Rohit Jyotiba Chougule, 19200240

Data Mining (COMP40370)

Assignment 8

Question 1. Categorical Variables

Data Table (Data Matrix):

Identifier	T1(Categorical)	T2(Ordinal)	T3(ratio scaled)
01	Code-A	Excellent	445
02	Code-B	Fair	22
03	Code-C	Good	164
04	Code-A	Excellent	1210

1. How many binary variables are needed for attribute T1?

- We have a data matrix, which we will need to create a dissimilarity matrix to identify the number of binary attributes required for the attribute T1(Categorical).
- Categorical attributes in T1 can be encoded using asymmetric binary attributes by using a new binary attribute for each of the m states.

2. Calculate the dissimilarity matrix, showing all the steps of your calculation.

- A dissimilarity matrix stores the proximities for all pairs of n objects in the data table and is represented by n x n table.
- $$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ . & . & . & & \\ d(n,1) & d(n,2) & d(n,3) & . & 0 \end{bmatrix}$$
- The above matrix shows the dissimilarity matrix where *d(i, j)* is the measured dissimilarity or difference between object i and j. If *d(i, j)* is close to 0 the objects are similar and if it is near to 1, they a different.
- We can evaluate the dissimilarity matrix for the categorical attribute T1 as we need to find the binary variables required for T1. From the description of dissimilarity measure for categorical variables from the question, it is the only categorical variable in the data matrix, hence, we can take p=1

  using $d(i,j) = \frac{p-m}{p}$
- The matrix is symmetric; therefore we do not use the *d(j, i)*
- $d(2,1) = \frac{1-0}{1} = 1$ as Code A is not same as Code B
- Similarly, $d(3,1) = \frac{1-0}{1} = 1$ $d(3,2) = \frac{1-0}{1} = 1$
- $d(4,1) = \frac{1-1}{1} = 0$ as Code A, the category of object 1 and 4 are same, we use m=1
- $d(4,2) = \frac{1-0}{1} = 1$ $d(4,3) = \frac{1-0}{1} = 1$
- Hence, after the above calculation, we can create the dissimilarity matrix as below:

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

From the above dissimilarity matrix, we can sense that object 1 and 4 are similar.

Question 2. Ordinal Variables

1. Explain how the ordinal variables are handled.

- The values in an ordinal attribute have meaningful order. For example: slow, medium, fast for a speed attribute. Discretization of numeric attributes into finite categories helps obtaining ordinal attributes. These attributes are ranked.
- The ordinal attributes are handled like numeric or interval scaled attributes when calculating the dissimilarity.

2. Describe briefly the necessary steps for handling this type of variables.

- The necessary steps for handling ordinal attributes are as follows:
  - i. If we consider *j* is an attribute from a set of ordinal attributes by rank *1* to M_{ij}, we can replace x_{ij} by their rank where rank is r_{ij} ∊ {1, … , M_{ij}}
  - ii. Data Normalization by mapping of each variable on to [0.0, 1.0] by replacing i^th object in j^th variable by: $Z_{ij} = \frac{r_{ij}-1}{M_j-1}$
  - iii. The dissimilarity can then is computed using method for interval scaled attributes/numeric attributes.

3. Assume that the Euclidean distance is used as a distance measure. Calculate the dissimilarity matrix for the attribute variable T2.

- From the data table written above, the object identifier and ordinal attribute T2 will be required to calculate the dissimilarity matrix using the steps mentioned in above answer.
- Here, in the data there are 3 unique values (states) for the attribute T2, which are fair, good, excellent. Hence, M_f = 3, considering the ranks our data table becomes:

Identifier	T2	T2(previously)
01	3	Excellent
02	1	Fair
03	2	Good
04	3	Excellent

- The T2 values range from 1 to 3 which can be normalized using $Z_{ij} = \frac{r_{ij}-1}{M_j-1}$

  For object 01: $Z_{13} = \frac{3-1}{3-1} = 1.0$ For Object 02: $Z_{21} = \frac{1-1}{3-1} = 0.0$

  For Object 03: $Z_{32} = \frac{2-1}{3-1} = 0.5$ For Object 04: $Z_{43} = \frac{3-1}{3-1} = 1.0$

- The Euclidean distance can be used as a distance measure to calculate the dissimilarity matrix:

$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}$, here i & j are two objects and 1 to *p* are the different ordinal measures.

$d(2,1) = \sqrt{(0-1)^2} = 1$ $d(3,1) = \sqrt{(0.5-1)^2} = 0.5$ $d(3,2) = \sqrt{(0.5-0)^2} = 0.5$

$d(4,1) = \sqrt{(1-1)^2} = 0$ $d(4,2) = \sqrt{(1-0)^2} = 1$ $d(4,3) = \sqrt{(1-0.5)^2} = 0.5$

The Dissimilarity matrix for the attribute T2 is as follows:

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

From the above dissimilarity matrix, objects 1 & 2 are dissimilar, and even objects 4 & 2 are dissimilar as the distance between them is 1.

Question 3. Ratio-scaled variables

1. Explain how you can handle the dissimilarity between objects of type ratio-scaled.

- A ratio-scaled attribute is a positive measurement on a nonlinear scale. Various methods to handle dissimilarity in such attributes include:
  - i. Treating them as interval scaled or numeric variables, which may not be good, unless normalized.
  - ii. Apply logarithmic transformation: $y_i = \log x_i$
  - iii. Treat x_i as continuous ordinal data and treat its rank as interval-scaled variable.

2. Give the necessary steps for calculating such dissimilarity.

- When calculating such dissimilarity measure using Euclidean distance measure, we need to ensure that the data is normalized.
- Once the data is normalized it can be treated as numeric or interval based for measuring distance.
- The dissimilarity matrix can be then evaluated using the results after the distance is measured.

3. Assume that the distance measure is chosen to be the Euclidian distance. Calculate the dissimilarity matrix for the attribute variable T3.

- We will need to first normalize the data, which we can do using min-max normalization using: $y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$, here $\max(x) = 1210$, $\min(x) = 22$
- For object 1: $y_1 = \frac{445-22}{1210-22} = 0.35$ For object 2: $y_2 = \frac{22-22}{1210-22} = 0.00$
- For object 3: $y_3 = \frac{164-22}{1210-22} = 0.12$ For Object 4: $y_i = \frac{1210-22}{1210-22} = 1.00$

  The new data table or matrix now becomes:

Identifier	T3	T3(previously)
01	0.35	445
02	0.00	22
03	0.12	164
04	1.00	1210

- Now, that the normalized value of the attribute T3 are obtained, Euclidean distance measure can be used to measure the distance between objects.

  $d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}$, here i & j are two objects and 1 to p are the different measures.

  $d(2,1) = \sqrt{(0 - 0.35)^2} = 0.35$ $d(3,1) = \sqrt{(0.12 - 0.35)^2} = 0.23$

  $d(3,2) = \sqrt{(0.12 - 0)^2} = 0.12$ $d(4,1) = \sqrt{(1 - 0.35)^2} = 0.65$

$d(4,2) = \sqrt{(1-0)^2} = 1$ $\qquad\qquad$ $d(4,3) = \sqrt{(1-0.12)^2} = 0.88$

The dissimilarity matrix for the T3 attribute can then be evaluated as:

$$\begin{bmatrix} 0 & & & \\ 0.35 & 0 & & \\ 0.23 & 0.12 & 0 & \\ 0.65 & 1 & 0.88 & 0 \end{bmatrix}$$

From the above dissimilarity matrix, we can infer that the object 2 & 4 are most dissimilar or the farthest values from each other.

Question 4. Mixed type variable

Mixed type approach is used for handling data where object attributes are of mixed type. The above computations were done separately for each of the type of objects.
In real world databases, the objects consist of different type of attributes which are not of same type, wherein, the mixed type approach is used to calculate the dissimilarity between objects.

We combine different attributes into a single dissimilarity matrix on a common scale of [0.0, 1.0]

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

p variables in the dataset

The dissimilarity matrices for each of the type of attribute are:

Attribute 1(Categorical)- T1 $\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$

Attribute 2(Ordinal)- T2 $\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$

Attribute 3(ratio-scaled)- T3 $\begin{bmatrix} 0 & & & \\ 0.35 & 0 & & \\ 0.23 & 0.12 & 0 & \\ 0.65 & 1 & 0.88 & 0 \end{bmatrix}$

In the above given dataset, we do not have any asymmetric binary attribute, hence, $\delta_{ij}^{f} = 1$ for all the computations below:

$$d(2,1) = \frac{1\,(1) + 1(1) + 1(0.35)}{3} = 0.7833$$

$$d(3,1) = \frac{1\,(1) + 1(0.5) + 1(0.23)}{3} = 0.5766$$

$$d(3,2) = \frac{1\,(1) + 1(0.5) + 1(0.12)}{3} = 0.54$$

$$d(4,1) = \frac{1\,(0) + 1(0) + 1(0.65)}{3} = 0.2167$$

$$d(4,2) = \frac{1\,(1) + 1(1) + 1(1)}{3} = 1$$

$$d(4,3) = \frac{1\,(1) + 1(0.5) + 1(0.88)}{3} = 0.7933$$

The result matrix of the mixed type approach is as follows:

$$\begin{bmatrix} 0 & & & \\ 0.7833 & 0 & & \\ 0.5766 & 0.54 & 0 & \\ 0.2167 & 1 & 0.7933 & 0 \end{bmatrix}$$

From the above dissimilarity matrix using mixed approach, we can conclude that when considering all the attributes of the given data matrix/data table, object 1 and 4 are most similar as the value is closest to 0.

Object 2 and 4 are most dissimilar as the value is 1.