Mathematics Learning Support Centre

Statistics: 3.1 Cluster Analysis

Rosie Cornish. 2007.

1 Introduction

This handout is designed to provide only a brief introduction to cluster analysis and how it is done. Books giving further details are listed at the end.

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. An example where this might be used is in the field of psychiatry, where the characterisation of patients on the basis of of clusters of symptoms can be useful in the identification of an appropropriate form of therapy. In marketing, it may be useful to identify distinct groups of potential customers so that, for example, advertising can be appropriately targetted.

WARNING ABOUT CLUSTER ANALYSIS
Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included.

2 Approaches to cluster analysis

There are a number of different methods that can be used to carry out a cluster analysis; these methods can be classified as follows:

- Hierarchical methods
  - Agglomerative methods, in which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions.
  - Divisive methods, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Agglomerative methods are used more often than divisive methods, so this handout will concentrate on the former rather than the latter.
- Non-hierarchical methods (often known as **k-means clustering** methods)

3 Types of data and measures of distance

The data used in cluster analysis can be interval, ordinal or categorical. However, having a mixture of different types of variable will make the analysis more complicated. This is because in cluster analysis you need to have some way of measuring the distance between observations and the type of measure used will depend on what type of data you have.

A number of different measures have been proposed to measure 'distance' for binary and categorical data. For details see the book by Everitt, Landau and Leese. Readers are also referred to this text for details of what to do if you have a mixture of different data types. For interval data the most common distance measure used is the **Euclidean distance**.

3.1 Euclidean distance

In general, if you have p variables $X_1, X_2, \ldots, X_p$ measured on a sample of n subjects, the observed data for subject i can be denoted by $x_{i1}, x_{i2}, \ldots, x_{ip}$ and the observed data for subject j by $x_{j1}, x_{j2}, \ldots, x_{jp}$. The Euclidean distance between these two subjects is given by

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

When using a measure such as the Euclidean distance, the scale of measurement of the variables under consideration is an issue, as changing the scale will obviously effect the distance between subjects (e.g. a difference of 10cm could being a difference of 100mm). In addition, if one variable has a much wider range than others then this variable will tend to dominate. For example, if body measurements had been taken for a number of different people, the range (in mm) of heights would be much wider than the range in wrist circumference, say. To get around this problem each variable can be standardised (converted to z-scores). However, this in itself presents a problem as it tends to reduce the variability (distance) between clusters. This happens because if a particular variable separates observations well then, by definition, it will have a large variance (as the between cluster variability will be high). If this variable is standardised then the separation between clusters will become less. Despite this problem, many textbooks do recommend standarisation. If in doubt, one strategy would be to carry out the cluster analysis twice — once without standardising and once with — to see how much difference, if any, this makes to the resulting clusters.

4 Hierarchical agglomerative methods

Within this approach to cluster analysis there are a number of different methods used to determine which clusters should be joined at each stage. The main methods are summarised below.

- Nearest neighbour method (single linkage method)
  In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but is often criticised because it doesn't take account of cluster structure and can result in a problem called **chaining** whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape.

- Furthest neighbour method (complete linkage method)
  In this case the distance between two clusters is defined to be the maximum distance between members — i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

- Average (between groups) linkage method (sometimes referred to as UPGMA)
  The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method.

- Centroid method
  Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.

- Ward's method
  In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

It is generally a good idea to try two or three of the above methods. If the methods agree reasonably well then the results will be that much more believable.

4.1 Selecting the optimum number of clusters

As stated above, once the cluster analysis has been carried out it is then necessary to select the 'best' cluster solution. There are a number of ways in which this can be done, some rather informal and subjective, and some more formal. The more formal methods will not be discussed in this handout. Below, one of the informal methods is briefly described.

When carrying out a hierarchical cluster analysis, the process can be represented on a diagram known as a **dendrogram**. This diagram illustrates which clusters have been joined at each stage of the analysis and the distance between clusters at the time of joining. If there is a large jump in the distance between clusters from one stage to another then this suggests that at one stage clusters that are relatively close together were joined whereas, at the following stage, the clusters that were joined were relatively far apart. This implies that the optimum number of clusters may be the number present just before that large jump in distance. This is easier to understand by actually looking at a dendrogram — see references for further information.

5 Non-hierarchical or k-means clustering methods

In these methods the desired number of clusters is specified in advance and the 'best' solution is chosen. The steps in such a method are as follows:

1. Choose initial cluster centres (essentially this is a set of observations that are far apart — each subject forms a cluster of one and its centre is the value of the variables for that subject).
2. Assign each subject to its 'nearest' cluster, defined in terms of the distance to the centroid.
3. Find the centroids of the clusters that have been formed
4. Re-calculate the distance from each subject to each centroid and move observations that are not in the cluster that they are closest to.
5. Continue until the centroids remain relatively stable.

Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster). Two disadvantages of non-hierarchical cluster analysis are: (1) it is often difficult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times and (2) it can be very sensitive to the choice of initial cluster centres. Again, it may be worth trying different ones to see what impact this has.

One possible strategy to adopt is to use a hierarchical approach initially to determine how many clusters there are in the data and then to use the cluster centres obtained from this as initial cluster centres in the non-hierarchical method.

6 Carrying out cluster analysis in SPSS

6.1 Hierarchical cluster analysis

– **Analyze**
– **Classify**
– **Hierarchical cluster**
– Select the variables you want the cluster analysis to be based on and move them into the **Variable(s)** box.
– In the **Method** window select the clustering method you want to use. Under **Measure** select the distance measure you want to use and, under **Transform values**, specify whether you want all variables to be standardised (e.g. to z-scores) or not.
– In the **Statistics** window you can specify whether you want to see the **Proximity Matrix** (this will give the distance between all observations in the data set — only really recommended for relatively small data sets!). You can also specify whether you want the output to include details of cluster membership — either for a fixed number of clusters or for a range of cluster solutions (e.g. 2 to 5 clusters).
– In the **Save** window you can specify whether you want SPSS to save details of cluster membership — again, either for a fixed number of clusters or for a range of cluster solutions (e.g. 2 to 5 clusters). If you ask it to do this, this information will be included as additional variables at the end of the data set.
– In the **Plots** window you can specify which plots you would like included in the output.
– **OK**

6.2 K-means cluster analysis

– **Analyze**
– **Classify**
– **K-means cluster**
– Select the variables you want the cluster analysis to be based on and move them into the **Variable(s)** box.
– Under **Method**, ensure that **Iterate and Classify** is selected (this is the default).
– In the **Iterate** window you can specify how many iterations you would like SPSS to perform before stopping. The default is ten. It might be worth leaving it as ten to start with and then increasing this if convergence doesn't occur (i.e. a stable cluster solution is not reached) within ten iterations.
– In the **Save** window you can specify whether you want SPSS to save details of cluster membership and distance to the cluster centre for each subject (observation).
– **OK**

7 References

- Everitt, B.S., Landau, S. and Leese, M. (2001), **Cluster Analysis**, Fourth edition, Arnold.
- Manly, B.F.J. (2005), **Multivariate Statistical Methods: A primer**, Third edition, Chapman and Hall.
- Rencher, A.C. (2002), **Methods of Multivariate Analysis**, Second edition, Wiley.