Chapter 448

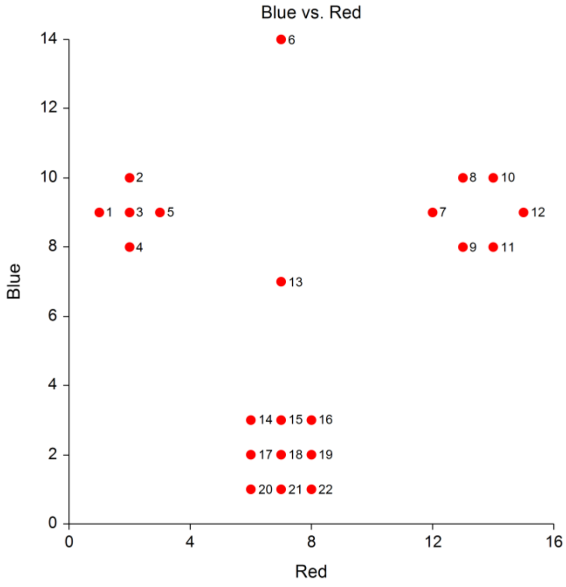
Fuzzy Clustering

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 $m_{i1}, m_{i2}, \ldots, m_{iK}$ 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 Statistics

One of the most difficult tasks in cluster analysis is choosing 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

Data Input Formats

A number of input formats are available.

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.

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.

Example 1 – Fuzzy Clustering

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

Setup

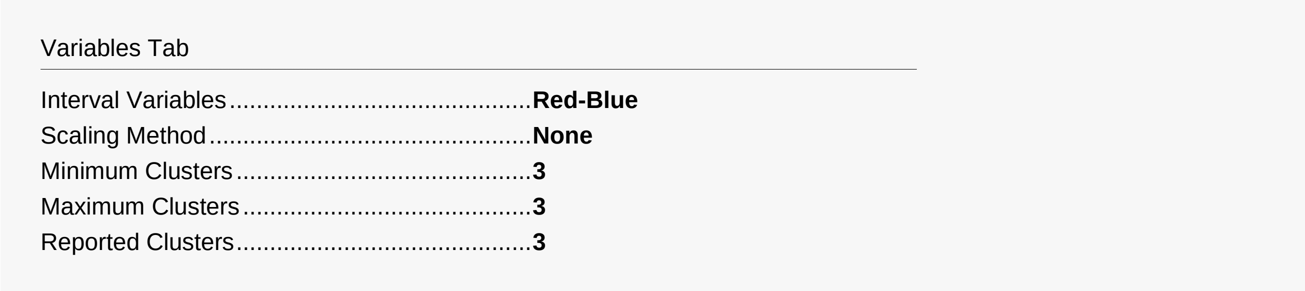
To run this example, complete the following steps:

1 Open the Fuzzy example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fuzzy** and click **OK**.

2 Specify the Fuzzy Clustering procedure options

- Find and open the **Fuzzy Clustering** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



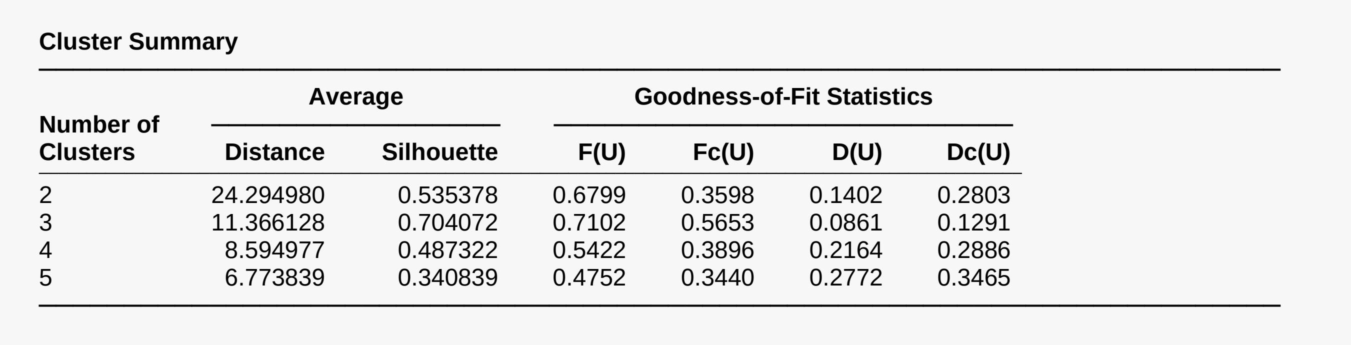
Variables Tab

Interval Variables..**Red-Blue**
Scaling Method...**None**
Minimum Clusters...**3**
Maximum Clusters...**3**
Reported Clusters...**3**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Cluster Summary



Cluster Summary

	Average		Goodness-of-Fit Statistics			
Number of Clusters	**Distance**	**Silhouette**	**F(U)**	**Fc(U)**	**D(U)**	**Dc(U)**
2	24.294980	0.535378	0.6799	0.3598	0.1402	0.2803
3	11.366128	0.704072	0.7102	0.5653	0.0861	0.1291
4	8.594977	0.487322	0.5422	0.3896	0.2164	0.2886
5	6.773839	0.340839	0.4752	0.3440	0.2772	0.3465

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.

Goodness-of-Fit Statistics: 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

```
Cluster Medoids (3 Clusters)
─────────────────────────────────────────────────────────
                  Cluster Medoid
                 ───────────────
Variable          1      2      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 the clusters. The last row of the report gives the row number (and label if designated) of each cluster's medoid.

Membership Summary

Membership Summary (3 Clusters)

	Cluster		Squared Memberships		Silhouette	
Row	**ID**	**Membership**	**Sum**	**Bar**	**Value**	**Bar**
3	1	0.9362	0.8786	|IIIIIIIIIIIIIIIIIIIIIIIIII	0.7337	|IIIIIIIIIIIIIIIIIIIIII
2	1	0.8785	0.7792	|IIIIIIIIIIIIIIIIIIIIIII	0.7313	|IIIIIIIIIIIIIIIIIIIIII
5	1	0.8742	0.7722	|IIIIIIIIIIIIIIIIIIIIIII	0.6840	|IIIIIIIIIIIIIIIIIIIII
1	1	0.8677	0.7618	|IIIIIIIIIIIIIIIIIIIIIII	0.6957	|IIIIIIIIIIIIIIIIIIIII
4	1	0.8606	0.7508	|IIIIIIIIIIIIIIIIIIIIIII	0.6400	|IIIIIIIIIIIIIIIIIII
6	1	0.4205	0.3531	|IIIIIIIIIII	0.1392	|IIII
10	2	0.8746	0.7728	|IIIIIIIIIIIIIIIIIIIIIII	0.8284	|IIIIIIIIIIIIIIIIIIIIIIIII
8	2	0.8718	0.7683	|IIIIIIIIIIIIIIIIIIIIIII	0.8168	|IIIIIIIIIIIIIIIIIIIIIIIII
11	2	0.8614	0.7517	|IIIIIIIIIIIIIIIIIIIIIII	0.8033	|IIIIIIIIIIIIIIIIIIIIIIII
9	2	0.8564	0.7438	|IIIIIIIIIIIIIIIIIIIIII	0.7854	|IIIIIIIIIIIIIIIIIIIIIIII
12	2	0.8386	0.7164	|IIIIIIIIIIIIIIIIIIIII	0.8023	|IIIIIIIIIIIIIIIIIIIIIIII
7	2	0.8188	0.6869	|IIIIIIIIIIIIIIIIIIIII	0.7523	|IIIIIIIIIIIIIIIIIIIIIII
18	3	0.9196	0.8489	|IIIIIIIIIIIIIIIIIIIIIIIII	0.8228	|IIIIIIIIIIIIIIIIIIIIIIIII
21	3	0.8668	0.7602	|IIIIIIIIIIIIIIIIIIIIIII	0.7976	|IIIIIIIIIIIIIIIIIIIIIIII
19	3	0.8599	0.7493	|IIIIIIIIIIIIIIIIIIIIII	0.7840	|IIIIIIIIIIIIIIIIIIIIIIII
17	3	0.8589	0.7478	|IIIIIIIIIIIIIIIIIIIIII	0.7790	|IIIIIIIIIIIIIIIIIIIIIII
15	3	0.8524	0.7375	|IIIIIIIIIIIIIIIIIIIIII	0.7834	|IIIIIIIIIIIIIIIIIIIIIIII
22	3	0.8226	0.6924	|IIIIIIIIIIIIIIIIIIIII	0.7630	|IIIIIIIIIIIIIIIIIIIIIII
20	3	0.8222	0.6920	|IIIIIIIIIIIIIIIIIIIII	0.7604	|IIIIIIIIIIIIIIIIIIIIIII
16	3	0.8012	0.6617	|IIIIIIIIIIIIIIIIIIII	0.7444	|IIIIIIIIIIIIIIIIIIIIII
14	3	0.7992	0.6593	|IIIIIIIIIIIIIIIIIIII	0.7342	|IIIIIIIIIIIIIIIIIIIIII
13	3	0.3734	0.3393	|IIIIIIIIII	0.1086	|III

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: ID

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.

Squared Memberships: Sum

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.

Squared Memberships: Bar

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: Value

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

Membership Matrix (3 Clusters)

		Cluster Membership Probability		
Row	**Cluster**	**1**	**2**	**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	0.8742	0.0549	0.0709
6	1	0.4205	0.3545	0.2250
7	2	0.0849	0.8188	0.0963
8	2	0.0618	0.8718	0.0664
9	2	0.0629	0.8564	0.0808
10	2	0.0596	0.8746	0.0658
11	2	0.0606	0.8614	0.0780
12	2	0.0733	0.8386	0.0880
13	3	0.3553	0.2713	0.3734
14	3	0.1156	0.0853	0.7992
15	3	0.0787	0.0689	0.8524
16	3	0.0972	0.1017	0.8012
17	3	0.0794	0.0617	0.8589
18	3	0.0424	0.0380	0.9196
19	3	0.0687	0.0714	0.8599
20	3	0.0982	0.0796	0.8222
21	3	0.0696	0.0636	0.8668
22	3	0.0873	0.0902	0.8226

This report displays the membership probability for each row in each cluster.