COMP40370 Practical 8

Types of Data Attributes

Prof. Tahar Kechadi

Academic year 2019-2020

Please submit your report in PDF format, and make sure it is 5 pages max.

Suppose that we have the sample data given in the table below. As you can see, there are three attributes of myxed types.

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

Question 1: Categorical Variables

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

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

where m is the number of matches and p is the total number of variables. The goal is to calculate the dissimilarity matrix (hint: note that categorical variables can be encoded by asymmetric binary variables).

1. How many binary variables are needed for the attribute variable T1?
2. Calculate the dissimilarity matrix, showing all the steps of your calculation.

Question 2: Ordinal Variables

1. Explain how are the ordinal variables handled.
2. Describe briefly the necessary steps for handling this type of variables.
3. Assume that the Euclidan distance is used as a distance measure. Calculate the dissimilarity matrix for the attribute variable T2.

Question 3: Ratio-scaled Variables

1. Explain how can you handle the dissimilarity between objects of type ratio-scaled.
2. Give the necessary steps for calculating such dissimilarity.
3. Asume the the distance measure is chosen to be the Euclidian distance. Calculate the dissimilarity matrix for the attribute variable T3.

Question 4: Mixed type Variables

A more preferable approach for handling mixed type variables is to process all variable types together. A such technique combines the different variables into a single dissimilarity matrix. For a dataset of p variables of mixed type, the dissmilarity between object i and j is defined by:

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

where:

$$\delta_{ij}^{(f)} = \begin{cases} 0, & \text{if } x_{if} \text{ or } x_{jf} \text{ are missing} \\ 0, & \text{if } x_{if} = x_{jf} = 0 \text{ and } f \text{ is asymmetric binary} \\ 1, & \text{otherwise} \end{cases}$$

The contribution of the variable f to the dissimilarity between i and j, that is $d_{ij}^{(f)}$, is calculated dependent on its type:

- If f is interval based:

  $$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$$

  where h runs over all non-missing objects for variable f.

- If f is binary or caregorical: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise $d_{ij}^{(f)} = 1$
- If f is ordinal: compute the rank r_{if} and $z_{if} = \frac{r_{if}-1}{M_f-1}$, and treat z_{if} as interval-scaled.
- If f is ratio-scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat f as continuous ordinal data, compute r_{if} and z_{if}, and then treat z_{if} as interval-scaled.

Calculate the dissimilarity matrix.