Overview of Cluster Analysis

1. Cluster Analysis:

"The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering"

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

2. Types of Data in Cluster Analysis

Data matrix (or *object-by-variable structure*): This represents n objects, such as persons, with p **variables** (also called *measurements* or *attributes*), such as age, height, weight, gender, and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects $\times p$ variables):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix (or *object-by-object structure*): This stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

where $d(i, j)$ is the measured **difference** or **dissimilarity** between objects i and j.

2.1 Interval-Scaled Variables

Interval-scaled variables are continuous measurements of a linear scale. Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature

Standardise the values using z-score normalization (see following formulas)

1. Calculate the **mean absolute deviation**, s_f:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|),$$

where $x_{1f}, \ldots, x_{nf}$ are n measurements of f, and m_f is the *mean* value of f, that is, $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$.

2. Calculate the **standardized measurement**, or **z-score**:

$$z_{if} = \frac{x_{if} - m_f}{s_f}.$$

After standardization, or without standardization in certain applications, the dissimilarity (or similarity) between the objects described by interval-scaled variables is typically computed based on the distance between each pair of objects. The most popular distance measure is **Euclidean distance**, which is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2},$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are two n-dimensional data objects.

Another well-known metric is **Manhattan (or city block) distance**, defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

Minkowski distance is a generalization of both Euclidean distance and Manhattan distance. It is defined as

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$$

where p is a positive integer. Such a distance is also called L_p norm, in some literature. It represents the Manhattan distance when $p = 1$ (i.e., L_1 norm) and Euclidean distance when $p = 2$ (i.e., L_2 norm).

If each variable is assigned a weight according to its perceived importance, the **weighted Euclidean distance** can be computed as

$$d(i, j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \cdots + w_m|x_{in} - x_{jn}|^2}.$$

Weighting can also be applied to the Manhattan and Minkowski distances.

2. Binary Variables

A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.

If all binary variables are thought of as having the same weight, we have the 2-by-2 contingency table

where q is the number of variables that equal 1 for both objects i and j,
r is the number of variables that equal 1 for object i but that are 0 for object j,
s is the number of variables that equal 0 for object i but equal 1 for object j,
t is the number of variables that equal 0 for both objects i and j.
The total number of variables is p, where p = q+r +s+t

A contingency table for binary variables.

		object *j*		
		1	0	sum
	1	q	r	$q+r$
object *i*	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

$$d(i, j) = \frac{r+s}{q+r+s+t}.$$

The dissimilarity based on such variables is called asymmetric binary dissimilarity, where the number of negative matches, t, is considered unimportant.

$$d(i, j) = \frac{r+s}{q+r+s}.$$

$$sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j).$$

The coefficient $sim(i, j)$ is called the **Jaccard coefficient**

Dissimilarity between binary variables. Suppose that a patient record table contains the attributes *name, gender, fever, cough, test-1, test-2, test-3*, and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

A relational table where patients are described by binary attributes.

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Mary, Jim) = \frac{1+2}{1+1+2} = 0.75$$

3. Categorical, Ordinal, and Ratio-Scaled Variables

Categorical Variable

A categorical variable is a generalization of the binary variable in that it can take on more than two states.

For example, map color is a categorical variable that may have, say, five states: red, yellow, green, pink, and blue.

Let the number of states of a categorical variable be M.

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

where m is the number of matches (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables.

$$d(i, j) = \frac{p - m}{p}$$

Dissimilarity between categorical variables. Suppose that we have the sample data of Table 7.3, except that only the object-identifier and the variable (or attribute) test-1 are available, where test-1 is categorical.

Table 7.3 A sample data table containing variables of mixed type.

object identifier	*test-1 (categorical)*	*test-2 (ordinal)*	*test-3 (ratio-scaled)*
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Since here we have one categorical variable, test-1, so that d(i, j) evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, we get

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Ordinal Variables

A discrete ordinal variable resembles a categorical variable, except that the M states of the ordinal value are ordered in a meaningful sequence.

y. For example, professional ranks are often enumerated in a sequential order, such as assistant, associate, and full for professors.

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

M_f ordered states

Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$.

Example

There are three states for test-2, namely fair, good, and excellent, that is Mf = 3. For step 1, if we replace each value fortest-2 by its rank, the four objects are assigned the ranks 3, 1, 2, and 3, respectively. Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0. For step 3, we can use, say, the Euclidean distance , which results in the following dissimilarity matrix:

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Ratio-Scaled Variables

A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately following the formula Ae^{Bt} or Ae^{-Bt}

Apply logarithmic transformation to a ratio-scaled variable f having value xi f for object i by using the formula $y_{if} = \log(x_{if})$. The y_{if} values can be treated as interval-valued

Example:

This time, we have the sample data of Table 7.3, except that only the object-identifier and the ratio-scaled variable, test-3, are available. Let's try a logarithmic transformation. Taking the log of test-3 results in the values 2.65, 1.34, 2.21, and 3.08 for the objects 1 to 4, respectively. Using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix:

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$