Chapter 448

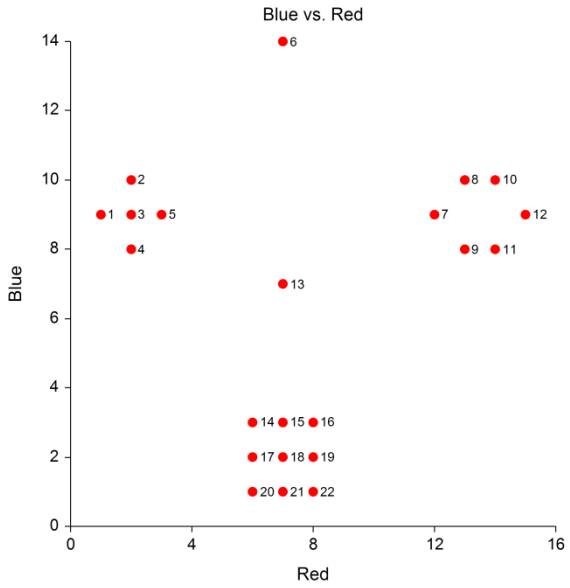
Fuzzy Clustering

Introduction

Fuzzy clustering generalizes partition clustering methods (such as k-means and medoid) by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual is a member of only one cluster. Suppose we have *K* clusters and we define a set of variables $m_{i1}, m_{i2}, \cdots, m_{iK}$ that represent the probability that object *i* is classified into cluster *k*. In partition clustering algorithms, one of these values will be one and the rest will be zero. This represents the fact that these algorithms classify an individual into one and only one cluster.

In fuzzy clustering, the membership is spread among all clusters. The m_{ik} can now be between zero and one, with the stipulation that the sum of their values is one. We call this a *fuzzification* of the cluster configuration. It has the advantage that it does not force every object into a specific cluster. It has the disadvantage that there is much more information to be interpreted.

To understand the reason that fuzzy clustering was developed, consider the following two-variable dataset whose values are plotted below.



The data have three obvious clusters and two outlier points (6 and 13). A regular clustering algorithm searching for three clusters will force these two points into specific clusters. This may cause distortion in the final solution. Fuzzy clustering, however, will assign a probability of about 0.33 for each cluster. This equal membership probability signals that these two points are outliers.

When you only have two variables, you can plot your data and see what the clusters are. Unfortunately, most clustering projects come with more than two variables, so plotting is not possible. Hence, we must use techniques like fuzzy clustering to deal with the anomalies that can occur.

Dissimilarities

The formation of the distances (dissimilarities) was described in the Medoid Clustering chapter and is not repeated here.

Fuzzy Algorithm

The fuzzy algorithm used by this program is described in Kaufman (1990). It seeks to minimize the following objective function, *C*, made up of cluster memberships and distances.

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ik}^2 m_{jk}^2 d_{ij}}{2\sum_{j=1}^{N} m_{jk}^2}$$

where m_{ik} represents the unknown membership of the object *i* in cluster *k* and d_{ij} is the dissimilarity between objects *i* and *j*. The memberships are subject to constraints that they all must be non-negative and that the memberships for a single individual must sum to one. That is, the memberships have the same constraints that they would if they were the probabilities that an individual belongs to each group (and they may be interpreted as such).

Goodness-of-Fit

One of the most difficult tasks in cluster analysis is choose the appropriate number of clusters. In fuzzy clustering, the following coefficients are used in conjunction with the silhouette values that are defined in the Medoid Clustering chapter.

The amount of 'fuzziness' in a solution may be measured by *Dunn's partition coefficient* which measures how close the fuzzy solution is to the corresponding hard solution. This *hard* solution is formed by classifying each object into the cluster which has the largest membership. The formula for Dunn's partition coefficient is

$$F(U) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} m_{ik}^2$$

This coefficient ranges from 1/*K* to 1. Its value is 1/*K* when all memberships are equal to 1/*K*. The value of one results when, for each object, the value of one membership is unity and the rest are zero.

Dunn's partition coefficient may be normalized so that it varies from 0 (completely fuzzy) to 1 (hard cluster). The normalized version is

$$Fc(U) = \frac{F(U) - (1/K)}{1 - (1/K)}$$

Another partition coefficient, given in Kaufman (1990), is

$$D(U) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}(h_{ik} - m_{ik})^2$$

This coefficient ranges from 0 (hard clusters) to 1-1/*K* (completely fuzzy). The normalized version of this equation is:

$$Dc(U) = \frac{D(U)}{1 - (1/K)}$$

Fc(U) and Dc(U) together give a good indication of an optimum number of clusters. You should choose *K* so that Fc(U) is large and Dc(U) is small.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

The data given in the following table were shown on the scatter plot displayed earlier and are found in the Fuzzy dataset. They are from a concocted database found in Kaufman (1990) designed specifically to show the usefulness of fuzzy clustering.

Fuzzy dataset (subset)

Red	Blue	ID
1	9	1
2	10	2
2	9	3
2	8	4
3	9	5
7	14	6
12	9	7
13	10	8
13	8	9

Missing Values

When an observation has missing values, appropriate adjustments are made so that the average dissimilarity across all variables with data may be computed. That is, rows with missing values are not omitted unless all variables have missing values. Note that the distances require that at least one variable have non-missing values for each pair of rows.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Interval Variables

Designates interval-type variables (if any) or the columns of the matrix if distance or correlation matrix input was selected. Interval variables are continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

In general, an interval should keep the same importance throughout the scale. For example, the length of time between 1905 and 1925 is the same as the length of time between 1995 and 2015.

Note that a nonlinear transformation of an interval variable is probably not an interval variable. For example, the logarithm of height is not an interval variable since the value of an interval along the scale changes depending upon where you are on the scale.

Ratio Variables

Specifies the ratio variables (if any). Ratio-type variables are positive measurements in which the distinction between two numbers is constant if their ratio is constant. For example, the distinction between 3 and 30 would have the same meaning as the distinction between 30 and 300. Examples are chemical concentration or radiation intensity.

The logarithms of ratio variables are analyzed as if they were interval variables.

Ordinal Variables

Specifies the ordinal-type variables (if any). Ordinal variables are measurements that may be ordered according to magnitude. For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1). Interval variables are ordinal, but ordinal variables are not necessarily interval.

The original values of ordinal variables are replaced by their ranks. These ranks are then analyzed as if they were interval variables.

Nominal Variables

Specifies the nominal-type variables (if any). Nominal variables are those in which the number represents the state of the variable. Examples include gender, race, hair color, country of birth, or zipcode. If a nominal variable has only two categories, it is often called a binary variable.

Nominal variables are analyzed using the number of matches between two individuals.

Symmetric-Binary Variables

Specifies the symmetric binary-type variables (if any). Symmetric binary variables have two possible outcomes, each of which carry the same information and weight. Examples include gender, marital status, or membership in a particular group. Usually, they are coded as 1 for yes or 0 for no, although this is not necessary. These variables are analyzed using the number of matches between two individuals.

Asymmetric-Binary Variables

Specifies the asymmetric binary-type variables (if any). Asymmetric binary-scaled variables are concerned with the presence or absences of a relatively rare event, the absence of which is unimportant.

These variables are analyzed using the number of matches in which both individuals have the trait of interest. Those cases in which both individuals do not have the trait are not of interest and are ignored.

Clustering Options

Distance Method

This option specifies whether Euclidean or Manhattan distance is used. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Scaling Method

Specify the type of scaling to be used on Interval, Ordinal, and Ratio variables. Possible choices are Standard Deviation, Average Absolute Deviation, Range, and None. These were discussed in the introduction to this chapter.

Max Iterations

This option sets a maximum number of iterations that are attempted before the algorithm terminates. This avoids the possible of the algorithm going into an infinite loop.

Fuzzifier Constant

Specifies the exponent of the memberships in the objective function that is being minimized. Normally, this value is set to two. In some situations, you may want to change this value. The value must be strictly greater than one. As this value is decreased from two towards one, the final solution will appear less and less fuzzy. That is, the membership values will be closer to either zero or one. Also, values of this option near one cause the algorithm to converge more slowly.

Minimum Change

When the change in the objective function from one iteration to the next is less than this amount, the algorithm terminates.

Maximum row

The maximum number of rows that will be analyzed by this procedure.

Clustering Options – Numbers of Clusters

Minimum Clusters

The minimum value of *K* to search. A separate analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters.

Maximum Clusters

The maximum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters.

Reported Clusters

The is the cluster configuration that is stored on the database if the Cluster Id or Membership Out Variables options are specified..

Format Options

Label Variable

This is an optional variable containing identification for each row (object). These labels are used to enhance the interpretability of the reports.

Input Format

Specify the type of data format that you have. Your choices are

- **Raw Data**

  The variables are in the standard format in which each row represents an object and each column represents a variable.

- **Distances**

  The variables containing a distance matrix are specified in the Interval Variables option. Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

- **Correlations 1**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = \frac{1 - r_{ij}}{2}$$

- **Correlations 2**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - \left| r_{ij} \right|$$

- **Correlations 3**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - r_{ij}^2$$

Note that all three types of correlation matrices must be completely specified. You cannot specify only the lower or upper triangular portions. Also, the rows correspond to variables. That is, the values along the first row represent the correlations of the first variable with each of the other variables. Hence, you cannot rearrange the order of the matrix.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Membership Report - Summary Report

Specify whether to display the indicated reports.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Storage Tab

These options let you specify where to store various row-wise statistics.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify columns that contain important data.

Store Membership Out in Variable

You can automatically store the row memberships into the columns specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Example 1 – Fuzzy Clustering

This section presents an example of how to run a cluster analysis. The data used found in the Fuzzy dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Fuzzy Clustering window.

1 Open the Fuzzy dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Fuzzy.NCSS**.
- Click **Open**.

2 Open the Fuzzy Clustering window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Fuzzy Clustering** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Fuzzy Clustering window, select the **Variables tab**.
- Double-click in the **Interval Variables** box. This will bring up the variable selection window.
- Select **Red** and **Blue** from the list of variables and then click **Ok**. "Red-Blue" will appear in the Interval Variables box.
- Under Clustering Options, set the **Scaling Method** to **None**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Summary Section

Summary Section

Number Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	24.294968	0.535378	0.6799	0.3598	0.1400	0.2800
3	11.366128	0.704072	0.7102	0.5653	0.0861	0.1291
4	8.594860	0.487322	0.5422	0.3896	0.2161	0.2881
5	6.768054	0.340839	0.4783	0.3479	0.2837	0.3546

This report actually appears last on the printout, but it is the first section that should be studied. This report lets you select the appropriate number of clusters. Select the number of clusters that maximizes the Average Silhouette and Fc(U) while minimizing Dc(U). In this case, three clusters are selected.

Average Distance

This is the value of the average dissimilarity. Note that this value has been rescaled as a percentage from the maximum distance in the dissimilarity matrix to improve readability.

Average Silhouette

This is the average of the silhouette values of all rows. The Silhouette statistic is discussed in the Medoid Partitioning chapter. It is used to aid in the search for the appropriate number of clusters by selecting the number of clusters that maximizes this value.

F(U), Fc(U), D(U), Dc(U)

The definitions of these statistics were presented earlier. Here we will note that we search for the number of clusters that maximizes Fc(U) and minimizes Dc(U). There will not always be an obvious choice as in this example.

Once the appropriate number of clusters has been determined, the solution can be studied in detail. Since three clusters is appropriate for this database, only the results for three clusters will be shown here.

Cluster Medoids Section

Cluster Medoids Section

Variable	Cluster1	Cluster2	Cluster 3
Red	2	14	7
Blue	9	10	2
Row	3	10	18

This report gives the medoid (most centrally located) of the nearest hard cluster configuration. It is provided to help you recognize and interpret cluster. The last row of the report gives the row number (and label if designated) of the each cluster's medoid.

Membership Summary Section (Clusters = 3)

Membership Summary Section for Clusters = 3

Row	Cluster	Cluster Membership	Sum of Squared Memberships	Bar of Squared Memberships	Silhouette Amount	Silhouette Bar
3	1	0.9362	0.8786	||||||||||||||||||||||||	0.7337	|||||||||||||||||||
2	1	0.8785	0.7792	||||||||||||||||||||	0.7313	|||||||||||||||||||
5	1	0.8741	0.7722	||||||||||||||||||||	0.6840	|||||||||||||||||
1	1	0.8677	0.7618	||||||||||||||||||||	0.6957	|||||||||||||||||
4	1	0.8606	0.7507	||||||||||||||||||||	0.6400	||||||||||||||||
6	1	0.4205	0.3531	||||||||||	0.1392	||||
10	2	0.8745	0.7727	||||||||||||||||||||	0.8284	|||||||||||||||||||||
8	2	0.8718	0.7683	||||||||||||||||||||	0.8168	|||||||||||||||||||||
11	2	0.8613	0.7517	||||||||||||||||||||	0.8033	|||||||||||||||||||
9	2	0.8564	0.7439	||||||||||||||||||	0.7854	|||||||||||||||||||
12	2	0.8386	0.7164	||||||||||||||||||	0.8023	|||||||||||||||||||
7	2	0.8188	0.6870	||||||||||||||||||	0.7523	||||||||||||||||||
18	3	0.9196	0.8489	||||||||||||||||||||||	0.8228	||||||||||||||||||||
21	3	0.8668	0.7602	||||||||||||||||||||	0.7976	|||||||||||||||||||
19	3	0.8599	0.7492	|||||||||||||||||||	0.7840	|||||||||||||||||||
17	3	0.8589	0.7478	|||||||||||||||||||	0.7790	||||||||||||||||||
15	3	0.8524	0.7375	|||||||||||||||||||	0.7834	|||||||||||||||||||
22	3	0.8226	0.6924	||||||||||||||||||	0.7630	||||||||||||||||||
20	3	0.8222	0.6921	||||||||||||||||||	0.7604	||||||||||||||||||
16	3	0.8012	0.6617	|||||||||||||||||	0.7444	|||||||||||||||||
14	3	0.7992	0.6593	||||||||||||||||||	0.7342	||||||||||||||||||
13	3	0.3734	0.3393	|||||||||	0.1086	|||

This report displays information about each row. The report is sorted by Silhouette Value within cluster. Notice how well the two outliers, rows six and thirteen, stand out on this report.

Row

The row number and, if designated, label of this individual. Each row of the database is represented on this report.

Cluster

This is the number of the cluster into which this row was classified.

Cluster Membership

This is the maximum of the memberships. It is the membership value for the cluster into which this row was assigned for the hard clustering.

Sum of Squared Memberships

All memberships for a given row are squared and summed. When a row is completely assigned to a single cluster, this value will be one. When the row is equally likely to be classified into each cluster, the value will be $1/K$. Hence, rows with high values here are near the center of a cluster. Rows with low values here are outliers.

Bar of Squared Memberships

This is a bar graph of the sum of squared membership values. It will help you to detect rows that are not well clustered.

Silhouette Amount

This is the value of the silhouette. Its interpretation was presented in the introduction to the Medoid Clustering chapter and will not be repeated here. We note that the value should be positive and most rows should be greater than 0.50.

Silhouette Bar

This is a bar graph of the silhouette values. It will help you to detect rows that are not well clustered.

Membership Matrix Section

Membership Matrix Section

Row	Cluster	Prob in 1	Prob in 2	Prob in 3
1	1	0.8677	0.0564	0.0759
2	1	0.8785	0.0551	0.0664
3	1	0.9362	0.0274	0.0364
4	1	0.8606	0.0562	0.0832
5	1	0.8741	0.0549	0.0709
6	1	0.4205	0.3545	0.2250
7	2	0.0849	0.8188	0.0963
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

This report displays the membership of each row in each cluster.