CHAPTER 3

Unit III

Measuring Data Similarity and Dissimilarity

Syllabus Topics

Measuring Data Similarity and Dissimilarity, Proximity Measures for Nominal Attributes and Binary Attributes, interval scaled; Dissimilarity of Numeric Data : Minkowski Distance, Euclidean distance and Manhattan distance; Proximity Measures for Categorical, Ordinal Attributes, Ratio scaled variables; Dissimilarity for Attributes of Mixed Types, Cosine Similarity.

Syllabus Topic : Measuring Data Similarity and Dissimilarity

3.1 Measuring Data Similarity and Dissimilarity

Data Mining Applications such as Clustering, Classification, outlier Analysis needs a way to assess of how alike or unalike are the objects from one another. For this some measures of similarity and dissimilarity are needed given below.

3.1.1 Data Matrix versus Dissimilarity Matrix

- Let us consider a set of n objects with p attributes given by $X_1=(X_{11},X_{12},.... X_{1p})$, $X_2 = (X_{21}, X_{22},.... X_{2p})$ and so on. Where X_{ij} is the value for i^{th} object with j^{th} attribute. These objects can be tuples in a relational database or feature vectors.
- There are mainly two types of data structures for main memory-based clustering algorithms :

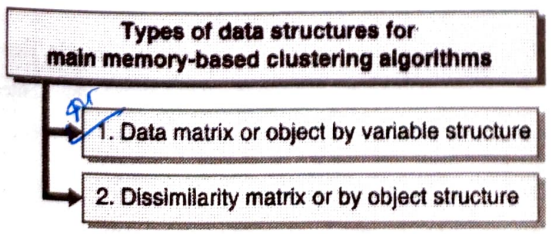


Fig. 3.1.1 : Types of data structures for main memory-based clustering algorithms

➜ **1. Data matrix or object by variable structure**

$$\begin{bmatrix} x_{11} & \dots & x_{11} & \dots & x_{11} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{i1} & \dots & x_{i1} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{n1} & \dots & x_{n1} \end{bmatrix}$$

The Data matrix stores the n data objects in the form of a relational table or in the form of a matrix as shown above.

➜ **2. Dissimilarity matrix or by object structure**

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

- In the above dissimilarity matrix d(i,j) refers to the measure of dissimilarity between objects i and j.

  d(i,j) is close to 0 when the objects i and j are similar.
- The distance d(i,j)==d(j,i), hence not shown as a part of the above matrix as the matrix is symmetric.
- **Similarity** : Similarity in data mining context refers to how much alike two data objects are which can be described by the distance with dimensions representing features of objects where a small distance indicating that the objects are highly similar and a large indicates they are not.

- Similarity can also be expressed as, $\mathbf{sim(i,j) = 1 - d(i,j)}$.
  - **Two mode matrix :** Data Matrix is also called as two mode matrix as it represents two entities objects which are its features.
  - **One mode matrix :** Dissimilarity matrix is called as one mode matrix as it only represents one dimension i.e. the distance.

Syllabus Topic : Proximity Measures for Nominal Attributes and Binary Attributes, Interval Scaled

3.2 Proximity Measures for Nominal Attributes and Binary Attributes, Interval Scaled

3.2.1 Proximity Measures for Nominal Attributes

- Nominal attributes are also called as **Categorical attributes** and allow for only qualitative classification.
- Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.
- The nominal attribute categories can be numbered arbitrarily.
- Arithmetic and logical operations on the nominal data cannot be performed.
- Typical examples of such attributes are :

Car owner :	1. Yes 2. No
Employment status :	1. Unemployed 2. Employed

- Proximity refers to either similarity or dissimilarity. As defined in Section 3.1 calculate similarity and dissimilarity of nominal attributes.
- Dissimilarity is given by,

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

where, p = Total number of attributes describing the objects

and m = Number of matches

- Similarity is given by,

$$sim(i,j) = 1 - d(i,j) = \frac{m}{p}$$

Table 3.2.1

Id	Types of Property
1	Houses
2	Condos
3	co-ops
4	bungalows

- The Table 3.2.1 represents nominal data for an estate agent classifying different types of property. The dissimilarity matrix for the above example can be calculated as follows :

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

- The value in the above matrix is 0 if the objects are similar and it is a 1 if the objects differ.

3.2.2 Proximity Measures for Binary Attributes

- Binary attributes are of two types, symmetric and asymmetric.
- A nominal attribute which has either of the two states 0 or 1 is called Binary attribute , where 0 means that the attribute is absent and 1 means that it is present.

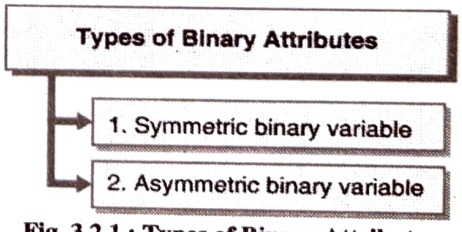


Fig. 3.2.1 : Types of Binary Attributes

➜ **1. Symmetric binary variable**

If both of its states i.e. 0 and 1 are equally valuable. Here we cannot decide which outcome should be 0 and which outcome should be 1.

For example : Marital status of a person is "Married or Unmarried". In this case both are equally valuable and difficult to represent in terms of 0(absent) and 1(present).

2. Asymmetric binary variable

If the outcome of the states are not equally important. An example of such a variable is the presence or absence of a relatively rare attribute.

For example : Person is "handicapped or not handicapped". The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).

- A contingency Table 3.2.2 for binary data :

Table 3.2.2

		Object n		
		1	0	Sum
Object m	1	A	b	a + b
	0	C	d	c + d
	Sum	a + c	b + d	P

- Here we are comparing two objects, object m and object n.
  - (a) would be the number of variables which are present for both objects.
  - (b) would be the number found in object m but not in object n.
  - (c) is just the opposite to b and d is the number that are not found in either object.
- Simple matching coefficient (invariant, if the binary variable is symmetric) as shown in Equation (3.2.1) :

$$d(i, j) = \frac{b + c}{a + b + c + d} \quad \ldots(3.2.1)$$

- Jaccard coefficient (non-invariant if the binary variable is asymmetric) as shown in Equation (3.2.2) :

$$d(i, j) = \frac{b + c}{a + b + c} \quad \ldots(3.2.2)$$

Example

Table 3.2.3 : A Relational table containing mostly binary values

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jai	M	Y	N	P	N	N	N
Raj	F	Y	N	P	N	P	N
Jaya	M	Y	P	N	N	N	N

- Gender is a symmetric attribute the remaining attributes are asymmetric binary.
- Let the values Y and P be set to 1, and the value N be set to 0 as shown in the Table 3.2.4.
- Using Equation (3.2.2) of asymmetric variable.

Table 3.2.4

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jai	M	1	0	1	0	0	0
Raj	F	1	0	1	0	1	0
Jaya	M	1	1	0	0	0	0

- Distance between Jai and Raj (i.e. d(Jai, Raj)) is calculated using Equation (3.2.2) and use contingency Table 3.2.4.

Consider attributes : Fever, cough, Test-1, Test-2, Test-3, Test-4

Consider Jai as object i and Raj as object j

a = Attribute values 1 in Jai and in Raj also = 2

b = Attribute values 1 in Jai but 0 in Raj = 0

c = Attribute values 0 in Jai but 1 in Raj = 1

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

$$d(\text{Jai, Raj}) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

Similarly, calculate distance for other combination

$$d(\text{Jai, Jaya}) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(\text{Jaya, Raj}) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

- So, Jai and Raj are most likely to have a similar disease with lowest dissimilarity value.

3.2.3 Interval Scaled

(SPPU - May 17)

Q. What are interval-scaled variables ? Describe the distance measures that are commonly used for computing the dissimilarity of objects described by such variables. **May 17, 8 Marks**

Interval-scaled attributes are continuous measurement on a linear scale.

- **Example :** weight, height and weather temperature. These attributes allow for ordering, comparing and quantifying the difference between the values. An interval-scaled attributes has values whose differences are interpretable.

- These measures include the Euclidean, Manhattan, and Minkowski distances.

1.	Euclidean L_2	$d_{Euc} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$
2.	City block L_1	$d_{CB} = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
3.	Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$

- The measurement unit can affect the clustering analysis.
- For example, changing measurement units for weight form kilograms to pounds or for height from meters to inches, may lead to a very dissimilar clustering structure. In general, state a variable in minor unit will lead to a larger range for that variable, and thus a larger effect on the resultant clustering structure. To assist avoid belief on the choice of measurement units, the data must be standardized. Standardizing measurements attempts to give all variables an equal weight. This is mainly helpful when given no previous knowledge of the data. However, in some applications, users can intentionally want to grant more weight to a certain set of variables than to others.
- For example, when clustering basketball player candidates, we may favor to give more weight to the variable height.

Syllabus Topic : Dissimilarity of Numeric Data : Minkowski Distance, Euclidean Distance and Manhattan Distance

3.3 Dissimilarity of Numeric Data : Minkowski Distance, Euclidean Distance and Manhattan Distance

Minkowski Distance

- It is used to determine the similarity or dissimilarity between two data objects.

Minkowski distance formula

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

where

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two objects with p number of attributes,

q is a positive integer

Euclidean distance and Manhattan distance

- If $q = 1$, then $d(i,j)$ is Manhattan distance

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

- If q = 2, then d (i,j) is Euclidean distance :

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

- Both the Euclidean distance and Manhattan distance satisfy the following mathematic requirements of a distance function :

$$d(i,j) \geq 0$$

$$d(i,i) = 0$$

- Supremum/Chebyshev (if $q = \infty$)

$$d(i,j) = \max_t \lvert i_t - j_t \rvert$$

- Let us consider the following data :

Customer ID	No. of Trans	Revenue	Tenure(Months)
101	30	1000	20
102	40	400	30
103	35	300	30
104	20	1000	35
105	50	500	1
106	80	100	10
107	10	1000	2

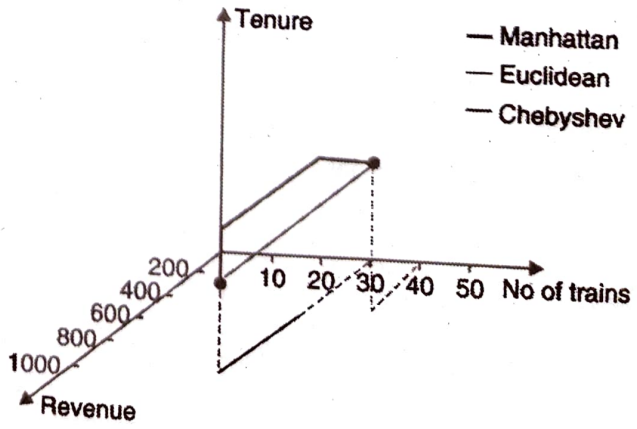


Fig. 3.3.1

$$d_1(cust101, cust102) = |30 - 40| + |11000 - 400| + |20 - 30| = 620$$

$$d_2(cust101, cust102) = \sqrt{(30-40)^2 + (1000-400)^2 + (20-30)^2} \approx 600.16$$

$$d_{max}(cust101, cust102) = |11000 - 400| = 600$$

Syllabus Topic : Proximity Measures for Categorical, Ordinal Attributes, Ratio Scaled Variables

3.4 Proximity Measures for Categorical, Ordinal Attributes, Ratio Scaled Variables

3.4.1 Categorical Attributes

Nominal attributes are also called as Categorical attributes. Described in section 3.2.1

3.4.2 Ordinal Attributes

- A discrete ordinal attribute is a nominal attribute, which have meaningful order or rank for its different states.
- The interval between different states is uneven due to which arithmetic operations are not possible, however logical operations may be applied.
- For example, Considering Age as an ordinal attribute, it can have three different states based on an uneven range of age value. Similarly income can also be considered as an ordinal attribute, which is categorised as low, medium, high based on the income value.
- An ordinal attribute can be discrete or continuous. The ordering of it is important e.g. a rank. These attributes can be treated like interval scaled variables.
- Let us consider f as an ordinal attribute having M_f states. These ordered states define the ranking :

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

- Map the range of each variable onto [0, 1] by replacing i^{th} object in the f^{th} variable by,

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

- Compute the dissimilarity using distance methods discussed in Section 3.2.3.
- Let us consider an example :

Emp Id	Income
1	High
2	Low
3	Medium
4	High

- The three states for the above income variable are low, medium and high, that is $M_f = 3$.
- Next we can replace these values by ranks 3(low), 2(medium) and 1(High).
- We can now normalise the ranking by mapping rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0.
- Next to calculate the distance we can use the Euclidean distance that results in a dissimilarity matrix as :

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

- From the above matrix it can be seen that objects 1 and 2 are most dissimilar so are the object 2 and 4.

3.4.3 Ratio Scaled Attributes

- Ratio scaled attributes are continuous positive measurements on a non linear scale. They are also interval scaled data but are not measured on a linear scale.
- Operations like addition, subtraction can be performed but multiplication and division are not possible.
- **For example :** For instance, if a liquid is at 40 degrees and we add 10 degrees, it will be 50 degrees. However, a liquid at 40 degrees does not have twice the temperature of a liquid at 20 degrees because 0 degrees does not represent "no temperature"

- There are three different ways to handle the ratio-scaled variables :
  - As interval scale variables. The drawback of handling them as interval scaled is that it can distort the results.
  - As continuous ordinal scale.
  - Transforming the data (for example, logarithmic transformation) and then treating the results as interval scaled variables.

3.4.4 Discrete Versus Continuous Attributes

- If an attribute can take any value between two specified values then it is called as **continuous** else it is **discrete**. An attribute will be continuous on one scale and discrete on another.
- **For example :** If we try to measure the amount of water consumed by counting the individual water molecules then it will be discrete else it will be continuous.
  - **Examples of continuous attributes** includes time spent waiting, direction of travel, water consumed etc.
  - **Examples of discrete attributes** includes voltage output of a digital device, a person's age in years.

Syllabus Topic : Dissimilarity for Attributes of Mixed Types

3.5 Dissimilarity for Attributes of Mixed Types

- In many of the applications, objects may be described by a mixture of attribute types.
- In such cases one of the most preferred approach is to combine all the attributes into a single dissimilarity matrix and computing on a common scale of [0.0 , 1.0]
- The dissimilarity may be calculated using

$$d(i,j) = \frac{\sum p_f \delta_{ij}(f)\, d_{ij}(f)}{\sum p_f = 1\ \delta_{ij}(f)}$$

$$\delta_{ij}(f) = 0$$

Where if either

X_{if} or X_{jf} is missing

$X_{if} = X_{jf} = 0$ and attribute f is asymmetric binary

Otherwise

$$\delta_{ij}(f) = 1$$

- The f attribute is computed based on the following :
  - If *f* is binary or nominal:

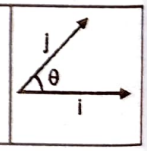
    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - If *f* is interval-based then use the normalized distance.
  - *If f* is ordinal or ratio-scaled then compute ranks r_{if} and treat z_{if} as interval-scaled.

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

Syllabus Topic : Cosine Similarity

3.6 Cosine Similarity

- Cosine similarity is a measure of similarity between two vectors. The data objects are treated as vectors. Similarity is measured as the angle θ between the two vectors. Similarity is 1 when θ = 0, and 0 when θ = 90°.
- Similarity function is given by,

$\cos(i,j) = \frac{i \bullet j}{\|i\| \times \|j\|}$	$i \bullet j = \sum_{k-1}^{n} i_k j_k$	$\|i\| = \sqrt{\sum_{k-1}^{n} i_k^2}$	

Let us consider an example

Given two data objects: x = (3, 2, 0, 5), and y = (1, 0, 0, 0)

Since,

$x \bullet y = 3*1 + 2*0 + 0*0 + 5*0 = 3$

$\|x\| = \text{sqrt}(3^2 + 2^2 + 0^2 + 5^2) \approx 6.16$

$\|y\| = \text{sqrt}(1^2 + 0^2 + 0^2 + 0^2) = 1$

Then, the similarity between

x and y : cos(x, y) = 3/(6.16 * 1) = 0.49

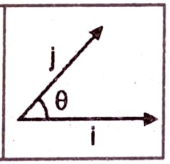
The dissimilarity between x and y : 1 – cos(x,y) = 0.51

Ex. 3.6.1 : Consider the following vectors x and y, x = [1, 1, 1, 1] y = [2, 2, 2, 2], Calculate :
(i) Cosine similarity
(ii) Euclidean distance **May 17, 3 Marks**

Soln. :

(I) Cosine similarity

- Cosine similarity is a measure of similarity between two vectors. The data objects are treated as vectors. Similarity is measured as the angle θ between the two vectors. Similarity is 1 when θ = 0, and 0 when θ = 90°.
- Similarity function is given by,

$\cos(i,j) = \frac{i \bullet j}{\|i\| \times \|j\|}$	$i \bullet j = \sum_{k-1}^{n} i_k j_k$	$\|i\| = \sqrt{\sum_{k-1}^{n} i_k^2}$	

- Let us consider an example

Given two data objects : x = (3, 2, 0, 5) and y = (1, 0, 0, 0)

Since,

$x \bullet y = 3*1 + 2*0 + 0*0 + 5*0 = 3$

$\|x\| = \text{sqrt}(3^2 + 2^2 + 0^2 + 5^2) \approx 6.16$

$\|y\| = \text{sqrt}(1^2 + 0^2 + 0^2 + 0^2) = 1$

Then, the similarity between

x and y : cos(x, y) = 3/(6.16 * 1) = 0.49

The dissimilarity between x and y : 1 – cos(x, y) = 0.51

(II) Euclidean Distance

To find the Euclidean Distance between two points or tuples, the formula is given below.

Let $Y_1 = =\{y_{11}, y_{12}, y_{13}, \ldots\ldots y_{1n}\}$ and $Y_2 = =\{y_{21}, y_{22}, y_{23}, \ldots\ldots y_{2n}\}$

$$\text{distance}(Y_1, Y_2) = \sqrt{\sum_{i=1}^{n} (y_{1i} - y_{2i})^2}$$

Here,

$x_1 = \{1,1,1,1\}$ and $x_2 = \{2,2,2,2\}$

$$\text{distance}(x_1, y_2) = \sqrt{\sum_{i=1}^{n} y_{1i} - y_{2i})^2}$$

$= \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2}$

$= \sqrt{4}$

$= 2$

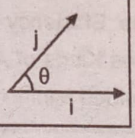
❑❑❑

Ex. 3.6.1 : Consider the following vectors x and y, x = [1, i, i, 1] y = [2, 2, 2, 2], Calculate :
(i) Cosine similarity
(ii) Euclidean distance **May 17, 3 Marks**

Soln. :

(I) Cosine similarity

- Cosine similarity is a measure of similarity between two vectors. The data objects are treated as vectors. Similarity is measured as the angle θ between the two vectors. Similarity is 1 when θ = 0, and 0 when θ = 90°.
- Similarity function is given by,

$\cos(i,j) = \frac{i \bullet j}{\|\|i\|\| \times \|\|j\|\|}$	$i \bullet j = \sum_{k-1}^{n} i_k j_k$	$\|\|i\|\| = \sqrt{\sum_{k-1}^{n} i_k^2}$	

- Let us consider an example

Given two data objects : x = (3, 2, 0, 5) and y = (1, 0, 0, 0)

Since,

$x \bullet y = 3*1 + 2*0 + 0*0 + 5*0 = 3$

$\|x\| = \text{sqrt}(3^2 + 2^2 + 0^2 + 5^2) \approx 6.16$

$\|y\| = \text{sqrt}(1^2 + 0^2 + 0^2 + 0^2) = 1$

Then, the similarity between

x and y : cos(x, y) = 3/(6.16 * 1) = 0.49

The dissimilarity between x and y : 1 – cos(x, y) = 0.51

(II) Euclidean Distance

To find the Euclidean Distance between two points or tuples, the formula is given below.

Let $Y_1 = \{y_{11}, y_{12}, y_{13}, \ldots y_{1n}\}$ and $Y_2 = \{y_{21}, y_{22}, y_{23}, \ldots y_{2n}\}$

$$\text{distance}(Y_1, Y_2) = \sqrt{\sum_{i=1}^{n} (y_{1i} - y_{2i})^2}$$

Here,

$x_1 = \{1,1,1,1\}$ and $x_2 = \{2,2,2,2\}$

$$\text{distance}(x_1, y_2) = \sqrt{\sum_{i=1}^{n} y_{1i} - y_{2i})^2}$$

$= \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2}$

$= \sqrt{4}$

$= 2$

□□□