4.2 Types of Data in Cluster Analysis

Spring 2010

Instructor: Dr. Masoud Yaghini
Outline

- Data Structures
- Interval-Valued (Numeric) Variables
- Binary Variables
- Categorical Variables
- Ordinal Variables
- Variables of Mixed Types
- References
Data Structures
Clustering algorithms typically operate on either of the following two data structures:

- Data matrix
- Dissimilarity matrix
This represents \( n \) objects, such as persons, with \( p \) variables (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 \( p \) variables)
Dissimilarity matrix

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

- It is often represented by an \( n \)-by-\( n \) where \( d(i, j) \) is the measured difference or dissimilarity between objects \( i \) and \( j \).
- In general, \( d(i, j) \) is a nonnegative number that is
  - close to 0 when objects \( i \) and \( j \) are highly similar or “near” each other
  - becomes larger the more they differ
- Where \( d(i, j) = d(j, i) \), and \( d(i, i) = 0 \)
Type of data in clustering analysis

- Dissimilarity can be computed for
  - Interval-scaled (numeric) variables
  - Binary variables
  - Categorical (nominal) variables
  - Ordinal variables
  - Ratio variables
  - Mixed types variables
Interval-Valued (Numeric) Variables
Interval-valued variables

- **Interval-scaled (numeric) variables** are continuous measurements of a roughly linear scale.

- **Examples**
  - weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.

- The measurement unit used can affect the clustering analysis
  - For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.
Data Standardization

- Expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure.
- To help avoid dependence on the choice of measurement units, the data should be standardized.
- Standardizing measurements attempts to give all variables an equal weight.
- To standardize measurements, one choice is to convert the original measurements to unitless variables.
Data Standardization

- Standardize data
  - Calculate the mean absolute deviation:
    \[ s_f = \frac{1}{n} \left( |x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f| \right) \]
  - where \( m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}) \).
  - Calculate the standardized measurement (z-score)
    \[ z_{if} = \frac{x_{if} - m_f}{s_f} \]
Data Standardization

- Using **mean absolute deviation** is more robust to outliers than using **standard deviation**.
- When computing the **mean absolute deviation**, the deviations from the mean are not squared; hence, the effect of outliers is somewhat reduced.
- Standardization may or may not be useful in a particular application.
  - Thus the choice of whether and how to perform standardization should be left to the user.
- **Methods of standardization** are also discussed under **normalization techniques** for data preprocessing.
Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects described by interval-scaled variables.
Dissimilarity Between Objects

- **Euclidean distance**: the most popular distance measure

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

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

- **Manhattan (city block) distance**: another well-known metric

  \[ d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}| \]
Dissimilarity Between Objects

- **Example**: Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects.

Euclidean distance
\[
= (2^2 + 3^2)^{1/2} = 3.61
\]

Manhattan distance
\[
= 2 + 3 = 5
\]
Dissimilarity Between Objects

- Properties of **Euclidean** and **Manhattan** distances:
  - \(d(i,j) \geq 0\) : Distance is a nonnegative number.
  - \(d(i,i) = 0\) : The distance of an object to itself is 0.
  - \(d(i,j) = d(j,i)\) : Distance is a symmetric function.
  - \(d(i,j) \leq d(i,k) + d(k,j)\) : Going directly from object \(i\) to object \(j\) in space is no more than making a detour over any other object \(h\) (**triangular inequality**).
Dissimilarity Between Objects

- **Minkowski distance**: a generalization of both Euclidean distance and Manhattan distance

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

- Where \( q \) is a positive integer
- It represents the Manhattan distance when \( q = 1 \) and Euclidean distance when \( q = 2 \)
Binary Variables

Types of Data in Cluster Analysis
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.
- Given the variable `smoker` describing a patient,
  - 1 indicates that the patient smokes
  - 0 indicates that the patient does not.
- Treating binary variables as if they are interval-scaled can lead to misleading clustering results.
- Therefore, methods specific to binary data are necessary for computing dissimilarities.
Binary Variables

- One approach involves computing a **dissimilarity matrix** from the given binary data.
- If all binary variables are thought of as having the **same weight**, we have the 2-by-2 **contingency table**
Contingency Table

<table>
<thead>
<tr>
<th>object $i$</th>
<th>1</th>
<th>$q$</th>
<th>$s$</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sum</td>
<td></td>
<td>$q+s$</td>
<td>$r+t$</td>
<td>$p$</td>
</tr>
</tbody>
</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 $j$, and
- $t$ is the number of variables that equal 0 for both objects $i$ and $j$.

$p$ is the total number of variables, $p = q+r+s+t$. 

Types of Data in Cluster Analysis
Symmetric Binary Dissimilarity

- A binary variable is **symmetric** if both of its states are **equally valuable** and carry the same weight
  - Example: the attribute **gender** having the states **male** and **female**.

- Dissimilarity that is based on symmetric binary variables is called **symmetric binary dissimilarity**.

- The dissimilarity between objects i and j:

\[
d(i, j) = \frac{r + s}{q + r + s + t}
\]
Asymmetric Binary Dissimilarity

- A binary variable is **asymmetric** if the outcomes of the states are not equally important,
  - Example: the **positive** and **negative** outcomes of a HIV test.
  - we shall code the most important outcome, which is usually the rarest one, by 1 (HIV positive)
- Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match).
- Therefore, such binary variables are often considered “monary” (as if having one state).
- The dissimilarity based on such variables is called **asymmetric binary dissimilarity**
Asymmetric Binary Dissimilarity

- In asymmetric binary dissimilarity the number of negative matches, $t$, is considered unimportant and thus is ignored in the computation:

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

- The asymmetric binary similarity between the objects i and j, or \( \text{sim}(i, j) \), can be computed as

\[
\text{sim}(i, j) = \frac{q}{q + r + s} = 1 - d(i, j)
\]

- The coefficient \( \text{sim}(i, j) \) is called the Jaccard coefficient.

- When both symmetric and asymmetric binary variables occur in the same data set, the mixed variables approach can be applied (described later).
Example: Dissimilarity Between Binary Variables

- Suppose that a patient record table contains the attributes:
  - **name**: an object identifier
  - **gender**: a symmetric attribute
  - **fever, cough, test-1, test-2, test-3, test-4**: the asymmetric attributes

<table>
<thead>
<tr>
<th></th>
<th>name</th>
<th>gender</th>
<th>fever</th>
<th>cough</th>
<th>test-1</th>
<th>test-2</th>
<th>test-3</th>
<th>test-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>M</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mary</td>
<td>F</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Jim</td>
<td>M</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
Example: Dissimilarity Between Binary Variables

- For asymmetric attribute values
  - let the values Y (yes) and P (positive) be set to 1, and
  - the value N (no or negative) be set to 0.
- Suppose that the distance between objects (patients) is computed based only on the asymmetric variables.
- The distance between each pair of the three patients, Jack, Mary, and Jim, is

\[ d(i, j) = \frac{r + s}{q + r + s} \]

- \[ d(\text{Jack, Mary}) = \frac{0+1}{2+0+1} = 0.33 \]
- \[ d(\text{Jack, Jim}) = \frac{1+1}{1+1+1} = 0.67 \]
- \[ d(\text{Mary, Jim}) = \frac{1+2}{1+1+2} = 0.75 \]
Example: Dissimilarity Between Binary Variables

- These measurements suggest that
  - Mary and Jim are unlikely to have a similar disease because they have the highest dissimilarity value among the three pairs.
  - Of the three patients, Jack and Mary are the most likely to have a similar disease.
Categorical Variables
Categorical Variables

- A categorical (nominal) variable is a generalization of the binary variable in that it can take on more than two states.
  - Example: map_color is a categorical variable that may have five states: red, yellow, green, pink, and blue.
- The states can be denoted by letters, symbols, or a set of integers.
Dissimilarity between categorical variables

- **Method 1: Simple matching**
  - The dissimilarity between two objects $i$ and $j$ can be computed based on the ratio of mismatches:

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

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

- Weights can be assigned to increase the effect of $m$ or to assign greater weight to the matches in variables having a larger number of states.
Example: Dissimilarity between categorical variables

- Suppose that we have the sample data
  - where \texttt{test-1} is categorical.

<table>
<thead>
<tr>
<th>object identifier</th>
<th>test-1 (categorical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>code-A</td>
</tr>
<tr>
<td>2</td>
<td>code-B</td>
</tr>
<tr>
<td>3</td>
<td>code-C</td>
</tr>
<tr>
<td>4</td>
<td>code-A</td>
</tr>
</tbody>
</table>
Categorical Variables

- Let’s compute the dissimilarity matrix

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

- Since here we have one categorical variable, test-1, we set \( p = 1 \) in

\[
d(i, j) = \frac{p - m}{p}
\]
Categorical Variables

- So that $d(i, j)$ evaluates to 0 if objects $i$ and $j$ match, and 1 if the objects differ. Thus,

\[
\begin{bmatrix}
0 \\
1 & 0 \\
1 & 1 & 0 \\
0 & 1 & 1 & 0
\end{bmatrix}
\]
Categorical Variables

- **Method 2: use a large number of binary variables**
  - creating a new asymmetric binary variable for each of the nominal states
  - For an object with a given state value, the binary variable representing that state is set to 1, while the remaining binary variables are set to 0.
  - For example, to encode the categorical variable `map_color`, a binary variable can be created for each of the five colors listed above.
  - For an object having the color `yellow`, the `yellow` variable is set to 1, while the remaining four variables are set to 0.
Ordinal Variables
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.
  - Example: professional ranks are often enumerated in a sequential order, such as assistant, associate, and full for professors.

- Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes.

- The values of an ordinal variable can be mapped to ranks.
  - Example: suppose that an ordinal variable $f$ has $M_f$ states.
  - These ordered states define the ranking $1, \ldots, M_f$.
Suppose that $f$ is a variable from a set of ordinal variables describing $n$ objects.

The dissimilarity computation with respect to $f$ involves the following steps:

Step 1:

- The value of $f$ for the $i$th object is $x_{if}$, and $f$ has $M_f$ ordered states, representing the ranking $1, \ldots, M_f$.
- Replace each $x_{if}$ by its corresponding rank:

$$r_{if} \in \{1, \ldots, M_f\}$$
Ordinal Variables

- **Step 2:**
  - Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto \([0.0, 1.0]\) so that each variable has equal weight.
  - This can be achieved by replacing the rank \(r_{if}\) of the \(i\)th object in the \(f\)th variable by:
    \[
    z_{if} = \frac{r_{if} - 1}{M_f - 1}
    \]

- **Step 3:**
  - Dissimilarity can then be computed using any of the distance measures described for interval-scaled variables.
Ordinal Variables

**Example:** Suppose that we have the sample data:

<table>
<thead>
<tr>
<th>object identifier</th>
<th>test-2 (ordinal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>excellent</td>
</tr>
<tr>
<td>2</td>
<td>fair</td>
</tr>
<tr>
<td>3</td>
<td>good</td>
</tr>
<tr>
<td>4</td>
<td>excellent</td>
</tr>
</tbody>
</table>

There are three states for test-2, namely fair, good, and excellent, that is $M_f = 3$. 
Example: Dissimilarity between ordinal variables

- Step 1: if we replace each value for test-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.
- Step 3: we can use, say, the Euclidean distance, which results in the following dissimilarity matrix:

\[
\begin{bmatrix}
0 & & \\
1 & 0 & \\
0.5 & 0.5 & 0 \\
0 & 1.0 & 0.5 & 0
\end{bmatrix}
\]
Variables of Mixed Types
Variables of Mixed Types

- A database may contain different types of variables
  - interval-scaled, symmetric binary, asymmetric binary, nominal, and ordinal
- We can combine the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval \([0.0, 1.0]\).
Variables of Mixed Types

Suppose that the data set contains $p$ variables of mixed type. The dissimilarity $d(i, j)$ between objects $i$ and $j$ is defined as

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

- $\delta_{ij}^{(f)} = 0$
  - if either (1) $x_{if}$ or $x_{jf}$ is missing (i.e., there is no measurement of variable $f$ for object $i$ or object $j$),
  - or (2) $x_{if} = x_{jf} = 0$ and variable $f$ is asymmetric binary;
- otherwise $\delta_{ij}^{(f)} = 1$
Variables of Mixed Types

- The contribution of variable $f$ to the dissimilarity between $i$ and $j$, that is, $d_{ij}^{(f)}$

- If $f$ is interval-based:
  - use the normalized distance so that the values map to the interval $[0.0, 1.0]$.

- If $f$ is binary or categorical:
  - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

- If $f$ is ordinal:
  - compute ranks $r_{if}$ and
Example: Dissimilarity between variables of mixed type

- The sample data:

<table>
<thead>
<tr>
<th>object</th>
<th>test-1 (categorical)</th>
<th>test-2 (ordinal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>code-A</td>
<td>excellent</td>
</tr>
<tr>
<td>2</td>
<td>code-B</td>
<td>fair</td>
</tr>
<tr>
<td>3</td>
<td>code-C</td>
<td>good</td>
</tr>
<tr>
<td>4</td>
<td>code-A</td>
<td>excellent</td>
</tr>
</tbody>
</table>
Example: Dissimilarity between variables of mixed type

- For test-1 (which is categorical) is the same as outlined above
- For test-2 (which is ordinal) is the same as outlined above
- We can now calculate the dissimilarity matrices for the two variables.

\[
\begin{pmatrix}
0.00 & 0.00 \\
1.00 & 0.00 \\
0.75 & 0.75 & 0.00 \\
0.00 & 1.00 & 0.75 & 0.00
\end{pmatrix}
\]
Example: Dissimilarity between variables of mixed type

- If we go back and look at the data, we can intuitively guess that objects 1 and 4 are the most similar, based on their values for test-1 and test-2.
- This is confirmed by the dissimilarity matrix, where $d(4, 1)$ is the lowest value for any pair of different objects.
- Similarly, the matrix indicates that objects 1 and 2 and object 2 and 4 are the least similar.
References
References

- J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier Inc. (2006). (Chapter 7)
The end