Chapter 447

Medoid Partitioning

Introduction

The objective of cluster analysis is to partition a set of objects into two or more clusters such that objects within a cluster are similar and objects in different clusters are dissimilar. The medoid partitioning algorithms presented here attempt to accomplish this by finding a set of representative objects called *medoids*. The *medoid* of a cluster is defined as that object for which the average dissimilarity to all other objects in the cluster is minimal. If *k* clusters are desired, *k* medoids are found. Once the medoids are found, the data are classified into the cluster of the nearest medoid.

Two algorithms are available in this procedure to perform the clustering. The first, from Spath (1985), uses random starting cluster configurations. The second, from Kaufman and Rousseeuw (1990), makes special use of silhouette statistics to help determine the appropriate number of clusters. Both of these algorithms will be explained in more detail later.

Dissimilarities

The fundamental value used in cluster analysis is the dissimilarity between two objects. This section discusses how the dissimilarity is computed for the various types of data.

For multivariate data, a critical issue is how the distance between individual variables is combined to form the overall dissimilarity. This depends on the variable type, scaling type, and distance type that is selected.

We begin with a brief discussion of the possible types of variables.

Types of Cluster Variables

Interval Variables

Interval variables are continuous measurements that follow a linear scale. Examples include height, weight, age, price, temperature, and time. These values may be positive or negative.

Ordinal Variables

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).

Ratio Variables

Ratio 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.

Nominal Variables

Nominal variables are those in which the number represents the state of the variable but does not represent magnitude. The number is used for identification purposes only. Examples include gender, race, hair color, city of birth, or zipcode.

Symmetric-Binary Variables

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 and 0 for no, although this is not necessary.

Assymetric-Binary Variables

Asymmetric-binary variables are concerned with the presence or absence of a relatively rare event, the absence of which is rather unimportant and uninformative. For example, if a person has a scar on his face, he might be more easily identified. But if you know the person does not have a scar, that will not help you identify him.

Distance Calculation

The dissimilarity (distance) between two objects is fundamental to cluster analysis since the technique's goal is to place similar objects in the same cluster and dissimilar objects in different clusters. Unfortunately, the measurement of dissimilarity depends on the type of variable. For interval variables, the distance between two objects is simply the difference in their values. However, how do you quantify the difference between males and females? Is it simply 1 - 0 = 1? How do you combine the difference between males and females with the difference in age to form an overall dissimilar? These are the questions that will be answered in this section. This discussion follows Kaufman and Rousseeuw (1990) very closely.

Assume that you have *N* rows (observations) which are separated to be clustered into *K* groups. Each row consists of *P* variables. Two types of distance measures are available in the program: Euclidean and Manhattan.

The *Euclidean distance* d_{jk} between rows *j* and *k* is computed using

$$d_{jk} = \sqrt{\frac{\sum_{i=1}^{P} \delta_{ijk}^2}{P}}$$

and *Manhattan distance* d_{jk} between rows *j* and *k* is computed using

$$d_{jk} = \frac{\sum_{i=1}^{P} \left| \delta_{ijk} \right|}{P}$$

where for interval, ordinal, and ratio variables

$$\delta_{ijk} = z_{ij} - z_{ik}$$

and for asymmetric-binary, symmetric-binary, and nominal variables

$$\delta_{ijk} = \begin{cases} 1 & \text{if } x_{ij} \neq x_{ik} \\ 0 & \text{if } x_{ij} = x_i k \end{cases}$$

with the exception that for asymmetric-binary, the variable is completely ignored (*P* is decreased by one for this row) if both x_{ij} and x_{ik} are equal to zero (the non-rare event).

The value of z_{ij} for interval, ordinal, and ratio variables is defined next.

Interval Variables

You most likely have variables with several different scales. For example, you might have percentages, ages, rates, income levels, and so on. In order to remove distortions due to these differences in scales, the data are transformed to a common scale.

Four types of scaling are available: absolute value, standard deviation, range, and none. Each of these have the general form:

$$z_{ij} = \frac{x_{ij} - A_i}{B_i}$$

where x_{ij} represents the original data value for variable *i* and row *j* and z_{ij} represents the corresponding scale value. The scaling choice determines the values used for A_i and B_i.

The following table shows the scaling mechanism used for each type of scaling.

Type of Scaling	**Value of A_i**	**Value of B_i**
Absolute Value	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\frac{\sum_{j=1}^{N} \lvert x_{ij} - A_i \rvert}{N}$
Standard Deviation	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\sqrt{\frac{\sum_{j=1}^{N} (x_{ij} - A_i)^2}{N-1}}$
Range	$\underset{\text{over } j}{\text{Min}}(x_{ij})$	$\underset{\text{over } j}{\text{Max}}(x_{ij}) - \underset{\text{over } j}{\text{Min}}(x_{ij})$
None	0	1

Ordinal and Ratio Variables

The distance calculations for the ordinal and ratio variables are the same as for interval variables except that the values are transformed to an interval scale before distance calculations begin. The ranks of the ordinal variables and the natural logarithms of the ratio variables are substituted for the actual values. Once these transformations are made, the interval distance formulas are used.

Algorithm Details

Medoid Algorithm of Spath

The first medoid algorithm is presented in Spath (1985). The method minimizes an objective function by swapping objects from one cluster to another. Beginning at a random starting configuration, the algorithm proceeds to a local minimum by intelligently moving objects from one cluster to another. When no object moving would result in a reduction of the objective function, the procedure terminates. Unfortunately, this local minimum is not necessarily the global minimum. To overcome this limitation, the program lets you rerun the algorithm using several random starting configurations and the best solution is kept.

The objective function *D* is the total distance between the objects within a cluster. Mathematically, it is represented as follows:

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij}$$

where *K* is the number of clusters, d_{ij} is the distance between objects *i* and *j*, and C_k is the set of all objects in cluster *k*.

Medoid Algorithm of Kaufman and Rousseeuw

Kaufman and Rousseeuw (1990) present a medoid algorithm which they call PAM (Partition Around Medoids). This algorithm also attempts to minimize the total distance *D* (formula given above) between objects within each cluster. The algorithm proceeds through two phases.

In the first phase, a representative set of *k* objects is found. The first object selected has the shortest distance to all other objects. That is, it is in the center. An addition *k-1* objects are selected one at a time in such a manner that at each step, they decrease *D* as much as possible.

In the second phase, possible alternatives to the *k* objects selected in phase one are considered in an iterative manner. At each step, the algorithm searches the unselected objects for the one that if exchanged with one of the *k* selected objects will lower the objective function the most. The exchange is made and the step is repeated. These iterations continue until no exchanges can be found that will lower the objective function.

Note that all potential swaps are considered and that the algorithm does not depend on the order of the objects on the database.

Silhouettes

Two of the most difficult tasks in cluster analysis are deciding on the appropriate number of clusters and deciding how to tell a bad cluster from a good one. Kaufman and Rousseeuw (1990) define a set of values called *silhouettes* that provide key information about both of these tasks. First, we will explain how these are calculated and then we will show how they are used.

Calculating Silhouettes

A silhouette value *s* is constructed for each object as follows.

1. Consider a particular object *i* which is in cluster *A*. Compute the value

   a = average dissimilarity of *i* to all other objects in *A*

   If *A* contains only one object, set *a* to zero.

2. For every other cluster not equal to *A*, find the cluster *B* that has the smallest average dissimilarity between its objects and *i*. Set

   b = average dissimilarity between *i* and the object in *B*.

   The cluster *B* is the nearest neighbor of object *i*.

3. Compute the silhouette *s* of object *i* as follows:

   If *A* contains only one object, set s = 0.

   If $a < b$, $s = 1 - a/b$.

   If $a > b$, $s = b/a - 1$.

   If $a = b$, $s = 0$.

Interpreting Silhouettes

A silhouette value is constructed for each object. The value can range from minus one to one. It measures how well an object has been classified by comparing its dissimilarity within its cluster to its dissimilarity with its nearest neighbor.

When *s* is close to one, the object is well classified. Its dissimilarity with other objects in its cluster is much less than its dissimilarity with objects in the nearest cluster.

When *s* is near zero, the object was just between clusters *A* and *B*. It was arbitrarily assigned to *A.*

When *s* is close to negative one, the object is poorly classified. Its dissimilarity with other objects in its cluster is much greater than its dissimilarity with objects in the nearest cluster. Why isn't it in the neighboring cluster?

Hence, the silhouette value summarizes how appropriate each object's cluster is.

Determining the Number of Clusters

One useful summary statistic is the average value of *s* across all objects. This summarizes how well the current configuration fits the data. An easy way to select the appropriate number of clusters is to choose that number of clusters which maximizes the average silhouette. We denote the maximum average silhouette across all values of *k* as *SC*.

Kaufman and Rousseeuw (1990) present the following table to aid in the interpretation of *SC*.

SC	Proposed Interpretation
0.71 to 1.00	A strong structure has been found.
0.51 to 0.70	A reasonable structure has been found.
0.26 to 0.50	The structure is weak and could be artificial. Try other methods on this database.
-1 to 0.25	No substantial structure has been found.

Finding Good Clusters

A bar chart of the silhouette values, sorted by cluster number and silhouette value, will show how well each cluster is doing. These charts will be discussed more in the output section.

Further Analysis

Once a cluster analysis has been run and an appropriate solution found, the cluster numbers should be saved to an empty variable so that the cluster solution can be further analyzed. What are some additional procedures that should be run? The most common is a discriminant analysis since it will let you study the impact of each of the variables on the solution. Discriminant analysis will also quantify how well the rows have been clustered. This will show up in the Wilks' lambda statistic.

In addition to discriminant analysis, you will want to produce various scatter plots in which the cluster number is used as a grouping variable. This will greatly increase your understanding of what the clusters that have been found look like.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable. A discussion of the types of variables will be presented shortly.

The data given in the following table contain information on twelve superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBall Dataset (Subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5
Johnson M	81.0	53.0	19.5	7.4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

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:

$$d_{ij} = 1 - |r_{ij}|$$

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.

Example 1 – Medoid Partitioning

This section presents an example of how to run a medoid partitioning analysis. The data used were shown above and are found in the BBall dataset.

Setup

To run this example, complete the following steps:

1 Open the BBall example dataset

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

2 Specify the Medoid Partitioning procedure options

- Find and open the **Medoid Partitioning** 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	**Height-Points,Rebounds**
Clustering Method	**Spath**
Distance Method	**Euclidean**
Scaling Method	**Standard Deviation**
Number of Random Starts	**4**
Random Seed	**5323623** (for reproducibility)
Best Starting Configuration	**Silhouette**
Weighting Method	**Regular**
Reported Clusters	**2**
Label Variable	**Player**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

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

Iteration Detail

Iteration Detail

Number of Clusters	Average Distance		Average Silhouette
	Value	Adjusted	
2	33.900841	5.650140	0.198731
2	34.862052	5.810342	0.170356
2	33.900841	5.650140	0.198731
2	33.446866	5.574478	0.215205
3	21.176572	5.294143	0.005228
3	20.528525	5.132131	0.075998
3	20.528525	5.132131	0.075998
3	20.528525	5.132131	0.075998
4	11.885038	3.961679	0.031361
4	11.885038	3.961679	0.031361
4	11.125179	3.708393	0.109662
4	12.446380	4.148793	0.022789
5	7.731403	3.221418	0.028077
5	9.053752	3.772397	-0.047373
5	7.651509	3.188129	0.023920
5	7.651509	3.188129	0.023920

You should select a number of clusters that minimizes average distance and maximizes average silhouette.

The results of this report may vary from run to run. This report shows the values of the objective functions for each iteration and number of clusters. This report is only generated when the Method option is set to Spath.

The report is especially useful in determining if you have set the number of random starts correctly. If you can see that two or three configurations at the desired number of clusters are identical then you have set the Number Random Starts large enough. Otherwise, you should increase this value and rerun the analysis.

In this example, we will conclude that *k* is two (determined from a later report). However, we notice that we have not achieved the maximum silhouette value (0.215205) more than once. We should change the Number Random Starts options to ten and rerun the analysis.

Average Distance Value

This is the value of the average dissimilarity. It is computed using

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij}$$

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

Adjusted Average Distance

This is the value of the adjusted average dissimilarity. It is computed using

$$D_{adjusted} = \frac{K}{N} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i \in C_k} d_{ij}$$

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.

Iteration Summary

Iteration Summary

Number of Clusters	Average Distance		Average Silhouette
	Value	Adjusted	
2	33.446866	5.574478	0.215205
3	20.528525	5.132131	0.075998
4	11.125179	3.708393	0.109662
5	7.731403	3.221418	0.028077

You should select a number of clusters that minimizes average distance and maximizes average silhouette.

This report shows the values of the objective functions for each number of clusters.

This report is used to determine the appropriate number of clusters. The number selected corresponds to the maximum value of the last (Average Silhouette) column. Usually, the row selected will have a respectable value of the Adjusted Average Distance (this value should be near its minimum).

The definitions of the columns were given above and will not be repeated here.

Cluster Medoids

Cluster Medoids (2 Clusters)

Variable	Cluster Medoid	
	1	2
Height	86	77
Weight	230	210
FgPct	55.9	48.5
FtPct	72.1	83.8
Points	24.6	25.7
Rebounds	11.2	7.5
Row	1 Jabbar K.A	10 Robertson

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

Notice that the players in cluster one are typically nine inches taller and pull down about four more rebounds than the players in cluster two. Apparently, cluster one represents centers (or tall forwards) and cluster two represents other players.

Row Detail

Row Detail

Row	Cluster	Nearest Neighbor	Average Distance		Silhouette	
			Within	Neighbor	Value	Bar
5 Chamberlai	1	2	57.62	75.18	0.2336	|||||||||||
11 Russell B	1	2	53.25	60.77	0.1237	||||||
1 Jabbar K.A	1	2	43.55	46.59	0.0652	|||
8 Johnson M	1	2	47.04	31.60	-0.3282	|
3 Baylor E	1	2	48.86	31.04	-0.3647	|
Cluster Average	1	(5)	50.06	49.04	-0.0541	
12 West J	2	1	26.66	54.72	0.5128	||||||||||||||||||||||||
10 Robertson	2	1	21.49	42.97	0.4999	||||||||||||||||||||||||
2 Barry R	2	1	25.20	46.45	0.4574	||||||||||||||||||||
9 Jordan M	2	1	31.50	52.94	0.4051	|||||||||||||||||
7 Erving J	2	1	23.96	39.78	0.3977	|||||||||||||||||
6 Cousy B	2	1	45.57	67.92	0.3291	|||||||||||||||
4 Bird L	2	1	28.82	38.48	0.2509	||||||||||||
Cluster Average	2	(7)	29.03	49.04	0.4076	
Overall Average		(12)	37.79	49.04	0.2152	= SC

Maximum Distance = 3.012578

This report displays information about each row that was clustered. The report is sorted by Silhouette Value within cluster.

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.

Nearest Neighbor

This is the identification number of the nearest cluster to this row (other than the one that it is in). This information is used in computing the silhouette value.

Average Distance: Within

This is the average distance between this object and all other objects in the cluster. This is the value of *a* in the computation of the silhouette.

Average Distance: Neighbor

This is the average distance between this object and the objects in the nearest neighbor. This is the value of *b* in the computation of the silhouette.

Silhouette: Value

This is the value of the silhouette. Its interpretation was presented in the introduction and will not be repeated here. We note that the value should be positive, and most rows should be greater than 0.50. The fact that several of the rows in this analysis have negative silhouette values would cause us to toss out this cluster configuration and look for a better one.

Silhouette: Bar

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