TYPE OF DATA IN CLUSTERING ANALYSIS

Data structure Data matrix (two modes) object by variable Structure

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

Dissimilarity matrix (one mode) object –by-object structure

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

We describe how object dissimilarity can be computed for object by Interval-scaled variables,

Binary variables, Nominal, ordinal, and ratio variables, Variables of mixed types

Interval-Scaled variables (continuous measurement of a roughly linear scale) Standardize data

Calculate the mean absolute deviation:

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

Where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$$

Calculate the standardized measurement (*z-score*)

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

Using mean absolute deviation is more robust than using standard deviation

Similarity and Dissimilarity Between Objects

Distances are normally used to measure the similarity or dissimilarity between two data objects

Some popular ones include: *Minkowski distance*:

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

- *If* $q = 2$, d is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

Properties

- $d(i,j) \geq 0$
- $d(i,i) = 0$
- $d(i,j) = d(j,i)$
- $d(i,j) \leq d(i,k) + d(k,j)$

Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Binary Variables

A contingency table for binary data

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Distance measure for symmetric binary variables:

$$d(i,j)= \frac{b+c}{a+b+c+d}$$

Distance measure for asymmetric binary variables:

$$d(i,j)= \frac{b+c}{a+b+c}$$

Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j)= \frac{a}{a+b+c}$$

Categorical variables

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green

Method 1: Simple matching

m: # of matches, *p*: total # of variables

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

Method 2: use a large number of binary variables

creating a new binary variable for each of the *M* nominal states

Ordinal Variables

An ordinal variable can be discrete or continuous Order is important, e.g., rank Can be treated like interval-scaled replace x_{if} by their rank

map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable

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

compute the dissimilarity using methods for interval-scaled variables

Ratio-scaled variable:

a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}

Methods:

treat them like interval-scaled variables—*not a good choice!* (why?—the scale can be distorted) apply logarithmic transformation $y_{if} = log(x_{if})$ treat them as continuous ordinal data treat their rank as interval-scaled

Variables of Mixed Types

A database may contain all the six types of variables symmetric binary, asymmetric binary,nominal, ordinal, interval and ratio

One may use a weighted formula to combine their effects vector objects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Vector objects: keywords in documents, gene features in micro-arrays, etc.

Broad applications: information retrieval, biologic taxonomy, etc.

Cosine measure

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{|\vec{X}||\vec{Y}|},$$

A variant: Tanimoto coefficient

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$