:

Rule 1' If $d_{1i2j} = min(d_{1l2l}, d_{1l2r}, d_{1r2l}, d_{1r2r})$, $(C_{1i} \in \{C_{1l}, C_{1r}\}, C_{2j} \in \{C_{2l}, C_{2r})$ the chains C_1 and C_2 are linked together in such a way that the nearest subclusters C_{1i} and C_{2j} of the different clusters C_1 and C_2 to be neighbours in the new chain.

Rule 2' If $d_{1i2j} = max(d_{1l2l}, d_{1l2r}, d_{1r2l}, d_{1r2r})$, $(C_{1i} \in \{C_{1l}, C_{1r}\}, C_{2j} \in \{C_{2l}, C_{2r})$ the chains C_1 and C_2 are linked together in such a way that the furthest subclusters C_{1i} and C_{2j} of the different clusters C_1 and C_2 to become opposite in the new chain.

It is clear that for applying these rules it is necessary to store information not only for between object distances δ but for the distances between the clusters of the current level and the lower level.

References

[1] Jambu Mishel, *Classification automatique pour l'analyse des donnees,* Bordas, Paris, 1978.

[2] Hand D. J., *Discrimination and Classification*, Wiley and Sons, 1992

[3] Hartigan J. A., *Clustering Algorithms*, John Wiley and Sons, New York, 1975

[4] Sneath P. H. A., Sokal R. R., *Numerical Taxonomy*, W. H. Freeman, San Francisco,1973

[5] Ward J. H. Jr., Hierarchical grouping to optimize an obective function, *J. Amer. Statist. Assoc.*, **58**, 236-244

[6] W. J.Dixon, M. B. Brown, L. Engelman, J. W. Frane, M. A. Hill, R. I. Jennrich, J. D. Toporek, *BMDP Statistical Software 1981,* University of California Press, Berkeley, 1981, p.456

[7] STATISTICA for Windows, Vol.1, 3, General conventions, StatSoft, Inc., 1994