統計與機械學習研討會

Clustering Ⅰ（陳君厚）
Clustering Ⅱ（吳漢銘）
中央研究院統計科學研究所

國立東華大學應用數學系

February 6, 2004

叢聚分析及相關問題

An Introduction to Clustering (Unsupervised Learning) Algorithms

Outline

A. HEURISTIC METHOD
   1). Leader.
B. NONHIERARCHICAL CLUSTERING: (k-mean)
   1). Algorithm I.
   2). Algorithm II.
   3). Algorithm III.
C. HIERARCHICAL CLUSTERING
   1). Agglomerative hierarchical clustering algorithms:
      1. Centroid method.
      2. Nearest-neighbor or single-linkage method.
      3. Farthest-neighbor or complete-linkage method
      4. Average-linkage method
      5. Ward's method.
   2). Divisive hierarchical clustering algorithms:
      1. monothetic.
      2. polythetic.

Outline (continued)

D. Selection of Proximity Measure

E. OTHER CLUSTERING METHODS

(to be covered by Dr. 吳漢銘)

1). Self Organizing Map (SOM).

2). Block Clustering.

3). Gene Shaving.

4). .....

F. CLUSTERING for CATEGORICAL DATA

(not covered in this lecture)

G. EXAMPLES & DISCUSSIONS

Clustering Analysis--materials extracted from:

- Papers:
  - P. MacNaughton-Smith, W. T. William, M. B. Dale and L. G. Mockett (1964), Dissimilarity analysis, Nature 202, 1034-1035.
  - R. M. Cormack (1971), A review of Classification (with discussion), JRSS-A, 321-367.
  - B. S. Everitt (1979), Unresolved Problems in Cluster Analysis, Biometrics 35, 169-181.
  - G. Punj and D. W. Stewart (1983), Cluster Analysis in Marketing Research: Review and Suggestions for Application, J. of Marketing Research, 20, 134-148.
  - J. A. Hartigan (1985), Statistical Theory in Clustering, J. of Classification 2, 63-76.

Clustering Analysis--materials extracted from:

- Papers:
  - D. Gordon (1987), A review of Hierachical Classification, JRSS-A 150, 119-137.
  - Richard P. Lippmann (1987), An Introduction to Computing with Neural Nets, IEEE ASSP Magazine, April 1987, 4-22.
  - Mark J. Schervish (1987), A Review of Multivariate Analysis, Statistical Science, 2, 396-433.
  - Panel on Discriminant Analysis and Clustering (1989), Discriminant Analysis and Clustering, Statistical Science, 4, 34-69

Clustering Analysis--materials extracted from:

- Books:
  - Peter H. A. Sneath and Robert R. Sokal (1973), Numerical Taxonomy: The principles and practice of numerical classification. Freeman.
  - John A. Hartigan (1975), Clustering Algorithms, John Wiley.
  - Helmuth Spath (1980), Cluster Analysis Algorithms: for data reduction and classification of objects. Ellis Horwood.
  - A. Afifi and Virginia Clark (1984), Computer-Aided Multivariate Analysis, Van Nostrand Reinhold.
  - Charles H. Romesburg (1984), Cluster Analysis for Researchers, Lifetime Learning.
  - Leonard Kaufman and Peter J. Rousseeuw (1990), Finding groups in data: an introduction to cluster analysis, John Wiley.

Clustering Analysis--materials extracted from:

- Books:
  - Richard A. Johnson and Dean W. Wichern (1992), Applied Multivariate Statistical Analysis, 3rd ed. Prentice Hall.
  - Brian S. Everitt (1993), Cluster Analysis, 3rd ed., Edward Arnold.
  - Subhash Sharma (1996), Applied Multivariate Techniques, John Wiley.

Clustering in (Statistical) Multivariate Analysis:

- In Mark J. Schervish (1987), A Review of Multivariate Analysis:
  - 1. Decision Theory and Bayesian Inference.
  - 2. Discriminant Analysis. *(LDA, k-NN, ML, CART, Aggregating, SVM, ...)* *Data: NCI60, B-Cell lymphoma, ...*
  - 3. Exploratory Methods:
    - 3.1 Cluster Analysis.
    - 3.2 Multidimensional Scaling. *3.1 + 3.3 ~ TreeViewer (M. Eisen)* *3,1 + 3.2 ~ S.O.M. (E. Lander, T. Kohone*
    - 3.3 Graphical Methods.
  - 4. Regression.
  - 5. Canonical Correlation. *(2 data sets)*
  - 6. Principal Components. *Cell Cycle (P. Brown, N. V. Fedoroff, K. C. Li, ...)*

Clustering in (Statistical) Multivariate Analysis:

- 7. Factor Analysis.
  - 7.3 Exploratory Factor Analysis.
  - 7.4 Confirmatory Factor Analysis.
  - 7.5 Interpretations.
- 8. Path Analysis and LISREL.
  - 8.1 Path Analysis. *separate genetic from environmental effect (family study) for pathway?*
  - 8.2 Linear Structural Relations.
  - 8.3 Interpretations .
- 9. Testing Hypotheses.
- 10. Discrete Multivariate Analysis *SNP (single nucleotide polymorphism)*
- 11. Design of Experiments. *+ ANOVA. chip/array (SWAP, LOOP, ...)*
- 12. MCMC, HMM.

Definition:

- A. A. Afifi and Virginia Clark:
  - Cluster analysis is a technique for grouping individuals or objects into unknown groups. It differs from other methods of classification, such as discriminant analysis, in that in cluster analysis the number and characteristics of the groups are to be derived from the data and are not usually known prior to the analysis.
- Leonard Kaufman and Peter J. Rousseeuw:
  - Cluster analysis is the art of finding groups in data.
- John A. Hartigan:
  - Clustering is the grouping of similar objects.

Definition:

- Charles H. Romesburg:
  - Cluster analysis is a generic name for a variety of mathematical methods, numbering in the hundreds, that can be used to find out which objects in a set are similar.

- Peter H. A. Sneath and Robert R. Sokal:
  - The relative vagueness of our definition of pattern makes it even more difficult to try and define cluster. We shall encounter a variety of criteria for measuring properties of clusters, and this variety opens the way to a multiplicity of different definitions of the term. For this reason we shall leave the definition of clusters conveniently vague: sets of OUT's in phenetic hyperspace that exhibit neither random nor regular distribution patterns and that meet one or more of various criteria imposed by a particular cluster definition.

Definition:

- Dianne Cook:
  - Cluster Analysis attempts to group data points into homogeneous groups. It is assumed that you don't know the groups *a priori* , so the first step is to examine the proximity of the points with respect to each other. Cluster Analysis can be considered to be an exploratory data analysis technique.

- Subhash Sharma:

  Cluster analysis is a technique used for combining observations into groups or clusters such that:
  - 1. Each group or cluster is homogeneous or compact with respect to certain characteristics. That is, observations in each group are similar to each other.
  - 2. Each group should be different from other groups with respect to the same characteristics; that is, observations of one group should be different from the observations of other groups.

Principles:

- Helmuth Spath:

  Basic problems in cluster analysis are:

  – 1. selection of distance *Euclidean (standardize), correlation, ...*

  – 2. selection of algorithm *k-mean, SOM, tree, ...*

  – 3. the number of clusters to be formed *???*

  – 4. the choice of variables, especially their scaling

  (feature selection, screening) *normalization*

Principles:

- Charles H. Romesburg:

  Its skeleton comprises six steps:
  - A. Obtain the data matrix. *** *Lab process 1. 2. 3…, Image process 1. 2. 3..*
  - B. Standardize the data matrix. *normalization*
  - C. Compute the resemblance matrix.
  - D. Execute the clustering method.
  - E. Rearrange the data and resemblance matrices. *Tree, SOM*
  - F. Compute the cophenetic correlation coefficient.

    goodness of a tree

Principles:

- Peter H. A. Sneath and Robert R. Sokal:

Eight aspects of clustering methods:

- Agglomerative versus Divisive Methods. *n -->1, 1--> n*
- Hierarchic versus Nonhierarchic Methods. *tree <--> k-mean*
- Nonoverlapping versus Overlapping Methods. *Probability <--> possibili*
- Sequential versus Simultaneous Methods. *K-mean <--> tree*
- Local versus Global Criteria. *tree <--> GAP*
- Direct versus Iterative Solutions. *tree <--> k-mean, SOM*
- Weighted versus Unweighted Clustering.
- Nonadaptive versus Adaptive Clustering.

Fuzzy

Gene Shavin

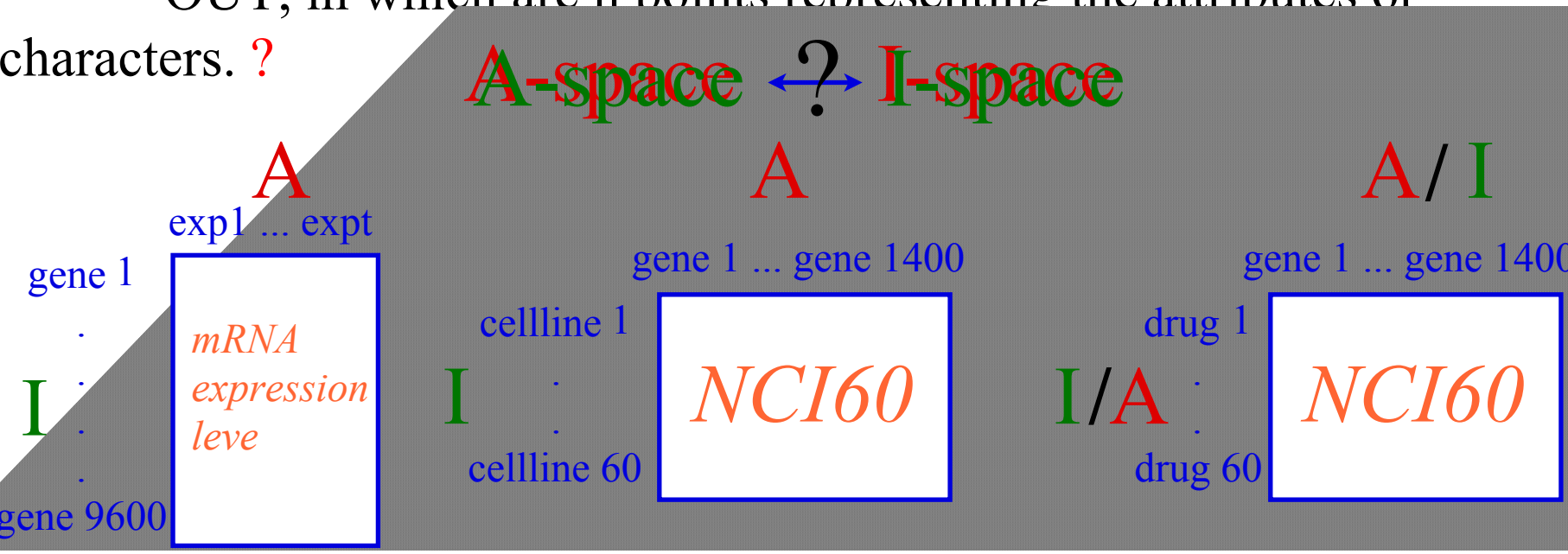
•OUT: operational taxonomic units.

•Pattern *(temporal exp. level, profile, ...)*

–(We shall mean by pattern, therefore, any describable properties of the distribution of OUT's and groups of OUT's in an A-space.)

•A-space (attribute space) has formally n dimensions, one for each attribute or character, in which there are t points that represent the OUT's. ?

•I-space (individual space) has formally t dimensions, one for each OUT, in which are n points representing the attributes or characters. ?



Applications:

- Taxonomy: Clustering species of bees into higher-level taxonomic groups
- Genetics: Studying genetic diversity within and between populations
- Medicine: Developing clusters of patients based on physiological variables
- Speech processing: Constructing a speaker-independent word recognition sys
- Glaciology: Mapping the Antarctic and Arctic regions in terms of clusters of types of sea ice and fern.
- Archaeology: Grouping broaches from an Iron Age site in Switzerland based on their attributes
- Education: Dividing up a class of workers in the telephone industry based on their common training
- Business: Clustering corporations according to their financial characteristics.
- Bioinformatics: Gene/protein expression, phenotype/genotype grouping, ….

Heuristic Methods:

- Helmuth Spath:
  - There are an arbitrarily large number of heuristic procedures (see Anderberg 1973). All have in common the way in which they are oriented towards visual, geometric representation of cluster formations of point objects distributed in a plane. This is very reasonable, since visual classification of points in a plane cannot be improved upon by algorithms. In fact cluster algorithms are superior to visual classification only for more than two dimensions.

Heuristic Methods: LEADER

- The LEADER program considers each object only once and immediately allocate it to a cluster. For a specified neighbor size ρ and maximum number of clusters NMAX to be formed:
  - A. i=1, k=1.
  - B. Case(i) is allocated to the k-th cluster.
  - C. i=i+1, stop if i > n
  - D. Assign case(i) to the first of the previously generated clusters for which the first (leading) element is at a distance less than ρ .
  - E. If there is no such cluster and k < NMAX, set k=k+1 and return to step 2, otherwise leave case(i) unallocated and return to step 3.
- ** If ρ is too small that NMAX clusters are insufficient to allocate all the points, the remaining points are not assigned → the result depends on the sequence in which the objects are taken.
- ** If ρ is too large, fewer than NMAX clusters are formed.

- From one cluster to one per cluster:

- Richard A. Johnson and Dean W. Wichern:

Possible Attributes:

suit:

rank:

color:

•••






- The number G(n, k) of possible ways of sorting n objects into k nonempty groups is a Stirling number of the second kind given by

$$\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{n}$$

- Adding these numbers for k=1,2,…, n groups, we obtain the total number of possible ways to sort n objects into groups.

(see Anderberg 1973, Jensen 1969, Luneburg 1971).

G(n, k)=kG(n-1,k) + G(n-1, k-1)

With G(1, 1)=1 and G(1, k) = 0 for n≠1.

For the case of two clusters: G(n, 2)=2^(n-1)-1

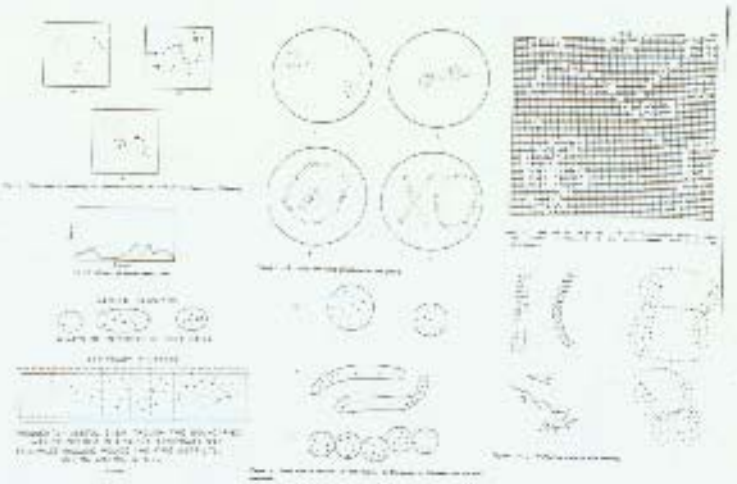
G(59, 2) ≈ 10^18

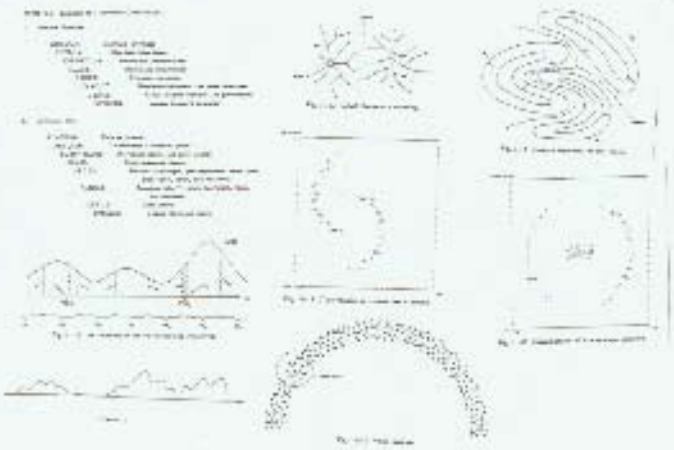
G(15, 3) = 2375101

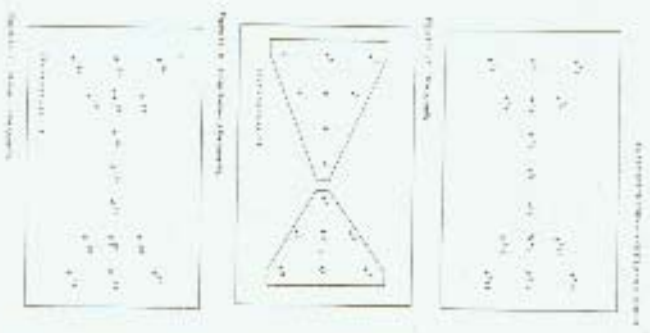
G(20, 4) = 45232115901

G(25, 8) = 69022372111836858O

G(100, 5) = 10^68







NONHIERARCHICAL CLUSTERING: Subhash Sharma (1996)

- In nonhierarchical clustering, the data are divided into k partitions or groups with each partition representing a cluster. Therefore, as opposed to hierarchical clustering, the number of clusters must be known a priori. Nonhierarchical clustering techniques basically follow these step:

  - 1. Select k initial cluster centroids or seeds, where k is the number of clusters desired.
  - 2. Assign each observation to the cluster to which it is the closest.
  - 3. Reassign or reallocate each observation to one of the k clusters according to a predetermined stopping rule.
  - 4. Stop if there is no reallocation of data points or if the reassignment satisfies the criteria set by the stopping rule. Otherwise go to Step 2.

NONHIERARCHICAL CLUSTERING:

- Most of the nonhierarchical algorithms differ with respect to: (1) the method used for obtaining initial cluster centroids or seeds; and (2) the rule used for reassigning observations. Some of the methods used to obtain initial seeds are

  - 1. Select the first k observations with nonmissing data as centroids or seeds for the initial clusters.
  - 2. Select the first nonmissing observation as the seed for the first cluster. The seed for second cluster is selected such that its distance from the previors seed is greater than a certain selected distance. The third seed is selected such that its distance from previously selected seeds is greater than the selected distance, and so on.
  - 3. Randomly select k nonmissing observations as cluster centers or seeds.

NONHIERARCHICAL CLUSTERING:

- 4. Refine the selected seeds using certain rules such that they are as far apart as possible. Some of these rules are discussed in Sections 7.7.2 and 7.8.
- 5. Use a heuristic that identifies cluster centers such that they are as far apart as possible.
- 6. Use seeds supplied by the researcher.

- Once the seeds are identified, initial clusters are formed by assigning each of the remaining n - k observations to the seed to which the observation is the closest.

it is possible to use non sample-point seeds

- Nonhierarchical algorithms also differ with respect to the procedure used for reassigning subjects to the k clusters. Some of the reassignment rules are

  – 1. Compute the centroid of each cluster and reassign subjects to the cluster whose centroid is the nearest. The Centroids are not updated while assigning each observation to the k clusters; they are recomputed after the assignment for all the observations have been made. If the change in the cluster centroids is greater than a selected convergence criterion then another pass at reassignment is made and cluster centroid are recomputed. The reassignment process is continued until the change in the centroids is less than the selected convergence criterion.

  – 2. Compute the centroid of each cluster and reassign subjects to the cluster whose centroid is the nearest. For the assignment of each observation, recompute the centroid of the cluster to which the observation is assigned and the cluster from which the observation is removed. Once again, reassignment is continued until the change in cluster centroid is less than the selected convergence criterion.

- 3. Reassign the observations such that some statistical criterion is minimized. These methods are commonly referred to as hill-climbing methods. Some of the objective functions or the statistical criteria that can be minimized are
  - (a) trace of the within-group SSCP $(= (X - \bar{X})'(X - \bar{X}))$ matrix (i.e., minimize ESS).
  - (b) determinant of the within-group SSCP matrix.
  - (c) trace of $W^{-1}B$, where W and B are, respectively, the within-group and between-group SSCP matrices.
  - (d) largest eigenvalue of the $W^{-1}B$ matrix.

SSPC: sum of squares and cross products

- As can be seen, a variety of clustering algorithms can be developed depending on the combination of the initial partitioning and the reassignment rule employed. Three popular types of nonhierarchical algorithms will be discussed and illustrated using the hypothetical data given in Table 7.1. For illustration purposes we will assume that three clusters are desired and that a convergence criterion of .02 has been specified.

Example Subhash Sharma (1996): Hypothetical Data

Subject	Income	Education
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

Algorithm I:

- 1. Selects the first k observations as the initial k centroids.
- 2. Calculate the (squared euclidean) distance of each observation from the initial k centroids.
- 3. Assign each observation to the cluster with minimum distance to that centroid
- 4. The next step is to compute the new centroid of each cluster.
- 5. For a given case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.
- 6. Repeat step 5 for each case.
- 7. Repeat steps 4, 5 and 6 until no cases are reassigned

Algorithm II:

- This algorithm differs from Algorithm I with respect to how the initial seeds are modified. The first three observations are selected as cluster seeds. Then each of the remaining observations is evaluated to determine if it can replace any of the previously selected seeds according to the following rule: The seed that is a candidate for replacement is from the two seeds (i.e., pair of seeds) that are closest to each other. An observation qualifies to replace one of the two identified seeds if the distance between the seeds is less than the distance between the observation and the nearest seed. If the observation quantifies, then the seed that is replaced is the one closest to the observation. This rule, and its variant, are used in the nonhierarchical clustering procedure in SAS.

Algorithm II:

- See example for detailed explanation:
- ** Nonhierarchical clustering techniques are quite sensitive to the selection of the initial seeds. Algorithms I and II are commonly referred to as K-means clustering.

Algorithm III:

- The nonhierarchical clustering programs differ with respect to initial partitioning and the reassignment rule. Here is an alternative heuristic for selecting the initial seeds and a reassignment rule that explicitly minimizes the ESS (ie., trace of the within-group SSCP matrix).
- Let Sum(i) be the sum of the values of the variables for each observation and k be the desired number of clusters. The initial allocation of observation *i* to cluster Ci is given by the integer part of the following equation:

$$C_i = \frac{(Sum(i) - Min)(k - 0.0001)}{Max - Min} + 1 \qquad Sum(i) = \sum_{j=1}^{p} x_{ij} = x_{i\bullet}$$

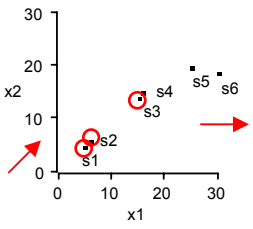
Algorithm III:

- Where Ci is the cluster to which observation I should be assigned, Max and Min are, respectively, the maximum and minimum of Sum(I), and k is the number of clusters desired.
- Next, the observations are reassigned such that the statistical criterion, ESS, is minimized. This change in ESS is computed for reassignment of the observation to each of the other clusters, and the observation is reassigned to the cluster that results in the greatest decrease in ESS. The procedure is repeated for all obs.

Subhash Sharma (1996):
Hypothetical Data

Obs.	x1	x2
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

first 3



PROXIMITY MATRIX WITH EUCLIDEAN DISTANCES

	S1	S2	S3	S4	S5	S6
S1	0	2	181	221	625	821
S2	2	0	145	181	557	745
S3	181	145	0	2	136	250
S4	221	181	2	0	106	212
S5	625	557	136	106	0	26
S6	821	745	250	212	26	0

Algorithm I:

Initial steps: 1, 2, 3

Initial Cluster Centroids:

Centroids

cluster	x1	x2
1	5	5
2	6	6
3	15	14

Distance from cluster centroids and initial assignment of observations

Obs.	Distance from cluster Centroid 1	2	3	assigned
S1	0	2	181	1
S2	2	0	145	2
S3	181	145	0	3
S4	221	181	2	3
S5	625	557	136	3
S6	821	745	250	3

Repeat steps: 4, 5, 6

Centroids of the 3 clusters and change in cluster centroid:

New Centroids

cluster	x1	x2
1	5	5
2	6	6
3	21.5	17.0

Centroids Change

	x1	x2
1	0	0
2	0	0
3	6.5	3.0

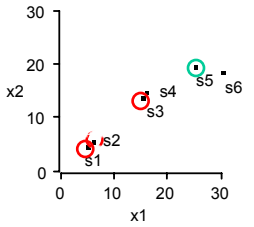
Distance from cluster centroids and 1st reassignment of observations

Obs.	Distance from cluster Centroid 1	2	3	Cluster Assignment prev	reasir
S1	0	2	416.25	1	1
S2	2	0	361.25	2	2
S3	181	145	51.25	3	3
S4	221	181	34.25	3	3
S5	625	557	21.25	3	3
S6	821	745	76.25	3	3

stop

Subhash Sharma (1996):
Hypothetical Data

Obs.	x1	x2
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



	S1	S2	S3	S4	S5	S6
S1	0	2	181	221	625	821
S2	2	0	145	181	557	745
S3	181	145	0	2	136	250
S4	221	181	2	0	106	212
S5	625	557	136	106	0	26
S6	821	745	250	212	26	0

Algorithm II:

Initial steps: 1, 2, 3

1. Initial seeds = (S1, S2, S3)
2. Smallest between seeds distance = min(d(seeds))=min(d12, d13, d23)=min(2, 181, 145)=2
3. S4 is not qualified since min(d(S4, seeds)) = min(d14, d24, d34)=min(221, 181, 2)= 2 /> 2 = min(d(seeds))
4. S5 quakified since min(d(S5, seeds)) = min(625, 557, 136)= 136 > 2
5. S5 replace S2 since d25 = 557 < 625 = d15 *** New seeds --> (S1, S3, S5)
6. Smallest between new seeds distance = min(d(seeds))=min(181, 625, 136) = 136
7. S6 is not qualified since min(d(S6, seeds)) = min(d16, d36, d56)=26 /> 136

Final seeds are S1, S3, S5

Initial assignment, Distance from cluster centroids

Obs.	Distance from cluster Centroid 1	2	3	assigned
S1	0	181	625	1
S2	2	145	557	1
S3	181	0	136	2
S4	221	2	106	2
S5	625	136	0	3
S6	821	250	26	3

Centroids

cluster	x1	x2
1	5.5	5.5
2	15.5	14.5
3	27.5	19.5

27.5 = mean(25, 30)

Repeat steps: 4, 5, 6

Distance from cluster centroids and 1st reassignment of observations

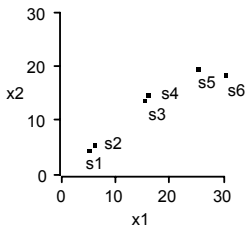
Obs.	Distance from cluster Centroid 1	2	3	Cluster Assignment prev	reasin
S1	0.5	200.5	716.5	1	1
S2	0.5	162.5	644.5	1	1
S3	162.5	0.5	186.5	2	2
S4	200.5	0.5	152.5	2	2
S5	590.5	120.5	6.5	3	3

=

stop

Subhash Sharma (1996):
Hypothetical Data

Obs.	x1	x2
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



	S1	S2	S3	S4	S5	S6
S1	0	2	181	221	625	821
S2	2	0	145	181	557	745
S3	181	145	0	2	136	250
S4	221	181	2	0	106	212
S5	625	557	136	106	0	26
S6	821	745	250	212	26	0

Algorithm III:

Initial steps: 1, 2, 3

Initial Assignment

Obs.	x1		x2		Sum(i)	Ci	Assigned to Cluster
S1	5	+	5	=	10	1	1
S2	6	+	6	=	12	1	1
S3	15	+	14	=	29	2	2
S4	16	+	15	=	31	2	2
S5	25	+	20	=	45	3	3
S6	30	+	19	=	49	3	3

Centroids

cluster	x1	x2
1	5.5	5.5
2	15.5	14.5
3	27.5	19.5

$$C_1 = \frac{(10-10)(3-0.0001)}{49-10}+1 = 0+1 = 1 \qquad C_3 = \frac{(29-10)(3-0.0001)}{49-10}+1 = \frac{56.9981}{39}+1 = 1.4615+1 = 2.4615$$

$$C_5 = \frac{(45-10)(3-0.0001)}{49-10}+1 = \frac{104.9965}{39}+1 = 2.6922+1 = 3.6922$$

$$C_i = \frac{(Sum(i)-Min)(k-0.0001)}{Max-Min}+1$$

Repeat steps: 4, 5, 6

The change in ESS if S1 belonging to cluster 1 is reassigned to cluster3:

$$\text{Change in ESS} = \frac{3}{2}[(5-27.5)^2+(5-19.5)^2]-\frac{1}{2}[(5-5.5)^2+(5-5.5)^2] = 1074.75-0.25 = 1074.5$$

weight(cluster3) weight(cluster1)

Change in ESS due to Reassignment

Obs.	Cluster	Change in ESS if Assigned to Cluster 3	2	1	Reassignment
S1	1	1074.50	300.50	-	1
S2	1	966.50	243.75	-	1
S3	2	279.50	-	243.50	2
S4	2	228.50	-	300.75	2
S5	3	-	177.50	882.50	3

Weight for cluster i =

#(obs) after reassign in cluster i

#(obs) before reassign in cluster i

HIERARCHICAL CLUSTERING: Subhash Sharma (1996)

- The hierarchical algorithms result in a tree-like dendrogram.
  - At the top of the tree each observation is represented as a separated "cluster".
  - At intermediate levels observations are grouped into fewer "cluster" than at the higher levels.
  - At the bottom, all of the observations are merged into one "cluster".
- The hierarchical clustering algorithm forms clusters in a hierarchical fashion. That is, the number of clusters at each stage is one less than the previous one. If there are n observations then at Step 1, Step2, …, Step n-1 of the hierarchical process the number of clusters, respectively, will be n-1, n-2, …, 1.
- Frequently the various steps or stages of the hierarchical clustering process are represented graphically in what is called a dendrogram or tree.

HIERARCHICAL CLUSTERING: Subhash Sharma (1996)

- In some problems, entire tree structure may be of interest.
  In others, tree is just a convenient tool for obtaining a partition.
  This is done by cutting the tree at a suitable level which forces a particular partition.
- Some hierarchical algorithms form the tree from the bottom up in a divisive fashion, but most work agglomeratively from the top down.

For M. Eisen TreeViewer, a tree is created to:

1. *obtain a partition: cluster of genes (according to expression profile)*
2. *obtain a permutation: "relative" positions of terminal nodes for sorting genes*
3. *Entire tree structure: not that important*

Agglomerative hierarchical clustering algorithms:

- A number of different rules or methods have been suggested for computing distances between two clusters. In fact, the various hierarchical clustering algorithms or methods differ mainly with respect to how the distances between the two clusters are computed. Some of the popular methods are:

  - 1. Centroid method. *TreeViewer (M. Eisen)*
  - 2. Nearest-neighbor or single-linkage method.
  - 3. Farthest-neighbor or complete-linkage method
  - 4. Average-linkage method *TreeViewer =Average Linkage (M. Eisen claimed)*
  - 5. Ward's method.

Single-Linkage or the Nearest-Neighbor Method

In the single-linkage method, the distance between two clusters is represented by the minimum of the distance between all possible pairs of subjects in the two clusters.

•Step1: Merge clusters 3 and 4, giving 1, 2, (34), and 5 at the value of e34=5.0.

Revise the distance matrix:

e1(34) = min(e13, e14) = min(20.6, 22.4) = 20.6;

e2(34) = min(e23, e24) = min(14.1, 11.2) = 11.2;

e5(34) = min(e35, e45) = min(25.0, 25.5) = 25.0

•Step2: Merge clusters 1 and 5, giving 2, (34), and (15) at the value of e15=7.07.

Revise the distance matrix:

e2(15) = min(e12, e25) = min(18.0, 18.0) = 18.0;

e(15)(34) = min(e13, e14, e35, e45) = min(20.6, 22.4, 25., 25.5) = 20.6;

•Step3: Merge clusters 2 and (34), giving (15) and (234) at the value of e2(34)=11.2.

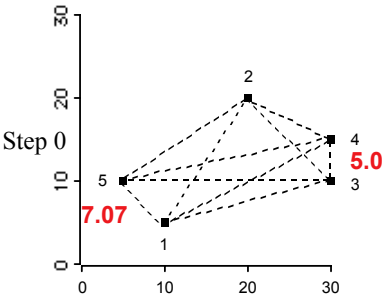
Revise the distance matrix:

e(15)(234) = min(e12, e13, e14, e25, e35, e45) = min(18.0, 20.6, 22.4, 25., 25.5) = 18.0;

Single Linkage

	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

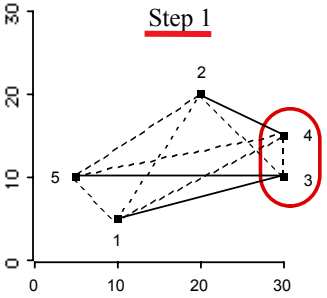
	1	2	3	4	5
1	0.00				
2	18.0	0.00			
3	20.6	14.1	0.00		
4	22.4	11.2	5.00	0.00	
5	7.07	18.0	25.0	25.5	0.00

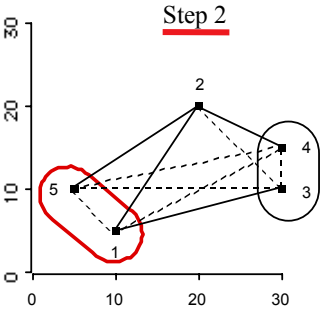


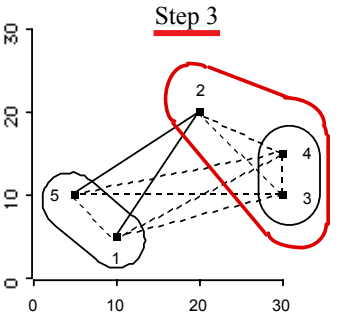
	1	2	5	(34)
1	0.00			
2	18.0	0.00		
5	7.07	18.0	0.00	
(34)	20.6	11.2	25.0	0.00

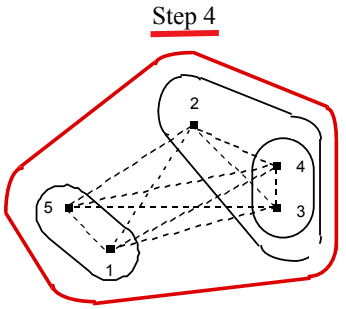
	2	(34)	(15)
2	0.00		
(34)	11.2	0.00	
(15)	18.0	20.6	0.00

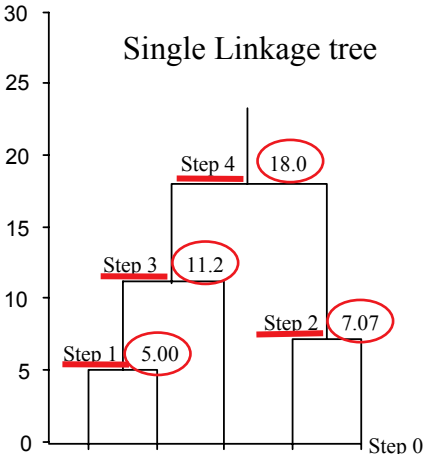
	(15)	(234)
(15)	0.00	
(234)	18.0	0.00











** solid lines show the minimum distances between clusters

Complete-Linkage or Farthest-Neighbor Method

The complete-linkage method is the exact opposite of the nearest-neighbor method. The distance between two clusters is defined as the maximum of the distances between all possible pairs of observations in the two clusters.

•Step1: Merge clusters 3 and 4, giving 1, 2, (34), and 5 at the value of e34=5.0.

Revise the distance matrix:

e1(34) = max(e13, e14) = max (20.6, 22.4) = 22.4;

e2(34) = max (e23, e24) = max (14.1, 11.2) = 14.1;

e5(34) = max (e35, e45) = max (25.0, 25.5) = 25.5.

•Step2: Merge clusters 1 and 5, giving 2, (34), and (15) at the value of e15=7.07.

Revise the distance matrix:

e2(15) = max(e12, e25) = max(18.0, 18.0) = 18.0;

e(15)(34) = max(e13, e14, e35, e45) = max(20.6, 22.4, 25., 25.5) = 25.5;

•Step3: Merge clusters 2 and (34), giving (15) and (234) at the value of e2(34)=14.1.

Revise the distance matrix:

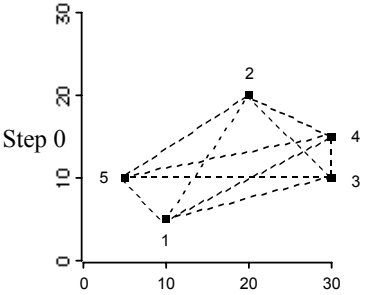
e(15)(234) = min(e12, e13, e14, e25, e35, e45) = min(18.0, 20.6, 22.4, 25., 25.5) = 18.0;

Complete Linkage

	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

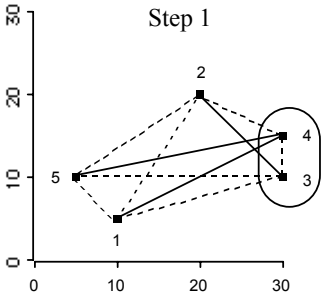
	1	2	3	4	5
1	0.00				
2	18.0	0.00			
3	20.6	14.1	0.00		
4	22.4	11.2	5.00	0.00	
5	7.07	18.0	25.0	25.5	0.00

Step 0



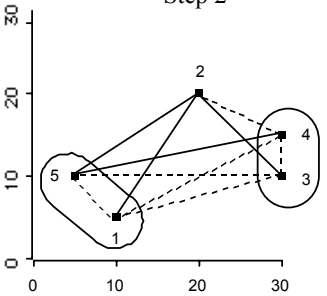
	1	2	5	(34)
1	0.00			
2	18.0	0.00		
5	7.07	18.0	0.00	
(34)	22.4	14.1	25.5	0.00

Step 1



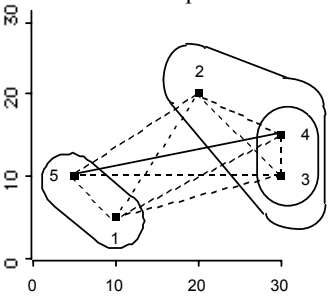
	2	(34)	(15)
2	0.00		
(34)	14.1	0.00	
(15)	18.0	25.5	0.00

Step 2

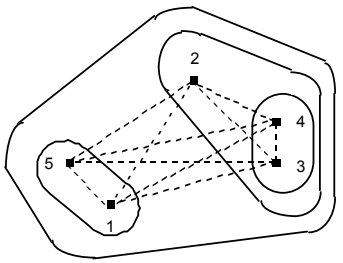


	(15)	(234)
(15)	0.00	
(234)	25.5	0.00

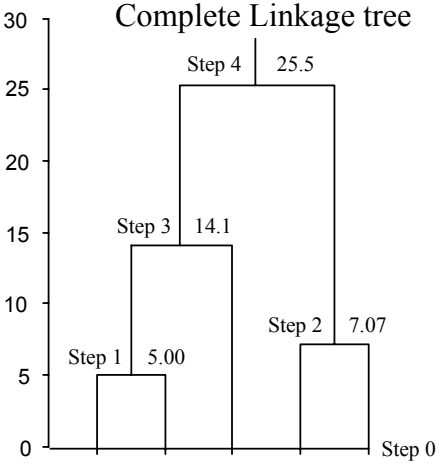
Step 3



Step 4



Complete Linkage tree



** solid lines show the maximum distances between clusters

Average-Linkage Method

In the average-linkage method the distance between two clusters is obtained by taking the average distance between all pairs of subjects in the two clusters.

•Step1: Merge clusters 3 and 4, giving 1, 2, (34), and 5 at the value of e34=5.0.

Revise the distance matrix:

e1(34) = (e13+e14)/2 = (20.6+22.4)/2 = 21.5;

e2(34) = (e23+e24)/2 = (14.1+11.2)/2 = 12.7;

e5(34) = (e35+e45)/2 = (25.0+25.5)/2 = 25.3.

•Step2: Merge clusters 1 and 5, giving 2, (34), and (15) at the value of e15=7.07.

Revise the distance matrix:

e2(15) = (e12+e25)/2 = (18.0, 18.0)/2 = 18.0;

e(15)(34) = (e13, e14, e35, e45)/4 = (20.6+22.4+25.+25.5)/4 = 23.4;

•Step3: Merge clusters 2 and (34), giving (15) and (234) at the value of e2(34)=12.7.

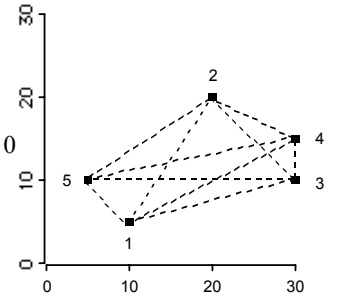
e(15)(234)=(e12,e13,e14,e25,e35,e45)/6= (18.0+20.6+22.4+18.0+25+25.5)/6 = 21.6;

Average Linkage (UPGMA)

	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

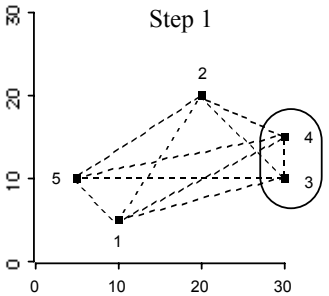
	1	2	3	4	5
1	0.00				
2	18.0	0.00			
3	20.6	14.1	0.00		
4	22.4	11.2	5.00	0.00	
5	7.07	18.0	25.0	25.5	0.00

Step 0



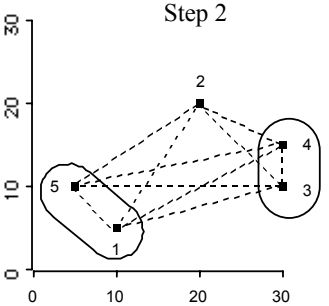
	1	2	5	(34)
1	0.00			
2	18.0	0.00		
5	7.07	18.0	0.00	
(34)	21.5	12.7	25.3	0.00

Step 1



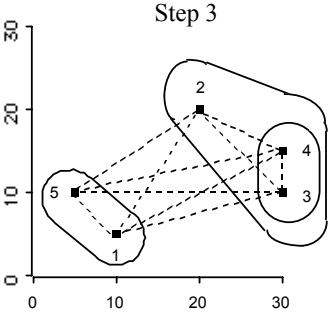
	2	(34)	(15)
2	0.00		
(34)	12.7	0.00	
(15)	18.0	23.4	0.00

Step 2

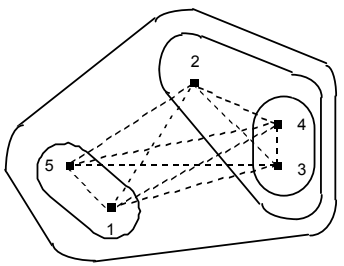


	(15)	(234)
(15)	0.00	
(234)	21.6	0.00

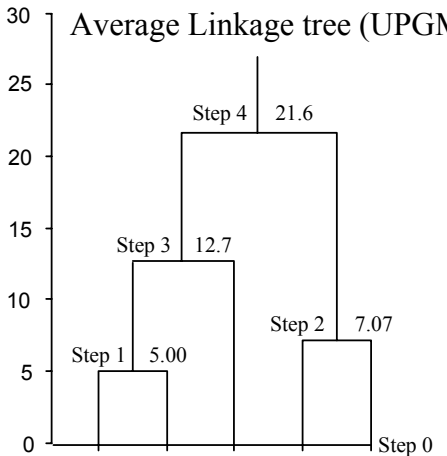
Step 3



Step 4



Average Linkage tree (UPGMA)



Example from Charles H. Romesburg (1984)

1.Centroid method *M. Eisen (TreeViewer)*

In the centroid method each group is replaced by an Average Subject which is the centroid of the group. For example, the first cluster, formed in step1 by combining subjects S3 and S4, is represented by the centoid of S3 and S4 [mean(30, 30), mean(10, 15)]=[30, 12.5]=S34.

See next page for n illustration.

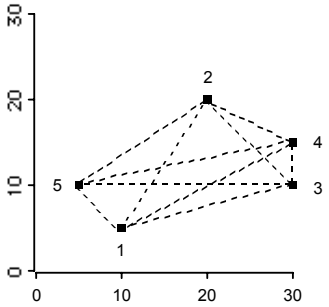
each gene cluster is represented bythe "average" gene profile for all gene profiles in that cluster

While intuitively appealing,he cenroid clustring method is not used much in practice, partly owing to is tendency to produce trees with reversals. Reversals occur when the values at which clusters merge do not increase from one clustering step to the next, but decrese instead. Thus, the tree can collapse onto itself and be difficult o interpret.

Centroid Method

Centroid Method-1

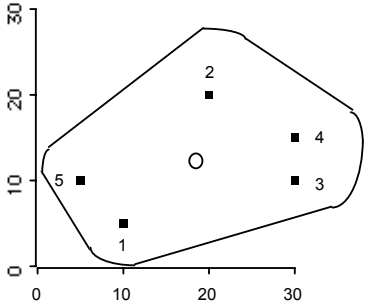
ep 0



	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

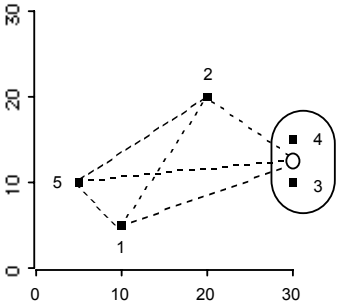
	1	2	3	4	5
1	0.00				
2	18.0	0.00			
3	20.6	14.1	0.00		
4	22.4	11.2	5.00	0.00	
5	7.07	18.0	25.0	25.5	0.00

Step 4



	x1	x2
(12345)	19	12

ep 1



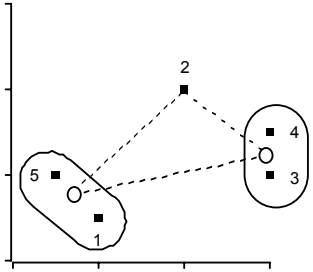
	x1	x2
1	10	5
2	20	20
5	5	10
(34)	30	12.5

	1	2	5	(34)
1	0.00			
2	18.0	0.00		
5	7.07	18.0	0.00	
(34)	21.4	12.5	25.1	0.00

(30+30)/2 (10+15)/2

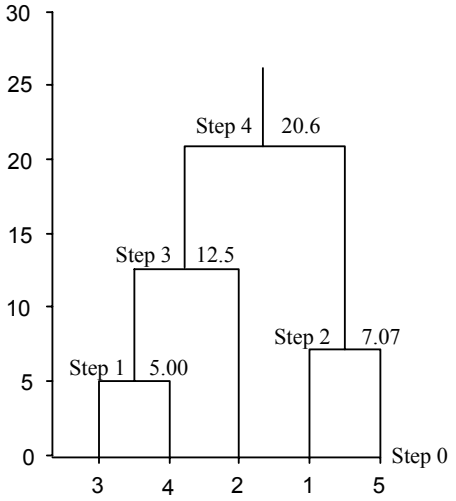
Centroid Method tree

ep 2

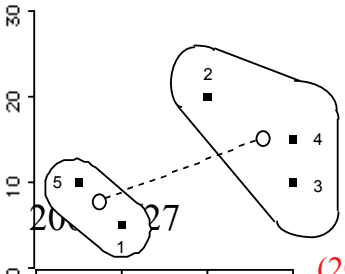


	x1	x2
2	20	20
(34)	30	12.5
(15)	7.5	7.5

	2	(34)	(15)
2	0.00		
(34)	12.5	0.00	
(15)	17.7	23.0	0.00



ep 3



	x1	x2
(15)	7.5	7.5
(234)	26.7	15

	(15)	(234)
(15)	0.00	
(234)	20.6	0.00

(20+30+30)/3 (20+10+15)/3

Ward's Method

The Ward's method does not compute distances between clusters. Rather, it forms clusters by maximizing within-clusters homogeneity. The within-group(i.e., within-cluster) sum of squares is used as the measure of homogeneity. That is, the Ward's method tries to minimize the total within-group or within-cluster sum of squares. Clusters are formed at each step such that the resulting cluster solution has the fewest within-clustersums of squares. The within-cluster sums of squares that is minimized is also known as the error sums of squares(ESS).

To begin Step1, we must first compute E for each of the ten possible mergers.

For example, take the first one: (12), 3, 4, 5. Calculate the cluster mean for (12) = mean(12) = [mean(10, 20), mean(5, 20)] = [15, 12.5].

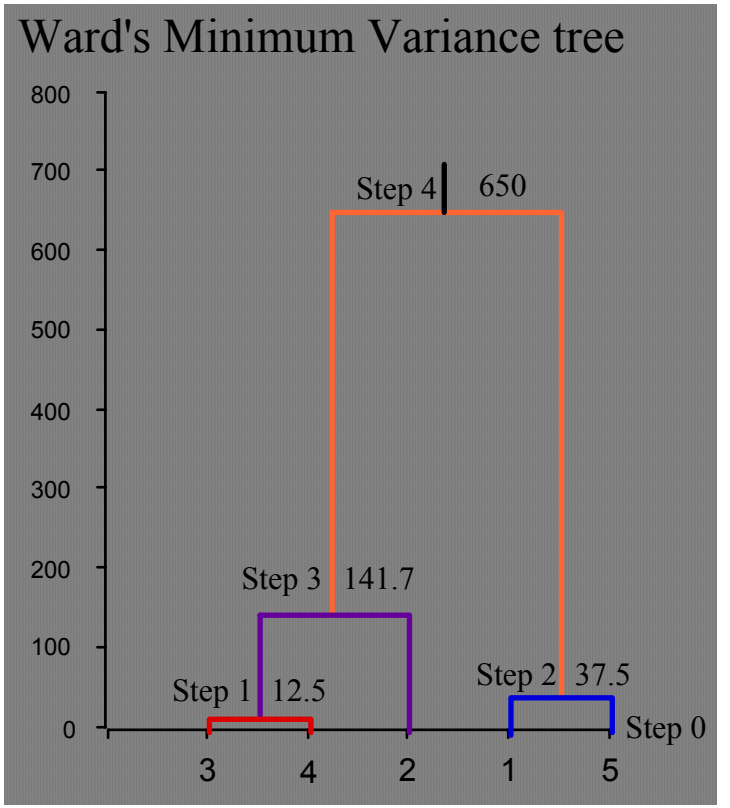
For the first possible merger the value of E is

$$E = (10-15)^2 + (5-12.5)^2 + (20-15)^2 + (20-12.5)^2$$
$$(30-30)^2 + (10-10)^2 + (30-30)^2 + (15-15)^2$$
$$(5-5)^2 + (10-10)^2$$
$$= 162.5$$

Ward's Minimum Variance Method

Example: Charles H. Romesburg (1984)

step	possible partitions				E
1	(12)	3	4	5	162.5
	(13)	2	4	5	212.5
	(14)	2	3	5	250.0
	(15)	2	3	4	25.0
	(23)	1	4	5	100.0
	(24)	1	3	5	62.5
	(25)	1	3	4	162.5
	(34)	1	2	5	12.5*
	(35)	1	2	4	312.5
	(45)	1	2	3	325.0
2	(34)	(12)	5		175.0
	(34)	(15)	2		37.5*
	(34)	(25)	1		175.0
	(134)	2	5		316.7
	(234)	1	5		116.7
	(345)	1	2		433.3
3	(234)	(15)			141.7*
	(125)	(34)			245.9
	(1345)	2			568.8
4	(12345)				650.0



Divisive hierarchical clustering algorithms: Brian S. Everitt (1993)

classification tree? *at root: 2^(n-1) - 1 possible splits*

The divisive clustering method is not used much in practice. Hartigan attributes this to the difficulty in finding effective splitting rules as expense involved in executing them. Divisive clustering techniques are essentially of two types,

–monothetic: which divide the data on the basis of the possession or otherwise of a single specified attribute.

–polythetic: where divisions are based on the values taken by all attributes.

Polythetic

The most feasible of the polythetic divisive methods is that described by MacNaughton-Smith et al. (1964):

A "splinter" group is accumulated by sequential addition of the individual whose total dissimilarity with the remainder, less its total dissimilarity with the splinter group, is a maximum. When this difference becomes negative the process is repeated on the two sub-groups.

Consider, for example, the following distance matrix, D, for seven individuals:

Divisive polythetic method

	1	2	3	4	5	6	7
1	0						
2	10	0					
3	7	7	0				
4	30	23	21	0			
5	29	25	22	7	0		
6	38	34	31	10	11	0	
7	42	36	36	13	17	9	0

D =

26	22.5	20.7	17.2	18.5	22.2	25.5

The individual used to initiate the splinter group is the one whose average distance from the other individuals is a maximum.

→

The initial groups (1) and (2, 3, 4, 5, 6, 7)

?Next the average distance of each individual in the main group to the individuals in the splinter group is found,
?Followed by the average distance of each individual in the main group to the other individuals in this group,
?the difference between these two averages is then found:

Individual	Average distance to splinter group (1)	Average distance to main group (2)	(2-1)
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	-15.2
5	29.0	16.4	-12.6
6	38.0	19.0	-19.0
7	42.0	22.2	-19.8

The maximum difference is 16.4 for individual 3, which is therefore accunulated into the splinter group giving the two groups: (1, 3) ans (2, 4, 5, 6, 7)

Repeated, the process gives the following:

Individual	Average distance to splinter group (1)	Average distance to main group (2)	(2-1)
2	8.5	29.5	21.0
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

So now individual 2 joins the splinter group to give groups: (1, 3, 2) ans (4, 5, 6, 7) and the process is repeated to give:

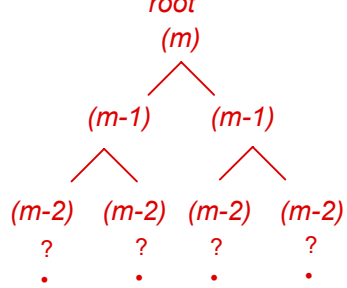
Individual	Average distance to splinter group (1)	Average distance to main group (2)	(2-1)
4	24.3	10.0	-14.3
5	25.3	11.7	-13.6
6	34.3	10.0	-24.3
7	38.0	13.0	-25.0

< 0

As all the differnces are now negative, the process would continue (if desired) on

Monothetic:

m variables



•Monothetic techniques are generally used where the data consists of binary variables. A division is then initially into those individuals who possess and those who lack, some one specified attribute. If divisions of this type only are considered the for a data set with m binary variables there are m potential divisions of the initial set, m-1 each of the two sub-sets thus formed and so on.

•Differences between methods arise because of the different criterion which may be used to choose the particular variable on which to divide. The most common division criteria are based on the chi-square type statistics derived from the 4-fold table for each pair of variables:

$$\chi^2_{jk} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

x_k \ $x_{j\neq k}$	
a	b
c	d

•For example, division might be on that variable, k (say), which makes $\sum_{j\neq k} x^2_{jk}$ a maximum. To illustrate this approach:

Consider the following data set consisting of three binary variables on five individuals.

	Variables		
Individual	1	2	3
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

The three chi-square statistics for each pair of variables can be calculated to be X2(12)=1.87, X2(13)=2.22, X2(23)=0.83, giving

X2(12)+X2(13)=4.09 X2(12)+X2(23)=2.70 X2(13)+X2(23)=3.05

Using the max $\sum_{j \neq k} x_{jk}^2$ criterion, then the first division of the data into two sub-sets is into those individuals who posses variable 1 and those who do not, giving the division (2, 3, 4) and (1, 5).

Such methods have had their widest use in ecological studies.

Charles H. Romesburg (1984):

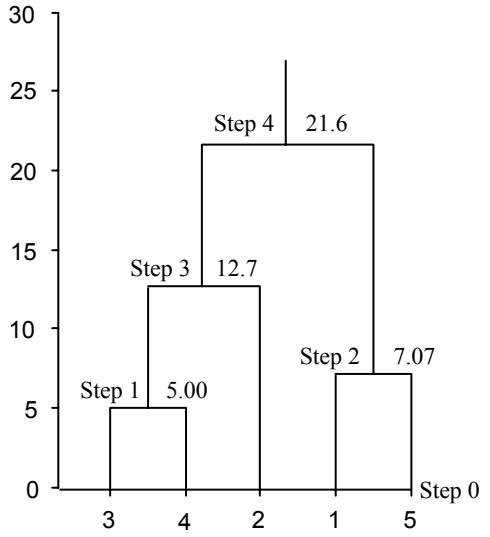
The cophenetic correlation coefficient measures how well the tree and the proximity matrix "say the same thing".

tree structure is displayed to represent (approximate) the input proximity matrix (MDS?)

Original Proximity Matrix

	1	2	3	4	5
1	0.00				
2	18.0	0.00			
3	20.6	14.1	0.00		
4	22.4	11.2	5.00	0.00	
5	7.07	18.0	25.0	25.5	0.00

Average Linkage tree (UPGMA)



Cophenetic Matrix from UPGMA Tree

	1	2	3	4	5
1	0.00				
2	21.6	0.00			
3	21.6	12.7	0.00		
4	21.6	12.7	5.00	0.00	
5	7.07	21.6	21.6	21.6	0.00

Input output

Cell	(2,1)	(3,1)	(4,1)	(5,1)	(3,2)	(4,2)	(5,2)	(4,3)	(5,3)	(5,4)
X	18.0	20.6	22.4	7.07	14.1	11.2	18.0	5.0	25.0	25.5
Y	21.6	21.6	21.6	7.07	12.7	12.7	21.6	5.0	21.6	21.6

The Cophenetic Correlation Coefficient = Correlation Coefficient (x, y)

goodness of fit

Which method?

Subhash Sharma (1996):

The decision depends on the objective of the study and the properties of the various clustering algorithms. Punj and Stewart (1983) have provided comprehensive summaries of the various clustering algorithms and the empirical studies which have compared those algorithms.

Hierarchical → advantage: do not require *a priori* knowledge of the number of clusters

Hierarchical → disadvantage: once an observation is assigned to a cluster it cannot be reassigned to another cluster.

→ Hierarchical methods are sometimes used in an exploratory sense and the resulting solution is submitted to a nonhierarchical method to further refine the cluster solution.

Hierarchical Methods:

1. Hierarchical Methods are susceptible to a chaining effect. That is, observations are sometimes assigned to existing clusters rather than being grouped in new clusters. This is more of a problem if chaining starts early in the clustering process. In general, the nearest neighbor is more susceptible to this problem than the complete-linkage technique. However, chaining sometimes becomes an advantage for identifying nonhomogeneous clusters.

2. Compared to the single-linkage method, the complete-linkage method is less affected by the presence of noise or outliers in the data.

3. The complete-linkage technique typically identifies compact clusters in which the observations are very similar to each other.

4. The Ward's method tends to find clusters that are compact and nearly of equal size and shape.

Nonhierarchical Methods:

- The nonhierarchical clustering algorithms, in general, are very sensitive to the initial partition. It should be further noted that since a number of starting partitions can be used, the final solution could result in local optimization of the objective function.

- Results of simulation studies have shown that K-mean algorithm and other nonhierarchical clustering algorithms perform poorly when random initial partitions are used. However, their performance is much superior when the results from hierarchical methods are used to form the initial partition.

- → It is recommended that for nonhierarchical clustering methods one should use an *a priori* initial partition or cluster solution. In other words. Hierarchical and nonhierarchical techniques should be viewed as complementary clustering techniques rather than as competing techniques.

Panel on Discriminant Analysis and Clustering (1989)

The single and complete linkage methods have the attractive feature that the topologies of the dendrograms are invariant under monotone transformations of the distances. However, the single linkage method is frequently shunned by practitioners because of its propensity to produce long, stringy clusters that are of little interest (see, e.g., Sneath and Sokal, 1973, page 223). The complete linkage method has the opposite problem of being "biased" in the direction of small compact clusters.

The average linkage method is a compromise between the extremes of the other two but it does not have their invariance feature.

- **Simplicity and availability** are probably the primary reasons for their frequent use rather than performance or optimality.

Select a Proximity Measure

- This is never an easy task.
- You must consider the characteristics of your data.
- The scale is important: quantitative, ordinal, nominal, binary or mixed data are very different.
- Before calculating the dissimilarity or similarity, some data transformation may need to apply, like z-score, range or ranking, etc. This equalize the effect of variables with uneven variations.
- Another big problem is: which variables are used for measure of the proximity.

Dissimilarity for Quantitative Data

dissimilarity	equation	metric	Eucl.
Minkowski	$d(i,j) = \left(\sum_k \left| x_{ik} - x_{jk} \right|^p \right)^{1/p}$	Y	Y if p=2
Canberra	$d(i,j) = \sum_k \left| x_{ik} - x_{jk} \right| / \left| x_{ik} + x_{jk} \right|$	Y	N
Soergel	$d(i,j) = \sum_k \left| x_{ik} - x_{jk} \right| / \sum_k \max(x_{ik}, x_{jk})$	Y	N
divergence	$d(i,j) = \sum_k (x_{ik} - x_{jk})^2 / (x_{ik} + x_{jk})^2$	Y	Y
Bary-Curtis	$d(i,j) = \frac{1}{p} \sum_k \left| x_{ik} - x_{jk} \right| / \sum_k (x_{ik} + x_{jk})$	N	N
Wave-Hedges	$d(i,j) = \frac{1}{p} \sum_k \left(1 - \frac{\min(x_{ik}, x_{jk})}{\max(x_{ik}, x_{jk})} \right)$	Y	N
Bhattacharyya	$d(i,j) = \left(\sum_k \left(\sqrt{x_{ik}} - \sqrt{x_{jk}} \right)^2 \right)^{1/2}$	Y	Y

Similarity for Quantitative Data

- Pearson correlation $$s(i,j) = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\mathrm{var}(x_i)\,\mathrm{var}(x_j)}}$$
- Spearman correlation $$s(i,j) = \frac{\mathrm{cov}(r_i, r_j)}{\sqrt{\mathrm{var}(r_i)\,\mathrm{var}(r_j)}}$$ r_i is ranked x_i
- Kendall's tau $$s(i,j) = \frac{1}{\binom{p}{2}} \sum_{k \neq k'} sign\left[(x_{ik} - x_{ik'})(x_{jk} - x_{jk'})\right]$$
- Angular separation $$s(i,j) = \frac{\sum_k x_{ik} x_{jk}}{\sqrt{\sum_k x_{ik}^2 \sum_k x_{jk}^2}}$$
- They can be transformed into dissimilarity by

  1-$s(i,j)$, $\sqrt{1 - s(i,j)}$ or $\sqrt{2(1 - s(i,j))}$

Similarity for Binary Data

a=#{(+,+)}, b=#{(+,-)}, c=#{(-,+)}, d=#{(-,-)}

similarity	equation	metric		Euclidean	
Kulczynski	a/(b+c)	$1-S$ Y	$\sqrt{1-S}$ Y	$1-S$ N	$\sqrt{1-S}$ Y
Rao	a/(a+b+c+d)	Y	Y	N	Y
Jaccard	a/(a+b+c)	Y	Y	N	Y
simple match	(a+d)/(a+b+c+d)	Y	Y	N	Y
Sneath	a/(a+2b+2c)	Y	Y	N	Y
Rogers	(a+d)/(a+2b+2c+d)	Y	Y	N	Y
Hamman	(a+d-b-c)/(a+b+c+d)	Y	Y	N	Y
Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	N	Y	N	Y
Yule	(ad-bc)/(ad+bc)	N	N	N	N

Proximity for Mixed Data

- To form a proximity measure for ordinal or nominal data is a challenge. What is worse, data may come in many different scales, i.e., mixed data.
- Gower (1971) introduce a similarity measure

$$s(i,j) = \frac{\sum_k 1_{ijk} s_k(i,j)}{\sum_k 1_{ijk}}$$

where $s_k(i,j)$ is the similarity between the *i-th* and *j-th* subjects alone the *k-th* variable, and 1_{ijk} is an indicator of whether *i-th* and *j-th* subjects can be compared on the *k-th* variable.

- Cox and Cox (2000) goes further to produce dissimilarity for subjects and for variables at the same time, using the idea of reciprocal averaging.

Measures Compatible with Algorithms

Algorithm	Euclidean metric	Non-Euclidean metric	Semi-metric
Single	✓	✓	✓
Complete	✓	✓	✓
Average	✓	✓	✓
Median	✓	✓	✓
Centroid	✓	?	?
Ward's	✓	✗	✗

Lance, G. N. and Williams, W. T. 1967. A general theory of classificatory sorting strategies. *Computer Journal*, **9**: 373-380.