Proximity Measures of Mixed Attributes in Data Mining

Dissimilarity for Attributes of Mixed Types

There are two approaches to compute the dissimilarity between objects of mixed attribute types.

1) One approach is to group each type of attribute together, performing separate data mining (e.g., clustering) analysis for each type. This is feasible if these analyses derive compatible results. However, in real applications, it is unlikely that a separate analysis per attribute type will generate compatible results.
2) A more preferable approach is to process all attribute types together, performing a single analysis. One such technique combines the different attributes into a single dissimilarity matrix, bringing all of the meaningful attributes onto a common scale of the interval [0.0,1.0].

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

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

where $\delta_{ij}^{f}=0$ if

- **i.** x_{if} or x_{jf} is missing,
- **ii.** $x_{if}=x_{jf}=0$ and attribute f is asymmetric binary.

Otherwise, $\delta_{ij}^{(f)}=1$.

$d_{ij}^{(f)}$ depends on type of attribute and can be calculated as with help of following formulas:

(1) if f is **numeric** $d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{max-min}$

(2) if f is **nominal or binary** $d_{ij}^{(f)}=0$ if $x_{if}=x_{jf}$, otherwise $d_{ij}^{(f)}=1$.

(3) if f is **ordinal**, compute the ranks (i.e assign values), r_{if} and $z_{if}=\frac{r_{if}-1}{M_f-1}$ and threat z_{if} as numeric

Example

Compute the dissimilarity between objects of mixed attribute types given in Table 1.

Table 1: A sample data table containing attributes of mixed type.

Object identifier	**test-1 (nominal)**	**test-2 (ordinal)**	**test-3 (numeric)**
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

Solution:

- Dissimilarity Matrix for Attribute **Test-1 (Nominal)**

  Find the dissimilarity matrix between objects for attribute **Test-1** which is of type **nominal** using **formula no 2**. Dissimilarity is 1 if two objects have different value for attribute **Test-1** otherwise 0

  Table 2: Dissimilarity Matrix between objects for attribute **Test-1 (Nominal)**.

Object identifier	**1**	**2**	**3**	**4**
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

- Dissimilarity Matrix for Attribute **Test-2 (Ordinal)**

  Find the dissimilarity matrix between objects for attribute Test-2 which is of type ordinal using formula no 3. Compute the ranks and treat the attribute as numeric

  Step 01: Count number of unique states of Test-2 attribute. Here 3 different unique states (fair, good, excellent) so, $M_f = 3$

  Step 02: Replace each ordinal data of test-2 by Rank. i.e, Fair = 1, Good = 2, and Excellent = 3

  Step 03: Normalize the rank or compute rank with **formula 3** $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

Fair = $z_{1f} = \frac{r_{1f} - 1}{M_f - 1} = \frac{1-1}{3-1} = 0$

Good = $z_{2f}=\frac{r_{2f}-1}{M_f-1}=\frac{2-1}{3-1}=\frac{1}{2}=0.5$

Excellent = $z_{3f}=\frac{r_{3f}-1}{M_f-1}=\frac{3-1}{3-1}=\frac{2}{2}=1$

Now attribute Test-2 becomes:

Object identifier	**test-2 (ordinal)**	**test-2 (ordinal)**
1	excellent	1
2	fair	0
3	good	0.5
4	excellent	1

Find the dissimilarity matrix between objects for attribute **Test-2** which is of type ordinal and converted to Numerical using **formula no 2**. Dissimilarity is find out using Manhattan distance

Table 3: Dissimilarity Matrix between objects for attribute **Test-2 (Ordinal)**.

Object identifier	**1**	**2**	**3**	**4**
1	0			
2	$\lvert 1-0\rvert=1$	0		
3	$\lvert 1-0.5\rvert=0.5$	$\lvert 0-0.5\rvert=0.5$	0	
4	$\lvert 0-0\rvert=0$	$\lvert 0-1\rvert=1$	$\lvert 0.5-1\rvert=0.5$	0

- Dissimilarity Matrix for Attribute **Test-3 (Numerical)**

Find the dissimilarity matrix between objects for attribute Test-3 which is of type numerical using formula no 1. normalize the values so that can be mapped $[0,1]$ using formula

$$d_{ij}^{(f)}=\frac{\lvert x_{if}-x_{jf}\rvert}{max-min}$$

$$d_{1,2}=\frac{\lvert x_{1f}-x_{2f}\rvert}{64-22}=\frac{\lvert 45-22\rvert}{42}=\frac{\lvert 23\rvert}{42}=0.548$$

$$d_{1,3}=\frac{\lvert x_{1f}-x_{3f}\rvert}{64-22}=\frac{\lvert 45-64\rvert}{42}=\frac{\lvert -19\rvert}{42}=0.452$$

$$d_{1,4}=\frac{\lvert x_{1f}-x_{4f}\rvert}{64-22}=\frac{\lvert 45-28\rvert}{42}=\frac{\lvert 17\rvert}{42}=0.405$$

$$d_{2,3}=\frac{\lvert x_{2f}-x_{3f}\rvert}{64-22}=\frac{\lvert 22-64\rvert}{42}=\frac{\lvert -42\rvert}{42}=1$$

Downloaded by Md Ataullha Saim (ataullha00@gmail.com)

$$d_{2,4}=\frac{|x_{2f}-x_{4f}|}{64-22}=\frac{|22-28|}{42}=\frac{|-6|}{42}=0.143$$

$$d_{3,4}=\frac{|x_{3f}-x_{4f}|}{64-22}=\frac{|64-28|}{42}=\frac{|36|}{42}=0.857$$

Table 4: Dissimilarity Matrix between objects for attribute **Test-3 (Numerical)**.

Object identifier	**1**	**2**	**3**	**4**
1	0			
2	0.548	0		
3	0.452	1	0	
4	0.405	0.143	0.857	0

Single dissimilarity matrix between all objects for all attributes

Now, combines the different attributes into a single dissimilarity matrix. The dissimilarity $d(i,j)$ between objects i and j is defined as

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

where $\delta_{ij}^{f}=0$ if

- **iii.** x_{if} or x_{jf} is missing,
- **iv.** $x_{if}=x_{jf}=0$ and attribute f is asymmetric binary.

Otherwise, $\delta_{ij}^{(f)}=1$.

As we can see that there is no any missing value and also no any asymmetric binary attribute, so $\delta_{ij}^{(f)}=1$ for all data points. And $p=3$ as there 3 attributes

Dissimilarity between Object 1 and Object 2

$$d(1,2)=\frac{\sum_{f=1}^{p=3}\delta_{1,2}^{(f)}d_{1,2}^{(f)}}{\sum_{f=1}^{p=3}\delta_{1,2}^{(f)}}=\frac{\left(\delta_{1,2}^{(Test-1)}d_{1,2}^{(Test-1)}\right)+\left(\delta_{1,2}^{(Test-2)}d_{1,2}^{(Test-2)}\right)+\left(\delta_{1,2}^{(Test-3)}d_{1,2}^{(Test-3)}\right)}{\delta_{1,2}^{(Test-1)}+\delta_{1,2}^{(Test-2)}+\delta_{1,2}^{(Test-3)}}$$

$$d(1,2)=\frac{(1*1)+(1*1)+(1*0.548)}{1+1+1}=\frac{2.548}{3}=0.849$$

Downloaded by Md Ataullha Saim (ataullha00@gmail.com)

Dissimilarity between Object 1 and Object 3

$$d(1,3)=\frac{\sum_{f=1}^{p=3}\delta_{1,3}^{(f)}d_{1,3}^{(f)}}{\sum_{f=1}^{p=3}\delta_{1,3}^{(f)}}=\frac{(\delta_{1,3}^{(Test-1)}d_{1,3}^{(Test-1)})+(\delta_{1,3}^{(Test-2)}d_{1,3}^{(Test-2)})+(\delta_{1,3}^{(Test-3)}d_{1,3}^{(Test-3)})}{\delta_{1,3}^{(Test-1)}+\delta_{1,3}^{(Test-2)}+\delta_{1,3}^{(Test-3)}}$$

$$d(1,3)=\frac{(1*1)+(1*0.5)+(1*0.405)}{1+1+1}=\frac{1.905}{3}=0.635$$

Dissimilarity between Object 1 and Object 4

$$d(1,4)=\frac{\sum_{f=1}^{p=3}\delta_{1,4}^{(f)}d_{1,4}^{(f)}}{\sum_{f=1}^{p=3}\delta_{1,2}^{(f)}}=\frac{(\delta_{1,4}^{(Test-1)}d_{1,4}^{(Test-1)})+(\delta_{1,4}^{(Test-2)}d_{1,4}^{(Test-2)})+(\delta_{1,4}^{(Test-3)}d_{1,4}^{(Test-3)})}{\delta_{1,4}^{(Test-1)}+\delta_{1,4}^{(Test-2)}+\delta_{1,4}^{(Test-3)}}$$

$$d(1,4)=\frac{(1*0)+(1*0)+(1*0.405)}{1+1+1}=\frac{0.405}{3}=0.135$$

Dissimilarity between Object 2 and Object 3

$$d(2,3)=\frac{\sum_{f=1}^{p=3}\delta_{2,3}^{(f)}d_{2,3}^{(f)}}{\sum_{f=1}^{p=3}\delta_{2,3}^{(f)}}=\frac{(\delta_{2,3}^{(Test-1)}d_{2,3}^{(Test-1)})+(\delta_{2,3}^{(Test-2)}d_{2,3}^{(Test-2)})+(\delta_{2,3}^{(Test-3)}d_{2,3}^{(Test-3)})}{\delta_{2,3}^{(Test-1)}+\delta_{2,3}^{(Test-2)}+\delta_{2,3}^{(Test-3)}}$$

$$d(2,3)=\frac{(1*1)+(1*0.5)+(1*1)}{1+1+1}=\frac{2.5}{3}=0.8$$

Dissimilarity between Object 2 and Object 4

$$d(2,4)=\frac{\sum_{f=1}^{p=3}\delta_{2,4}^{(f)}d_{2,4}^{(f)}}{\sum_{f=1}^{p=3}\delta_{2,4}^{(f)}}=\frac{(\delta_{2,4}^{(Test-1)}d_{2,4}^{(Test-1)})+(\delta_{2,4}^{(Test-2)}d_{2,4}^{(Test-2)})+(\delta_{2,4}^{(Test-3)}d_{2,4}^{(Test-3)})}{\delta_{2,4}^{(Test-1)}+\delta_{2,4}^{(Test-2)}+\delta_{2,4}^{(Test-3)}}$$

$$d(2,4)=\frac{(1*1)+(1*1)+(1*0.143)}{1+1+1}=\frac{2.143}{3}=0.714$$

Dissimilarity between Object 3 and Object 4

$$d(3,4)=\frac{\sum_{f=1}^{p=3}\delta_{3,4}^{(f)}d_{2,4}^{(f)}}{\sum_{f=1}^{p=3}\delta_{3,4}^{(f)}}=\frac{(\delta_{3,4}^{(Test-1)}d_{3,4}^{(Test-1)})+(\delta_{3,4}^{(Test-2)}d_{3,4}^{(Test-2)})+(\delta_{3,4}^{(Test-3)}d_{3,4}^{(Test-3)})}{\delta_{3,4}^{(Test-1)}+\delta_{3,4}^{(Test-2)}+\delta_{3,4}^{(Test-3)}}$$

Downloaded by Md Ataullha Saim (ataullha00@gmail.com)

$$d(3,4)=\frac{(1*1)+(1*0.5)+(1*0.857)}{1+1+1}=\frac{2.357}{3}=0.952$$

Table 5: Dissimilarity Matrix between objects for attributes (**Test-1, Test-2, Test-3)**.

Object identifier	**1**	**2**	**3**	**4**
1	0			
2	0.849	0		
3	0.635	0.8	0	
4	0.135	0.714	0.952	0

Downloaded by Md Ataullha Saim (ataullha00@gmail.com)