Summer School in Statistics for Astronomers & Physicists, II
June 6-10, 2006

Cluster Analysis

Donald Richards
Department of Statistics
Center for Astrostatistics
Penn State University

Cluster analysis: Statistical methods for grouping multivariate data into subgroups (or clusters) of similar observations in order to discover structure in the data.

The goal is to find good groupings which allow us to gain additional information about the overall data and the clusters.

Cluster analysis methods are of many types

Some cluster analysis methods are well-grounded in proper statistical inference (hypothesis testing, maximum likelihood estimation)

Other methods are improvised or impromptu

When a cluster analysis is based on a choice of statistical model, the type of clustering may depend on the choice of model.

Chatfield and Collins, *An Introduction to Multivariate Analysis*, Chapman & Hall.

Johnson and Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall.

R. Smith's very nice lecture notes on *Multivariate Analysis*, UNC Chapel Hill, 1999. (You may find it via an Internet search)

Internet search "divisive clustering"

Kaufman and Rousseeuw (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York.

The Ruspini data set

Num.	x	y	Num.	x	y	Num.	x	y
1	4	53	26	41	150	51	98	124
2	5	63	27	38	145	52	99	119
3	10	59	28	38	143	53	99	128
4	9	77	29	32	143	54	101	115
5	13	49	30	34	141	55	108	111
6	13	69	31	44	156	56	110	111
7	12	88	32	44	149	57	108	116
8	15	75	33	44	143	58	111	126
9	18	61	34	46	142	59	115	117
10	19	65	35	47	149	60	117	115
11	22	74	36	49	152	61	70	4
12	27	72	37	50	142	62	77	12
13	28	76	38	53	144	63	83	21
14	24	58	39	52	152	64	61	15
15	27	55	40	55	155	65	69	15
16	28	60	41	54	124	66	78	16
17	30	52	42	60	136	67	66	18
18	31	60	43	63	139	68	58	13
19	32	61	44	86	132	69	64	20
20	36	72	45	85	115	70	69	21
21	28	147	46	85	96	71	66	23
22	32	149	47	78	94	72	61	25
23	35	153	48	74	96	73	76	27
24	33	154	49	97	122	74	72	31
25	38	151	50	98	116	75	64	30

A plot of the data set suggests that it consists of at least four clusters.

y
0
50
100
150
x
0
20
40
60
80
100
120

Hierarchical clustering

An advantage of hierarchical clustering is that it requires no assumptions about the model.

Agglomerative hierarchical clustering: Start with each individual object classified as a separate cluster. Find the minimal distance between each cluster, and merge neighboring clusters which are "most similar." Repeat the process until all subgroups are merged into a single cluster.

Divisive hierarchical clustering: A reverse of agglomerative clustering: Start from a single "cluster" covering all the data; then successively split the cluster into smaller clusters according to the particular distance function being used.

Much information is available from the Internet regarding hierarchical and other clustering methods. For instance, from
http://fconyx.ncifcrf.gov/~lukeb/diclust.html
we read:

"Divisive clustering starts by placing all objects into a single group. Before we start the procedure, we need to decide on a threshold distance. Once this is done, the procedure is as follows:

"1. The distance between all pairs of objects within the same group is determined and the pair with the largest distance is selected.

“2. This maximum distance is compared to the threshold distance.

“If it is larger than the threshold, this group is divided in two. This is done by placing the selected pair into different groups and using them as seed points. All other objects in this group are examined, and are placed into the new group with the closest seed point. The procedure then returns to Step 1.

“If the distance between the selected objects is less than the threshold, the divisive clustering stops.

“To run a divisive clustering, you simply need to decide upon a method of measuring the distance between two objects.”

The results of these procedures are represented by a *dendogram*.

Smith (1999) gives the following:

The *S-PLUS* or *R* commands required to create dendogram plots for divisive clustering:

```
x<-matrix(scan(file='ruspini.dat'),ncol=2,byrow=T)
y<-dist(x,metric="euclidean")
y1<-hclust(y,method="connected")
plclust(y1)
```

"Here, the first line reads the data into a matrix x, and the second creates a distance vector y. If x has n rows and $1 \leq i < j \leq n$, the distance between the ith and jth rows of x is the $\{(n(i-1) - \frac{i(i-1)}{2} + j - i\}$th entry of y. The distance here is the usual Euclidean distance between two vectors but we could also specify metric="manhattan" (the sum of absolute differences of the components of the two rows) or metric="binary" (the proportion of non-zero elements that the two vectors do not have in common).

"The third line applies the hierarchical clustering algorithm using the "connected" method (i.e., using the minimum distance between the points of two clusters to define the distance between the cluster – this is also called single link clustering) and the fourth line draws the dendogram.

"Apart from the single link clustering algorithm just mentioned, there are a number of other hierarchical clustering algorithms based on different distance measures between clusters:

"*Average distance method* ("average" in SPlus) – the distance between two clusters is the average of the distances between the members of the clusters;

"*Complete linkage method* ("compact" in SPlus) – the distance between two clusters is the maximum of the individual distances between points of the cluster;

"*Centroid method* (not implemented as part of the SPlus hclust algorithm described above, but it is available through the "mclust" algorithm described in the next section, where it is given by the option method="centroid") – the distance between two clusters is the distance between their centroids;

"*Sum of squares method* (also known as *Ward's method*, or the *trace method*: method="sum of squares" or method="trace" within the "mclust" algorithm) – this splits clusters in a way which minimizes the total within-cluster sum of squares."

Model-Based Clustering

Clusters $C_1, \ldots, C_G$

G is *unknown*

We have some belief about the probability models for the clusters $C_1, \ldots, C_G$

We believe that each cluster C_k represents a population with probability density function $f(x; \theta_k)$ where θ_k is a set of unknown parameters.

For instance, we may believe that each C_k is a multivariate normal population with mean vector μ_k and covariance matrix $\mathbf{\Sigma}_k$, so $\theta_k = \{\mu_k, \mathbf{\Sigma}_k\}$.

We want to estimate G and all parameters such that the resulting likelihood function is maximized.

We also need to "develop" an assignment rule, γ. The rule γ determines the assignment of observations to clusters.

For example, if $\gamma_3 = 7$ then the third observation is assigned to the seventh cluster.

The likelihood function:

$$L(\gamma, \theta_1, \ldots, \theta_k) = \prod_{x \in C_1} f(x; \theta_1) \cdots \prod_{x \in C_G} f(x; \theta_G)$$

Let n_k denote the number of observations assigned to C_k.

In the multivariate normal case, the maximum likelihood estimator of μ_k is $\bar{x}_k$, the sample mean of the kth cluster. Also, the maximum likelihood estimator of $\mathbf{\Sigma}_k$ is $(n_k - 1)S_k/n_k$, where S_k is the sample covariance matrix of the kth cluster.

Assume that γ, all n_k, and G are known

Follow the rules for maximizing likelihood functions

$$\begin{aligned} L(\gamma, \widehat{\theta}_1, \ldots, \widehat{\theta}_k) &= \prod_{x \in C_1} f(x; \widehat{\theta}_1) \cdots \prod_{x \in C_G} f(x; \widehat{\theta}_G) \\ &\propto \prod_{k=1}^{G} |S_k|^{-n_k/2}. \end{aligned}$$

Problem: Find the value of G, $n_1, \ldots, n_G$, and the corresponding optimal allocation rule $\widehat{\gamma}$ such that $L(\widehat{\gamma}, \widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is maximized.

Note: We cannot use calculus to find G.

There is a large literature on the problem of maximizing L.

To maximize L, we must search over a *very* large set of possibilities.

Genetic algorithms for cluster analysis

Lozano and Larrañaga, Pattern Recognition Letters, 1999; and other articles.

Robust Methods for Cluster Analysis

The problem with astronomers: Their data always seem to contain outliers, measurements with errors, influential points, (and worst of all) *non-random* samples.

The problem with statisticians: They insist on cluster analysis without a strict definition of a cluster. O ur personal observations: Standard clustering methods often do not work well in the presence of outliers, etc.

Kaufman and Rousseeuw (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York.

Antwerp Group On Robust & Applied Statistics

http://www.agoras.ua.ac.be/

Rousseeuw, P.J. and Van Driessen, K. (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator, Technometrics, 41, 212-223.

ftp://ftp.win.ua.ac.be/pub/preprints/99/Fasalg99.pdf

Page 214, Problem 2: Analysis of data from the Digitized Palomar Sky Survey