Symbolic Data Analysis: Dissimilarity/Similarity/Distance Measures (for Clustering)

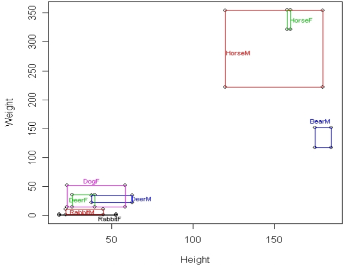
Lynne Billard

Department of Statistics
University of Georgia
lynne@stat.uga.edu

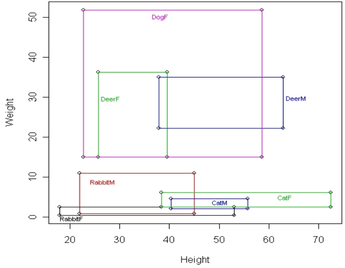
COMPSTAT - August 2010

Consider Veterinary Data (Table 7.5)

ω_u	Animal	Y_1 Height	Y_2 Weight
ω_1	Horse M	[120.0, 180.0]	[222.2, 354.0]
ω_2	Horse F	[158.0, 160.0]	[322.0, 355.0]
ω_3	Bear M	[175.0, 185.0]	[117.2, 152.0]
ω_4	Deer M	[37.9, 62.9]	[22.2, 35.0]
ω_5	Deer F	[25.8, 39.6]	[15.0, 36.2]
ω_6	Dog F	[22.8, 58.6]	[15.0, 51.8]
ω_7	Rabbit M	[22.0, 45.0]	[0.8, 11.0]
ω_8	Rabbit F	[18.0, 53.0]	[0.4, 2.5]
ω_9	Cat M	[40.3, 55.8]	[2.1, 4.5]
ω_{10}	Cat F	[38.4, 72.4]	[2.5, 6.1]



All animals $\omega_u, u = 1, \ldots, 10$

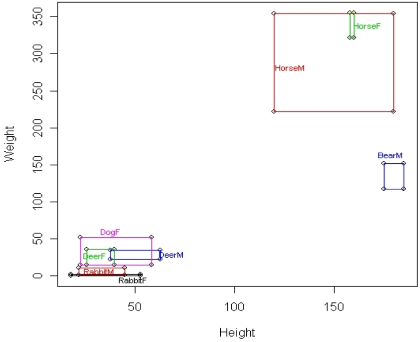


Animals $\omega_u, u = 4, \ldots, 10$

Distance Measures, Similarity/Dissimilarity Matrices:

Goal is to subdivide the complete set of observations E into subsets $P_r = (C_1, \ldots, C_r) \equiv E$ with $\cup C_k = E$, and $C'_k \cap C_k = \phi, k' \neq k$

Mathematically,
use distance measures to produce what we see visually in veterinary data:



Dis/Similarity / Distance Measures

Let the dissimilarity measure between objects a and b be $d(a, b)$, and the corresponding similarity measure be $s(a, b)$.

[Typically, $d(a, b)$ and $s(a, b)$ have reciprocal /inverse relationship, e.g., $d(a, b) = 1s(a, b)$. So, consider $d(a, b)$.]

Definition 7.1: Let a and b be any two objects in E. Then, a dissimilarity measure $d(a, b)$ is a measure that satisfies

(i) $d(a, b) = d(b, a)$;

(ii) $d(a, a) = d(b, b) < d(a, b)$ for all $a \neq b$;

(iii) $d(a, a) = 0$ for all $a \in E$.

Definition 7.2: A distance measure (or metric) is a dissimilarity measure as defined in Definition 7.1 which further satisfies

(iv) $d(a, b) = 0$ implies $a = b$;

(v) $d(a, b) \leq d(a, c) + d(c, b)$ for all $a, b, c \in E$.

Then from property (i), dissimilarity $d(a, b)$ is symmetric, and (v) is the triangle property

Definition 7.3: An ultrametric measure is a distance measure as defined in Definition 7.2 which also satisfies

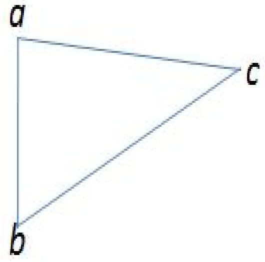
(vi) $d(a, b) \leq Max\{d(a, c), d(c, b)\}$ for all $a, b, c \in E$.

Definition 7.3: An ultrametric measure is a distance measure as defined in Definition 7.2 which also satisfies

(vi) $d(a, b) \leq Max\{d(a, c), d(c, b)\}$ for all $a, b, c \in E$.

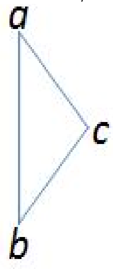
Ultrametrics and hierarchies are in 1-1 correspondence; so need ultrametrics to compare hierarchies.

E.g.,



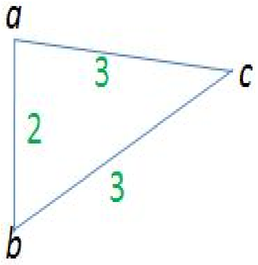
$d(a, b) \leq \max\{d(a, c), d(b, c)\}$
- ultrametric

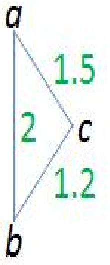
However,



$d(a, b) \geq \max\{d(a, c), d(b, c)\}$
- NOT ultrametric

Definition 7.4: For the collection of objects $a_1, \ldots, a_m \in E$, the dissimilarity matrix (or, distance matrix) is the $m \times m$ matrix D with elements $d(a_i, a_j), i, j = 1, \ldots, m$.





$d(a, b) \leq \max\{d(a, c), d(b, c)\}$
- ultrametric

$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 3 \\ 2 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}$$

$d(a, b) \geq \max\{d(a, c), d(b, c)\}$
- NOT ultrametric

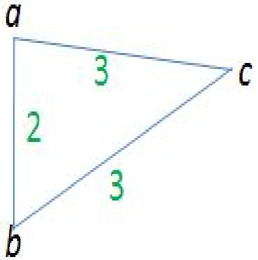
$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 1.5 \\ . & 0 & 1.2 \\ . & . & 0 \end{bmatrix}$$

Notice property (v) $d(a, b) \leq d(a, c) + d(c, b)$ for all a, b, c, holds.

Dis/Similarity / Distance Measures

Definition 7.5: A dissimilarity (or distance) matrix whose elements $d(a, b)$ monotonically increase as they move away from the diagonal (by column and by row) is called a Robinson matrix. (Some use monotonically non-decreasing)

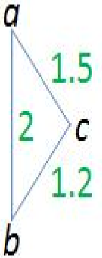
Robinson matrices are in 1-1 correspondence with indexed pyramids.



- ultrametric

$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 3 \\ 2 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}$$

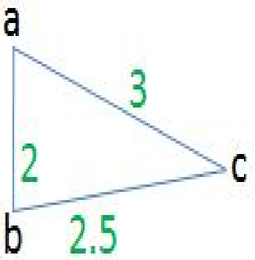
(Not ?) Robinson



- NOT ultrametric

$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 1.5 \\ . & 0 & 1.2 \\ . & . & 0 \end{bmatrix}$$

Not Robinson



- ultrametric

$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 3 \\ . & 0 & 2.5 \\ . & . & 0 \end{bmatrix}$$

Robinson

Definition 7.6: The Cartesian join $A \bigoplus B = (A_1 \bigoplus B_1, \ldots, A_p \bigoplus B_p)$ between two sets A and B is their componentwise union where $A_j \bigoplus B_j =$ "$A_j \cup B_j$". When A and B are multi-valued objects with $A_j = \{a_{j1}, \ldots, a_{js_j}\}$ and $B_j = \{b_{j1}, \ldots, b_{jt_j}\}$, then

$$A_j \bigoplus B_j = \{a_{j1}, \ldots, b_{jt_j}\},\ j = 1, \ldots, p, \tag{7.1}$$

is the set of values in A_j, B_j or both. When A and B are interval-valued objects with $A_j = [a_j^A, b_j^A]$ and $B_j = [a_j^B, b_j^B]$, then

$$A_j \bigoplus B_j = [Min(a_j^A, a_j^B),\ Max(b_j^A, b_j^B)] \tag{7.2}$$

Definition 7.7: The Cartesian meet $A \bigotimes B = (A_1 \bigotimes B_1, \ldots, A_p \bigotimes B_p)$ between two sets A and B is their componentwise intersection where $A_j \bigotimes B_j =$ "$A_j \cap B_j$". When A and B are multi-valued objects, then $A_j \bigotimes B_j$ is the list of possible values from Y_j common to both. When A and B are interval-valued objects forming overlapping interval on Y_j,

$$A_j \bigotimes B_j = [Max(a_j^A, a_j^B), Min(b_j^A, b_j^B)] \tag{7.3}$$

and when $A_j \cap B_j = \phi$, then $A_j \bigotimes B_j = 0$.

E.g.1, multi-valued variables ...
$A = (\{\text{blue, gray, pink, green}\}, \{\text{shirt, dress}\}, \{\text{small, large}\})$
$B = (\{\text{ blue, white}\}, \{\text{shirt, slacks, dress}\}, \{\text{small, medium}\})$

Then, the join is
$A \bigoplus B = (\{\text{blue, gray, pink, green, white}\}, \{\text{shirt, slacks, dress}\}, \{\text{small, medium, large}\})$,
and the meet is
$A \bigotimes B = (\{\text{blue}\}, \{\text{shirt, dress}\}, \{\text{small}\})$.

E.g.2, interval-valued variables ...
$A = ([6, 12], [16, 22])$, $B = ([8, 10], [18, 24])$

Then the join is
$A \bigoplus B = ([6, 12], [16, 24])$,
and the meet is
$A \bigotimes B = ([8, 10], [18, 22])$.

E.g.3, mixed variables (multi- and interval-valued) ...
Let $A = ([6, 12], \{\text{shirt, dress}\})$, $B = ([8, 10], \{\text{shirt, slacks, dress}\})$.

Then, $A \bigoplus B = ([6, 12], \{\text{shirt, slacks, dress}\})$, $A \bigotimes B = ([8, 10], \{\text{shirt, dress}\})$

Multi-valued Variables:
Write observations $\xi(\omega_u)$ as

$$\xi(\omega_u) = (\{Y_{u1k_1}, k_1 = 1, \ldots, k_1^u\}; \ldots; \{Y_{u1k_p}, k_p = 1, \ldots, k_p^u\}). \tag{7.14}$$

Definition 7.15: The Gowda-Diday dissimilarity measure between two multi-valued observations $\xi(\omega_1)$ and $\xi(\omega_2)$ of the form (7.14) is

$$D(\omega_1, \omega_2) = \sum_{j=1}^{p} [D_{1j}(\omega_1, \omega_2) + D_{2j}(\omega_1, \omega_2)]$$

where

$$D_{1j}(\omega_1, \omega_2) = (|k_j^1 - k_j^2|)/k_j, \quad j = 1, \ldots, p, \tag{7.15}$$

$$D_{2j}(\omega_1, \omega_2) = (k_j^1 + k_j^2 - 2k_j^*)/k_j, \quad j = 1, \ldots, p, \tag{7.16}$$

where k_j is the number of values from $\mathcal{Y}_j$ in the join and k_j^* is the number in the meet of $\xi(\omega_1)$ and $\xi(\omega_2)$, respectively.

$D_{1j}(\omega_1, \omega_2)$ is a span distance (relative sizes) component, and
$D_{2j}(\omega_1, \omega_2)$ is a relative content component, of the distance

Write, $D(\omega_1, \omega_2) = \sum_j \phi_j(\omega_1, \omega_2)$

E.g., Color and Habitat of Birds (Table 7.2)
Y_1 = Color, Y_2 = Habitat

ω_u	Species	Y_1 = Color	Y_2 = Habitat
ω_1	species1	{red, black}	{urban, rural}
ω_2	species2	{red}	{urban}
ω_3	species3	{red, black, blue}	{rural}
ω_4	species4	{red, black,blue}	{urban, rural}

Recall $D(\omega_1, \omega_2) = \sum_{j=1}^{p} [D_{1j}(\omega_1, \omega_2) + D_{2j}(\omega_1, \omega_2)] = \sum_j \phi_j(\omega_1, \omega_2)$

$D_{1j}(\omega_1, \omega_2) = (|k_j^1 - k_j^2|)/k_j, \quad D_{2j}(\omega_1, \omega_2) = (k_j^1 + k_j^2 - 2k_j^*)/k_j, \; j = 1, \ldots, p, \;$ (7.14–7.15)

where k_j is the number of values from $\mathcal{Y}_j$ in the join and k_j^* is the number in the meet of $\xi(\omega_1)$ and $\xi(\omega_2)$, respectively, and k_j^u is the number of values from $\mathcal{Y}_j$ in ω_u.

For Y_1 : $D_{11}(\omega_1, \omega_3) = (|2 - 3|)/3 = 1/3$; $D_{21}(\omega_1, \omega_3) = (2 + 3 - 2 \times 2)/3 = 1/3$.

For Y_2 : $D_{12}(\omega_1, \omega_3) = (|2 - 1|)/2 = 1/2$; $D_{22}(\omega_1, \omega_3) = (2 + 1 - 2 \times 1)/2 = 1/2$.

$\phi_1(\omega_1, \omega_3) = D_{11}(\omega_1, \omega_3) + D_{21}(\omega_1, \omega_3) = 1/3 + 1/3 = 2/3$;
$\phi_2(\omega_1, \omega_3) = D_{12}(\omega_1, \omega_3) + D_{22}(\omega_1, \omega_3) = 1/2 + 1/2 = 1$;

$D(\omega_1, \omega_3) = \sum_j \phi_j(\omega_1, \omega_3) = 2/3 + 1 = 5/3$.

The complete table of Gowda-Diday distances, $D(\omega_u, \omega_{u'}) \equiv \phi(\omega_u, \omega_{u'})$:

	Y_1 = Color			Y_2 = Habitat			(Y_1, Y_2)
$(\omega_u, \omega_{u'})$	$D_1(.,.)$	$D_2(.,.)$	$\phi_1(\omega_u, \omega_{u'})$	$D_1(.,.)$	$D_2(.,.)$	$\phi_2(\omega_u, \omega_{u'})$	$\phi(\omega_u, \omega_{u'})$
(ω_1, ω_2)	1/2	1/2	1	1/2	1/2	1	2
(ω_1, ω_3)	1/3	1/3	2/3	1/2	1/2	1	5/3
(ω_1, ω_4)	1/3	1/3	2/3	0	0	0	2/3
(ω_2, ω_3)	2/3	2/3	4/3	0	1	1	7/3
(ω_2, ω_4)	0	2/3	2/3	1/2	1/2	1	5/3
(ω_3, ω_4)	0	0	0	1/2	1/2	1	1

Distance matrix is: $$\mathbf{D} = \begin{bmatrix} 0 & 2 & 5/3 & 2/3 \\ . & 0 & 7/3 & 5/3 \\ . & . & 0 & 1 \\ . & . & . & 0 \end{bmatrix}$$

This is not normalized for scale differences.

To account for scale differences, use $\phi'(\omega_u, \omega_{u'}) = \phi(\omega_u, \omega_{u'})/|\mathcal{Y}|$ where $|\mathcal{Y}|$ is number of possible values from $|\mathcal{Y}|$ covered by E

The complete table of Gowda-Diday distances, $D(\omega_u, \omega_{u'}) \equiv \phi(\omega_u, \omega_{u'})$:

	Y_1 = Color		Y_2 = Habitat		(Y_1, Y_2)	
$(\omega_u, \omega_{u'})$	$\phi_1(.,.)$	$\phi'_1(.,.)$	$\phi_2(.,.)$	$\phi'_2(.,.)$	$\phi(\omega_u, \omega_{u'})$	$\phi'(\omega_u, \omega_{u'})$
(ω_1, ω_2)	1	1/3	1	1/2	2	5/6
(ω_1, ω_3)	2/3	2/9	1	1/2	5/3	13/18
(ω_1, ω_4)	2/3	2/9	0	0	2/3	2/9
(ω_2, ω_3)	4/3	4/9	1	1/2	7/3	17/18
(ω_2, ω_4)	2/3	2/9	1	1/2	5/3	13/18
(ω_3, ω_4)	0	0	1	1/2	1	1/2

$|\mathcal{Y}_1| = 3$ and $|\mathcal{Y}_2| = 2$

Gowda-Diday distance matrix:

Normalized :

$$\mathbf{D}' = \begin{bmatrix} 0 & 5/6 & 13/18 & 2/9 \\ . & 0 & 17/18 & 13/18 \\ . & . & 0 & 1/2 \\ . & . & . & 0 \end{bmatrix}$$

Non-Normalized:

$$\mathbf{D} = \begin{bmatrix} 0 & 2 & 5/3 & 2/3 \\ . & 0 & 7/3 & 5/3 \\ . & . & 0 & 1 \\ . & . & . & 0 \end{bmatrix}$$

Recall observations $\xi(\omega_u)$ written as

$$\xi(\omega_u) = (\{Y_{u1k_1}, k_1 = 1, \ldots, k_1^u\}; \ldots; \{Y_{u1k_p}, k_p = 1, \ldots, k_p^u\}). \qquad (7.14)$$

Definition 7.16: The Ichino-Yaguchi dissimilarity measure between two multi-valued observations $\xi(\omega_1)$ and $\xi(\omega_2)$ of the form of Equation (7.14) for the variable Y_j, $j = 1, \ldots, p$, is

$$\phi_j(\omega_1, \omega_2) = k_j - k_j^* + \gamma(2k_j^* - k_j^1 - k_j^2),\ j = 1, \ldots, p, \qquad (7.17)$$

where k_j is the number of values from $\mathcal{Y}_j$ in the join and k_j^* is the number in the meet of $\xi(\omega_1)$ and $\xi(\omega_2)$, respectively, with k_j^u the number of values from $\mathcal{Y}_j$ in observation ω_u; and where $0 \leq \gamma \leq 0.5$ is a prespecified constant.

For the Bird Data (Table 7.4)

$(\omega_u, \omega_{u'})$	$\phi_j(\omega_u, \omega_{u'})$		Non-Normalized		Normalized†	
	Y_1 = Color	Y_2 = Habitat	$q = 1$	$q = 2$	$q = 1$	$q = 2$
(ω_1, ω_2)	$1 + \gamma(-1)$	$1 + \gamma(-1)$	0.500	0.707	0.208	0.300
(ω_1, ω_3)	$1 + \gamma(-1)$	$1 + \gamma(-1)$	0.500	0.707	0.208	0.300
(ω_1, ω_4)	$1 + \gamma(-1)$	0	0.250	0.500	0.083	0.167
(ω_2, ω_3)	$2 + \gamma(-2)$	$2 + \gamma(-2)$	1.000	1.414	0.417	0.601
(ω_2, ω_4)	$2 + \gamma(-2)$	$1 + \gamma(-1)$	0.750	1.118	0.181	0.417
(ω_3, ω_4)	0	$1 + \gamma(-1)$	0.250	0.500	0.125	0.250

† Normalized by $\mathcal{Y}_j$

Interval-valued data -

$$\xi_u \equiv \xi(\omega_u) = ([a_{uj}, b_{uj}],\ j = 1, \ldots, p), u = 1, \ldots, m$$

Definition 7.18: The Ichino-Yaguchi dissimilarity measure between two interval-valued observations $\xi(\omega_{u_1})$ and $\xi(\omega_{u_2})$ $\xi(\omega_u) = [a_{uj}, b_{uj}]$, $u = 1, ..., m$ for the variable Y_j, $j = 1, \ldots, p$, is

$$\phi_j(\omega_{u_1}, \omega_{u_2}) = |\omega_{u_1 j} \oplus \omega_{u_2 j}| - |\omega_{u_1 j} \otimes \omega_{u_2 j}| + \gamma(2|\omega_{u_1 j} \otimes \omega_{u_2 j}| - |\omega_{u_1 j}| - |\omega_{u_2 j}| \quad (7.27)$$

where $|A|$ is the length of the interval $A = [a, b]$, i.e., $|A| = b - a$, and $0 \leq \gamma \leq 0.5$ is a prespecified constant.

Definition 7.19: The generalized Minkowski distance of order $q \geq 1$ between two interval-valued objects ω_{u_1} and ω_{u_2} is

$$d_q(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^q)^{1/q} \quad (7.28)$$

where $\phi_j(\omega_{u_1}, \omega_{u_2})$ is the Ichino-Yaguchi distance (of Definition 7.18, eqn(7.27)) and w_j^* is an appropriate weight function associated with $Y_j, j = 1, \ldots, p$.

When $q = 1 \rightarrow$ City Block distance
When $q = 2 \rightarrow$ Euclidean distance

Take the first 3 observations only of veterinary data:

ω_u	Animal	Y_1 Height	Y_2 Weight
ω_1	Horse M	[120.0, 180.0]	[222.2, 354.0]
ω_2	Horse F	[158.0, 160.0]	[322.0, 355.0]
ω_3	Bear M	[175.0, 185.0]	[117.2, 152.0]

$$\phi_j(\omega_{u_1}, \omega_{u_2}) = |\omega_{u_1j} \oplus \omega_{u_2j}| - |\omega_{u_1j} \otimes \omega_{u_2j}| + \gamma(2|\omega_{u_1j} \otimes \omega_{u_2j}| - |\omega_{u_1j}| - |\omega_{u_2j}| \tag{7.27}$$

$$A_j \oplus B_j = [Min(a_j^A, a_j^B), Max(b_j^A, b_j^B)] \tag{7.2}$$

$$A_j \otimes B_j = [Max(a_j^A, a_j^B), Min(b_j^A, b_j^B)] \tag{7.3}$$

For $(HorseF, BearM)$ and Y_1,

$$\begin{aligned}\phi_1(\omega_{u_1}, \omega_{u_2}) = & |Min(158, 175), Max(160, 185)| - |Max(158, 175), Min(160, 185)| \\ & + \gamma(2|Max(158, 175), Min(160, 185)| - |160 - 158| - |185 - 175|) \\ = & |158, 185| - |175, 160| + \gamma(2 \times 0 - 2 - 12) \\ = & 27 - 0 + \gamma(2 \times 0 - 12) = 27 + \gamma(-12)\end{aligned}$$

Note, the meet $|175, 160|$ is empty.

For the first 3 observations only of veterinary data:

The complete set of Ichino-Yaguchi Dissimilarity measures is:

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j = 1$	$j = 2$	$\gamma = 1/2$ $j = 1$	$j = 2$
(HorseM, HorseF)	58 + γ(-58)	100.8 + γ(-100.8)	29	50.4
(HorseM, BearM)	60 + γ(-60)	236.8 + γ(-166.6)	30	153.5
(HorseF, BearM)	27 + γ(-12)	237.8 + γ(-67.8)	21	203.9

Definition 7.19: The generalized Minkowski distance of order $q \geq 1$ between two interval-valued objects ω_{u_1} and ω_{u_2} is

$$d_q(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^q)^{1/q} \tag{7.28}$$

where $\phi_j(\omega_{u_1}, \omega_{u_2})$ is the Ichino-Yaguchi distance (of Definition 7.18, eqn(7.27)) and w_j^* is an appropriate weight function associated with $Y_j, j = 1, \ldots, p$.
$q = 1 \rightarrow$ City Block distance $\quad q = 2 \rightarrow$ Euclidean distance

The normalized Euclidean distance of order q between two objects ω_{u_1} and ω_{u_2} is

$$d_2(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^q)^{1/q} \tag{7.30}$$

where $\phi_j(\omega_{u_1}, \omega_{u_2})$ is the Ichino-Yaguchi distance (of Definition 7.18, eqn(7.27)) and w_j^* is an appropriate weight function associated with $Y_j, j = 1, \ldots, p$.

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j = 1$	$j = 2$	$\gamma = 1/2$ $j = 1$	$j = 2$
(HorseM, HorseF)	58 + γ(-58)	100.8 + γ(-100.8)	29	50.4
(HorseM, BearM)	60 + γ(-60)	236.8 + γ(-166.6)	30	153.5
(HorseF, BearM)	27 + γ(-12)	237.8 + γ(-67.8)	21	203.9

$$\phi_j(\omega_{u_1}, \omega_{u_2}) = |\omega_{u_1 j} \oplus \omega_{u_2 j}| - |\omega_{u_1 j} \otimes \omega_{u_2 j}| + \gamma(2|\omega_{u_1 j} \otimes \omega_{u_2 j}| - |\omega_{u_1 j}| - |\omega_{u_2 j}|$$

$$d_2(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^2)^{1/2}, \quad w_j^* = |\mathcal{Y}_j|$$

Unweighted (i.e., $w_j^* = 1$), the normalized Euclidean distance for (HorseF, BearM) is,

$$d_2(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} \omega_j^* [\phi_j(HorseF, BearM)]^2)^{1/2}$$

$$= ((1/2)[(21)^2 + (203.9)^2])^{1/2} = 144.94$$

Weighted (i.e., $w_j^* = \mathcal{Y}_j$), the normalized Euclidean distance for (HorseF, BearM) is,

$$d_2(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} w_j^* \omega_j^* [\phi_j(HorseF, BearM)]^2)^{1/2}$$

$$= ((1/2)[(1/65)(21)^2 + (1/237.8)(203.9)^2])^{1/2} = 144.94$$

Normalized Euclidean distances
using Ichino-Yaguchi Dissimilarity measures is ($\gamma = 1/2$):

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j = 1$	$j = 2$	$d_2(\omega_{u_1}, \omega_{u_2})$ Unweighted	Weighted
(HorseM, HorseF)	29	50.4	41.117	3.437
(HorseM, BearM)	30	153.5	110.594	7.514
(HorseF, BearM)	21	203.9	144.942	9.529

Normalized Euclidean Distance matrix:

$$\mathbf{D}' = \begin{bmatrix} 0 & 41.117 & 110.595 \\ . & 0 & 144.942 \\ . & . & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 0 & 3.437 & 7.514 \\ . & 0 & 9.529 \\ . & . & 0 \end{bmatrix}$$

Unweighted ($w_j^* = 1$)

Weighted ($w_j^* = 1/|\mathcal{Y}_j|$)

Normalized Weighted Euclidean Distance Matrix
using Ichino-Yaguchi Dissimilarity measures is $(\gamma = 1/2)$:

$$\mathbf{D} = \begin{bmatrix} 0 & 2.47 & 5.99 & 11.16 & 11.76 & 11.28 & 12.37 & 12.45 & 12.06 & 11.85 \\ . & 0 & 7.74 & 13.07 & 13.62 & 13.16 & 14.25 & 14.35 & 13.97 & 13.77 \\ . & . & 0 & 8.13 & 9.04 & 8.52 & 9.36 & 9.35 & 8.74 & 8.39 \\ . & . & . & 0 & 0.98 & 0.70 & 1.26 & 1.31 & 0.98 & 0.95 \\ . & . & . & . & 0 & 0.67 & 0.78 & 1.08 & 1.19 & 1.48 \\ . & . & . & . & . & 0 & 1.11 & 1.23 & 1.26 & 1.36 \\ . & . & . & . & . & . & 0 & 0.37 & 0.81 & 1.21 \\ . & . & . & . & . & . & . & 0 & 0.69 & 1.09 \\ . & . & . & . & . & . & . & . & 0 & 0.51 \\ . & . & . & . & . & . & . & . & . & 0 \end{bmatrix}$$

For the first 3 animals **(HorseM, HorseF, BearM)** we had:

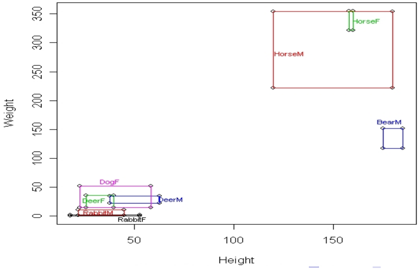
$$\mathbf{D} = \begin{bmatrix} 0 & 3.437 & 7.514 \\ . & 0 & 9.529 \\ . & . & 0 \end{bmatrix}$$

– difference is due to differing weights

Normalized Weighted Euclidean Distance Matrix
using Ichino-Yaguchi Dissimilarity measures is ($\gamma = 1/2$):

D =

										Animal
0	2.47	5.99	11.16	11.76	11.28	12.37	12.45	12.06	11.85	Horse M
.	0	7.74	13.07	13.62	13.16	14.25	14.35	13.97	13.77	HorseF
.	.	0	8.13	9.04	8.52	9.36	9.35	8.74	8.39	BearM
.	.	.	0	0.98	0.70	1.26	1.31	0.98	0.95	DeerM
.	.	.	.	0	0.67	0.78	1.08	1.19	1.48	DeerF
.	.	.	.	.	0	1.11	1.23	1.26	1.36	DogF
.	.	.	.	.	.	0	0.37	0.81	1.21	RabbitM
.	.	.	.	.	.	.	0	0.69	1.09	RabbitF
.	.	.	.	.	.	.	.	0	0.51	CatM
.	.	.	.	.	.	.	.	.	0	CatF



$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j=1$	$j=2$	Euclidean: $d_2(\omega_{u_1}, \omega_{u_2})$ Unweighted	Weighted	City Block: $d_1(\omega_{u_1}, \omega_{u_2})$ Unweighted	Weighted
(HorseM, HorseF)	29	50.4	41.117	3.437	39.70	0.329
(HorseM, BearM)	30	153.5	110.594	7.514	91.75	0.554
(HorseF, BearM)	21	203.9	144.942	9.529	112.45	0.590

Ichino-Yaguchi measures:

$$\phi_j(\omega_{u_1}, \omega_{u_2}) = |\omega_{u_1 j} \oplus \omega_{u_2 j}| - |\omega_{u_1 j} \otimes \omega_{u_2 j}| + \gamma(2|\omega_{u_1 j} \otimes \omega_{u_2 j}| - |\omega_{u_1 j}| - |\omega_{u_2 j}|$$

Normalized weighted Minkowski distance:

$$d_q(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^q)^{1/q}$$

Unweighted: $w_j^* = 1$; Weighted $w_j^* = 1/|\mathcal{Y}_j|$: $w_1^* = 1/65$, $w_2^* = 1/237.8$

City Block: $d_1(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} c_j w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})])$
City Block factor/weight: $c_j = 1/p = 1/2$

Normalized Euclidean: $d_2(\omega_{u_1}, \omega_{u_2}) = ([1/p] \sum_{j=1}^{p} w_j^* [\phi_j(\omega_{u_1}, \omega_{u_2})]^2)^{1/2}$

These are important for Divisive Clustering methodology

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j=1$	$j=2$	Euclidean: $d_2(\omega_{u_1}, \omega_{u_2})$ Unweighted	Weighted	City Block: $d_1(\omega_{u_1}, \omega_{u_2})$ Unweighted	Weighted
(HorseM, HorseF)	29	50.4	41.117	3.437	39.70	0.329
(HorseM, BearM)	30	153.5	110.594	7.514	91.75	0.554
(HorseF, BearM)	21	203.9	144.942	9.529	112.45	0.590

City Block Distance Matrix

Euclidean Distance Matrix

$$\mathbf{D} = \begin{bmatrix} 0 & 39.70 & 91.75 \\ . & 0 & 112.45 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 0.33 & 0.55 \\ . & 0 & 0.59 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 41.12 & 110.59 \\ . & 0 & 144.94 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 0.35 & 0.56 \\ . & 0 & 0.65 \\ . & . & 0 \end{bmatrix}$$

Unweighted Weighted Unweighted Weighted

None appear to be Robinson matrices

However,

$$\mathbf{D} = \begin{bmatrix} 0 & 39.70 & 112.45 \\ . & 0 & 91.75 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 0.33 & 0.59 \\ . & 0 & 0.55 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 41.12 & 144.94 \\ . & 0 & 110.59 \\ . & . & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 0.35 & 0.65 \\ . & 0 & 0.56 \\ . & . & 0 \end{bmatrix}$$

ALL are Robinson matrices

Hausdorff Distances for interval-valued data:

- Hausdorff
- Euclidean Hausdorff
- Normalized Euclidean Hausdorff
- Span Normalized Euclidean Hausdorff

(Important for Divisive Clustering methodology)

Definition 7.20: The Hausdorff distance between two interval-valued objects ω_{u_1} and ω_{u_2}, with $\xi_{uj} = [a_{uj}, b_{uj}]$, $j = 1, \ldots, p$, $u = 1, \ldots, m$, for Y_j, is

$$\phi_j(\omega_{u_1}, \omega_{u_2}) = Max[|a_{u_1 j} - a_{u_2 j}|, |b_{u_1 j} - b_{u_2 j}| \qquad (7.31)$$

Definition 7.21: The Euclidean Hausdorff distance between two interval-valued objects ω_{u_1} and ω_{u_2}, with $\xi_{uj} = [a_{uj}, b_{uj}]$, is

$$d(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p} [\phi_j(\omega_{u_1}, \omega_{u_2})]^2)^{1/2} \qquad (7.32)$$

Definition 7.22: The Normalized Euclidean Hausdorff distance between two interval-valued objects ω_{u_1} and ω_{u_2}, with $\xi_{uj} = [a_{uj}, b_{uj}]$, is

$$d(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p} [\{\phi_j(\omega_{u_1}, \omega_{u_2})\}/H_j]^2)^{1/2} \tag{7.33}$$

$$H_j^2 = (1/[2m^2]) \sum_{u_1=1}^{m} \sum_{u_2=1}^{m} [\phi_j(\omega_{u_1}, \omega_{u_2})]^2 \tag{7.34}$$

The Normalized Euclidean Hausdorff distance is also called a Dispersion Normalization

If the data are classical, then this Normalized Euclidean distance is equivalent to a Euclidean distance on $\mathcal{R}^2$, with H_j corresponding to the standard deviation of Y_j.

Definition 7.23: The Span Normalized Euclidean Hausdorff distance between two interval-valued objects ω_{u_1} and ω_{u_2}, with $\xi_{uj} = [a_{uj}, b_{uj}]$, is

$$d(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p} [\{\phi_j(\omega_{u_1}, \omega_{u_2})\}/|\mathcal{Y}_j|]^2)^{1/2} \tag{7.35}$$

where from (7.26) the span is $|\mathcal{Y}_j| = max_u(b_{uj}) - min_u(a_{uj})$.

This Span Normalization is also called a maximum deviation distance.

ω_u	Animal	Y_1 Height	Y_2 Weight
ω_1	Horse M	[120.0, 180.0]	[222.2, 354.0]
ω_2	Horse F	[158.0, 160.0]	[322.0, 355.0]
ω_3	Bear M	[175.0, 185.0]	[117.2, 152.0]

Hausdorff distance: $\phi_j(\omega_{u_1}, \omega_{u_2}) = Max[|a_{u_1j} - a_{u_2j}|, |b_{u_1j} - b_{u_2j}| \quad (7.31)$

For (HorseF, BearM) and Y_1, we have
$\phi_1(HorseF, BearM) = Max[|158 - 175|, |160 - 185|] = Max[17, 25] = 25$

For (HorseF, BearM) and Y_2, we have
$\phi_2(HorseF, BearM) = Max[|322 - 117.2|, |355 - 152|] = Max[204.8, 203] = 204.8$

Complete set of Hausdorff Distances – (First 3 animals) –

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j = 1$	$j = 2$
(HorseM, HorseF)	38	99.8
(HorseM, BearM)	55	202.0
(HorseF, BearM)	25	204.8

Complete set of Hausdorff Distances – (First 3 animals) –

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j=1$	$j=2$	Euclidean $d(\omega_{u_1}, \omega_{u_2})$	Normalized Euclidean $d^n(\omega_{u_1}, \omega_{u_2})$
(HorseM, HorseF)	38	99.8	106.790	2.653
(HorseM, BearM)	55	202.0	209.354	4.314
(HorseF, BearM)	25	204.8	206.320	3.217

Hausdorff distance: $\phi_j(\omega_{u_1}, \omega_{u_2}) = Max[|a_{u_1j} - a_{u_2j}|, |b_{u_1j} - b_{u_2j}|$ (7.31)

Euclidean Hausdorff distance: $d(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p}[\phi_j(\omega_{u_1}, \omega_{u_2})]^2)^{1/2}$ (7.32)

Normalized Euclidean Hausdorff distance:

$$d^n(\omega_{u_1}, \omega_{u_2}) = (\sum_{j=1}^{p}[\{\phi_j(\omega_{u_1}, \omega_{u_2})\}/H_j]^2)^{1/2}, \quad (7.33)$$

$$H_j^2 = (1/[2m^2]) \sum_{u_1=1}^{m}\sum_{u_2=1}^{m}[\phi_j(\omega_{u_1}, \omega_{u_2})]^2 \quad (7.34)$$

$H_1^2 = (1/[2 \times 3^2])[38^2 + 55^2 + 25^2] = 283 \quad H_1 = 16.823$

$H_2^2 = (1/[2 \times 3^2])[99.8^2 + 202^2 + 204.8^2] = 5150.39; \quad H_2 = 71.766$

For (HorseF, BearM), we have
$d^n(HorseF, BearM) = [(25/16.823)^2 + (204.8/71.766)^2]^{1/2} = 3.217$

Set of Span/Normalized/Euclidean Hausdorff Distances – Veterinary Clinic Data –

$(\omega_{u_1}, \omega_{u_2})$	$\phi_j(\omega_{u_1}, \omega_{u_2})$ $j = 1$	$j = 2$	Euclidean $d(\omega_{u_1}, \omega_{u_2})$	Normalized Euclidean $d^n(\omega_{u_1}, \omega_{u_2})$	SpanNormalized Euclidean $d^s(\omega_{u_1}, \omega_{u_2})$
(HorseM, HorseF)	38	99.8	106.790	2.653	0.720
(HorseM, BearM)	55	202.0	209.354	4.314	1.199
(HorseF, BearM)	25	204.8	206.320	3.217	0.943

Hausdorff distance: $\phi_j(\omega_{u_1}, \omega_{u_2}) = Max[|a_{u_1j} - a_{u_2j}|, |b_{u_1j} - b_{u_2j}|$ (7.31)

Euclidean Hausdorff Distance Matrix D_1:
Normalized Euclidean Hausdorff Distance Matrix D_2:
Span Normalized Euclidean Hausdorff Distance Matrix D_3:

$\mathbf{D_1} = \begin{bmatrix} 0 & 106.790 & 206.354 \\ . & 0 & 206.320 \\ . & . & 0 \end{bmatrix}$ $\mathbf{D_2} = \begin{bmatrix} 0 & 2.653 & 4.314 \\ . & 0 & 3.217 \\ . & . & 0 \end{bmatrix}$ $\mathbf{D_3} = \begin{bmatrix} 0 & 0.720 & 1.199 \\ . & 0 & 0.943 \\ . & . & 0 \end{bmatrix}$

ALL Robinson matrices

Definition 7.17: The Gowda-Diday dissimilarity measure between two interval-valued observations $\xi(\omega_{u_1})$ and $\xi(\omega_{u_2})$ of the form $\xi(\omega_u) = [a_{uj}, b_{uj}]$ is

$$D(\omega_1, \omega_2) = \sum_{j=1}^{p} [D_{j1}(\omega_1, \omega_2) + D_{j2}(\omega_1, \omega_2) + D_{j3}(\omega_1, \omega_2)]$$

where, for $j = 1, \ldots, p$,

$$D_{j1}(\omega_1, \omega_2) = (||b_{u_1j} - a_{u_1j}| - |b_{u_2j} - a_{u_2j}||)/k_j, \quad (7.23)$$
$$D_{j2}(\omega_1, \omega_2) = (|b_{u_1j} - a_{u_1j}| + |b_{u_2j} - a_{u_2j}| - 2I_j)/k_j, \quad (7.24)$$
$$D_{j3}(\omega_1, \omega_2) = (|a_{u_1j} - a_{u_2j}|)/|\mathcal{Y}_j| \quad (7.25)$$

where

$$k_j = |Max(b_{u_1j}, b_{u_2j}), Min(a_{u_1j}, a_{u_2j})|$$
$$I_j = |Max(a_{u_1j}, a_{u_2j}) - Min(b_{u_1j}, b_{u_2j}|$$
$$|\mathcal{Y}_j| = max_u(b_{uj}) - min_u(a_{uj}).$$

Here, k_j is the length of the entire distance spanned by ω_{u_1} and ω_{u_2}, I_j is the length of the intersection of the intervals $[a_{u_1j}, b_{u_1j}]$ and $[a_{u_2j}, b_{u_2j}]$, and $|\mathcal{Y}_j|$ is the total length in $\mathcal{Y}$ covered by observed values of Y_j.
So, $D_{j1}(\omega_1, \omega_2)$ is the span component, $D_{j2}(\omega_1, \omega_2)$ is the relative content component, and $D_{j3}(\omega_1, \omega_2)$ is the relative position component of the distance measure.

Gowda-Diday distances:

$(\omega_{u_1}, \omega_{u_2})$	Y_1 = Height				Y_2 = Weight				(Y_1, Y_2)
	D_{11}	D_{12}	D_{13}	D_1	D_{21}	D_{22}	D_{23}	D_2	D
(HorseM, HorseF)	.967	.967	.584	2.518	.744	.759	.442	1.922	4.440
(HorseM, BearM)	.769	.923	.846	2.538	.409	.703	.021	1.554	4.093
(HorseF, BearM)	.296	.444	.262	1.002	.008	.285	.861	1.154	2.156

$$\mathbf{D} = \begin{bmatrix} 0 & 4.440 & 4.093 \\ . & 0 & 2.156 \\ . & . & 0 \end{bmatrix}$$

Clustering:

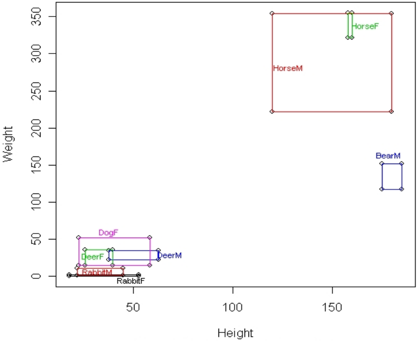
Use the Distance matrices, **D,** calculated from symbolic data in the same way as the Distance matrices, **D,** calculated from classical data are used to

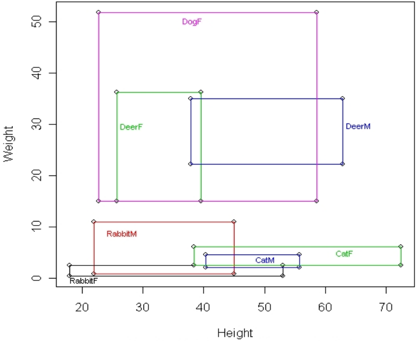
construct

- partitions
- hierarchies
- pyramids

Clustering

E.g., Veterinary dataset –



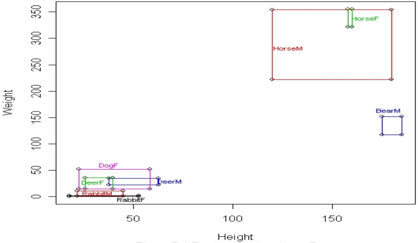


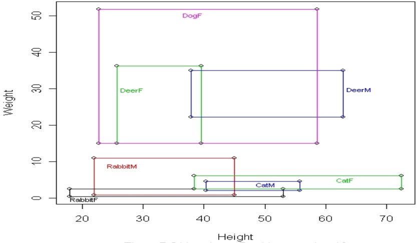
Denote r_{th} partition by $P_r = (C_1, \ldots, C_r)$.

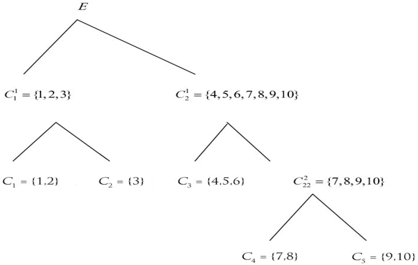
$P_1 = C_1 : \quad E \equiv C_1 = \{1, \ldots, 10\} =$
$\{$HorseM,HorseF,BearM,DeerM,DeerF,DogF,RabbitM,RabbitF,CatM,CatF$\}$
$P_4 = (C_1, \ldots, C_4) : \;\; C_1 = \{1, 2\},\; C_2 = \{3\},\; C_3 = \{4, 5, 6\},\; C_4 = \{7, 8, 9, 10\}$
$P_5 = (C_1, \ldots, C_5) : \;\; C_1 = \{1, 2\},\; C_2 = \{3\},\; C_3 = \{4, 5, 6\},\; C_4 = \{7, 8\},\; C_5 = \{9, 10\}$
OR, $P'_5 = (C_1, \ldots, C_5)$:
$C_1 = \{1, 2\},\; C_2 = \{3\},\; C_3 = \{4, 5, 6\},\; C_4 = \{7, 8\},\; C_5 = \{8, 9, 10\}$
P_5 is a hierarchy; and P'_5 is a pyramid

Veterinary dataset:

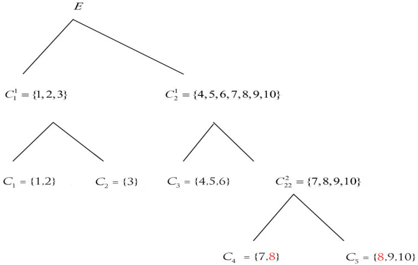
{HorseM,HorseF,BearM,DeerM,DeerF,DogF,RabbitM,RabbitF,CatM,CatF}







Hierarchy



Pyramid