# 8 Análise de agrupamento

## 8.1. Introdução

As análises rudimentares e exploratórias de dados como os procedimentos gráficos auxiliam em geral o entendimento da complexa natureza da análise multivariada. No presente capítulo serão discutidas algumas técnicas gráficas adicionais para agrupar objetos (itens ou variáveis) e também apresentar os algoritmos que devem ser usados para efetivamente realizá-los. Encontrar nos dados uma estrutura natural de agrupamento é uma importante técnica exploratória. A análise de agrupamento deve ser distinguida da análise discriminante, pelo fato desta última ser aplicada a um número de grupos já conhecidos, tendo por objetivo a discriminação de um novo indivíduo a um destes grupos. A análise de agrupamento por sua vez não considera o número de grupos e é realizada com base na similaridade ou dissimilaridade (distâncias).

Objetivo então é de agrupar objetos semelhantes segundo suas características (variáveis). Mas nada impede o agrupamento de variáveis semelhantes segundo os valores obtidos pelos seus objetos. Um outro problema para o qual se

buscará uma resposta é de como verificar se um indivíduo A é mais parecido com B do que com C. Quando o número de variáveis envolvidas é pequeno, a inspeção visual poderá responder. Assim, por exemplo, na Figura 8.1 observa-se uma situação em que A é mais parecido com C do que com B. Intuitivamente para fazer tal inferência usou-se o conceito de distância euclidiana, o qual definiu a idéia de parecença.

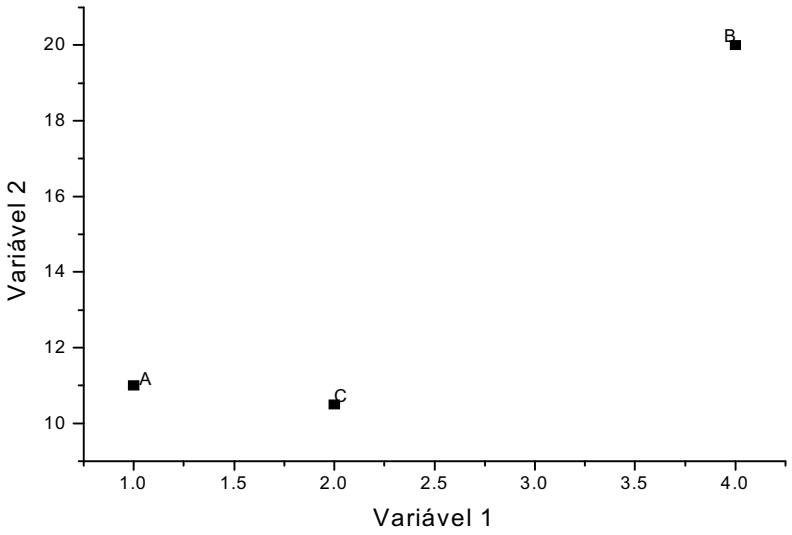


Figura 8.1. Dispersão bidimensional entre três indivíduos.

## 8.2. Medidas de parecença (similaridade e dissimilaridade)

Como foi visto no exemplo da Figura 8.1, é necessário especificar um coeficiente de parecença que indique a proximidade entre os indivíduos. É importante considerar, em todos os casos semelhantes a este, a natureza da variável (discreta, contínua, binária) e a escala de medida (nominal, ordinal, real ou razão).

No capítulo 1, foi discutida a noção de distância, e foi apresentada a distância euclidiana entre dois objetos no espaço p-dimensional. Seja $\underset{\sim}{X}' = \left[ x_1 \; x_2 \cdots x_p \right]$ e $\underset{\sim}{Y}' = \left[ y_1 \; y_2 \cdots y_p \right]$ observações destes objetos, então a distância euclidiana entre eles é dada por:

$$d(\underset{\sim}{X}, \underset{\sim}{Y}) = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2 + ... + \left(x_p - y_p\right)^2} = \sqrt{(x-y)'(x-y)} \qquad (8.1)$$

A distância estatística entre estas duas observações pode ser obtida, e é conhecida como distância de Mahalanobis.

$$d(\underset{\sim}{X}, \underset{\sim}{Y}) = \sqrt{\left(\underset{\sim}{X} - \underset{\sim}{Y}\right)' S^{-1} \left(\underset{\sim}{X} - \underset{\sim}{Y}\right)} \qquad (8.2)$$

em que, $S^{-1}$ é a inversa da matriz de variância e covariância amostral. Outra medida de distância é a matriz métrica de Minkowski, a qual depende de funções modulares.

$$d(\underset{\sim}{X}, \underset{\sim}{Y}) = \left[ \sum_{i=1}^{p} \left| X_i - Y_i \right|^m \right]^{1/m} \qquad (8.3)$$

Para m=2 (8.3) representa a distância euclidiana, e em geral variações de m troca os pesos dados a pequenas e grandes diferenças. Sempre que possível é ideal usar distâncias verdadeiras, ou seja, aquelas que obedecem a desigualdade triangular, no agrupamento de objetos.

Seja $x_{ij}$ as observações do i-ésimo objeto na j-ésima variável, e $x_{kj}$ as observações do k-ésimo objeto na j-ésima variável, e sejam $z_{ij}$ e $z_{kj}$ estes valores padronizados, então podem ser definidas a partir destas notações as seguintes distâncias.

Distância euclidiana média,

$$d_{ik} = \sqrt{\sum_{j=1}^{p}\left(X_{ij} - X_{kj}\right)^2 / p} \qquad (8.4)$$

Distância euclidiana padronizada,

$$d_{ik} = \sqrt{\sum_{j=1}^{p}\left(\frac{X_{ij} - X_{kj}}{\sqrt{S_{jj}}}\right)^2} = \sqrt{\left(\underset{\sim}{X}_i - \underset{\sim}{X}_k\right)' D^{-1}\left(\underset{\sim}{X}_i - \underset{\sim}{X}_k\right)} \qquad (8.5)$$

em que, D é uma matriz diagonal tendo o j-ésimo componente como a variância $S_{jj}$, ou seja,

$$D = \begin{bmatrix} S_{11} & 0 & \cdots & 0 \\ 0 & S_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_{pp} \end{bmatrix}$$

De modo análogo pode-se definir a distância euclidiana padronizada média,

$$d_{ik} = \sqrt{\sum_{j=1}^{p}\left(\frac{X_{ij} - X_{kj}}{\sqrt{S_{jj}}}\right)^2 \Big/ p} = \sqrt{\left(\underset{\sim}{X}_i - \underset{\sim}{X}_k\right)' D^{-1}\left(\underset{\sim}{X}_i - \underset{\sim}{X}_k\right) \Big/ p} \qquad (8.6)$$

Outros tipos de coeficientes podem ser encontrados na literatura (Bussab, Miazaki e Andrade, 1990). Um exemplo é o coeficiente de Gower, baseado na proporção da variação em relação a maior discrepância possível.

$$d_{ik} = -\log_{10}\left[1 - \frac{1}{p}\sum_{j=1}^{p}\left[\frac{\left|X_{ij} - X_{kj}\right|}{X_{maxj} - X_{min.j}}\right]\right] \qquad (8.7)$$

Muitas vezes os objetos não podem ser medidos por variáveis quantitativas, sendo que eles são comparados com base na presença ou ausência de uma determinada característica. A presença e ausência de uma característica podem ser representadas por uma variável binária, a qual assume valor 1 se a característica estiver presente e o valor zero se estiver ausente. Como exemplo consideram-se duas linhagens de milho que foram estereotipadas através de marcadores moleculares denominados RAPD, para as quais o melhorista tinha interesse na similaridade genética dos materiais. Cinco bandas foram utilizadas e o resultado, para presença e ausência das mesmas foram:

| | Bandas | | | | |
|---|---|---|---|---|---|
| Linhagens | 1 | 2 | 3 | 4 | 5 |
| A | 1 | 0 | 0 | 1 | 1 |
| B | 1 | 1 | 0 | 1 | 0 |

Existem neste exemplo duas concordâncias, uma com 1-1 e outra com 2-2 e duas discordâncias. Representando o escore (1 ou 0) da j-ésima variável binária no i-ésimo objeto por $x_{ij}$ e da mesma forma $x_{kj}$ representa o escore do k-ésimo objeto na j-ésima variável, j=1, 2, ..., p. Consequentemente, um determinado desvio ao quadrado terá apenas o valor zero ou 1, da seguinte forma:

$$\left(\mathbf{X}_{ij} - \mathbf{X}_{kj}\right)^2 = \begin{cases} 0 & \text{se } \mathbf{X}_{ij} = \mathbf{X}_{kj} = 1 \text{ ou } \mathbf{X}_{ij} = \mathbf{X}_{kj} = 0 \\ 1 & \mathbf{X}_{ij} \neq \mathbf{X}_{kj} \end{cases} \tag{8.8}$$

Dessa forma a distância euclidiana quadrática representa a contagem do número de pares não coincidentes. Grandes distâncias correspondem a muitas diferenças e, portanto, a objetos dissimilares. Para o exemplo em questão, tem-se:

$$d_{A,B} = 2$$

A equação (8.4) pode ser usada muitas vezes como base para distância, mas muitas vezes sofre por dar aos pares (1-1) e (0-0) o mesmo peso, pois em muitos casos 1-1 é uma forte evidência de similaridade, mas o (0-0) não o é, ou vice versa. Muitos esquemas existem na literatura, dando diferentes tratamentos a este problema. Para introduzir estes conceitos serão apresentados os resultados de coincidências e divergências dos objetos i e k, em uma tabela de contingência.

| | Item k | | |
|---|---|---|---|
| | 1 | 0 | Totais |
| Item i 1 | a | b | a + b |
| Item i 0 | c | d | c + d |
| Totais | a + c | b + d | p = a + b +c + d |

Nesta Tabela, a representa a freqüência de coincidências (1-1), b a freqüência de (1-0), e assim sucessivamente. No exemplo em questão a=2, b=c=d=1.

A Tabela 8.1 apresenta alguns dos coeficientes de semelhança em termos das freqüências apresentadas anteriormente, para variáveis binárias. Os valores para o exemplo, a variação de cada uma, o nome comum na literatura e explicação racional para as mesmas foram apresentados.

Na Tabela 8.1, estão apresentados os coeficientes de similaridades, com exceção da distância binária de Sokal. Muitas vezes podemos transformar, as medidas de dissimilaridade em similaridade pela relação apresentada em Johnson e Wichern (1988), por exemplo.

$$s_{ik} = \frac{1}{1 + d_{ik}} \qquad (8.9)$$

Outras formas de se obter coeficientes de similaridades a partir da distância euclidiana com variáveis padronizadas, pode ser obtida pelo coeficiente de Cattel (Bussab, Miazaki, Andrade, 1990).

$$s_{i\,k} = \frac{2\left(p - \frac{2}{3}\right) - d_{ik}^2}{2\left(p - \frac{2}{3}\right) + d_{ik}^2} \qquad (8.10)$$

Uma outra derivada é a de Cattel e Coulter, apresentada a seguir.

$$s_{i\,k} = \frac{\sqrt{2p} - d_{ik}^2}{\sqrt{2p} + d_{ik}^2} \qquad (8.11)$$

No entanto, nem sempre é possível construir distâncias a partir de similaridades. Isso só pode ser feito se a matriz de similaridades for não negativa definida. Com a condição de que $s_{i\,i} = 1$, máximo das similaridades, e que a matriz de similaridades é não negativa definida, então (8.12) tem as propriedades de distância.

$$d_{i\,k} = \sqrt{2(1 - s_{i\,k})} \qquad (8.12)$$

Tabela 8.1. Alguns coeficientes de parecença para variáveis dicotômicas.

| Nome | Expressão | Explicação | Variação | Ex. |
|---|---|---|---|---|
| Coincidência simples | $\frac{a+d}{p}$ | Pesos iguais para 1-1 e 0-0 | 0-1 | 0,60 |
| Sokal e Sneath | $\frac{2(a+d)}{2(a+d)+b+c}$ | Peso duplo para 1-1 e 0-0 | 0-1 | 0,75 |
| Rogers e Tanimoto | $\frac{a+d}{a+2(b+c)+d}$ | Duplo peso para pares não coincidentes | 0-1 | 0,43 |
| Russel e Rao | $\frac{a}{p}$ | Nenhum 0-0 no numerador | 0-1 | 0,40 |
| Jaccard | $\frac{a}{a+b+c}$ | As coincidências 0-0 são tratadas como irrelevantes | 0-1 | 0,50 |
| Sorenson | $\frac{2a}{2a+b+c}$ | 0-0 é irrelevante e duplo peso para 1-1. | 0-1 | 0,66 |
| - | $\frac{a}{a+2(b+c)}$ | 0-0 é irrelevante e duplo peso para não coincidência. | 0-1 | 0,33 |
| - | $\frac{a}{b+c}$ | Razão entre coincidências e não coincidências - Exceto 0-0 | 0-(p-1) | 1,00 |
| Dist. Binária de Sokal | $\sqrt{\frac{b+c}{p}}$ | Única medida de dissimilaridade. | 0-1 | 0,63 |
| Ochiai | $\frac{a}{\sqrt{(a+b)(a+c)}}$ | Concordâncias positivas sobre adaptação da média geométrica de discordâncias | 0-1 | 0,67 |
| Baroni-Urbani-Buser | $\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$ | Concordâncias positivas e a média geom. de concordância positivas e negativas | 0-1 | 0,63 |
| Haman | $\frac{(a+d)-(b+c)}{p}$ | Proporção de coincidências menos a proporção de discordâncias | -1 - +1 | 0,20 |
| Yule | $\frac{ad-bc}{ad+bc}$ | proporção de ad menos a de bc | -1 - +1 | 0,33 |
| $\phi$ | $\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ | Produto de momento de correlação aplicado a variáveis binárias | -1 - +1 | 0,17 |
| Ochiai II | $\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ | Proporção de coincidências em relação à média geom. Total modificada | 0 -1 | 0,33 |

Em algumas aplicações é necessário agrupar variáveis ao invés de objetos. As medidas de similaridades para agrupar variáveis usadas na prática são baseadas nos coeficientes de correlação amostral. Em algumas aplicações de agrupamento, as correlações negativas são trocadas pelos seus valores absolutos. Quando, as variáveis são binárias esta correlação está apresentada na Tabela 8.1 ($\phi$). Este coeficiente de correlação está associado à estatística de qui-quadrado, para testar a independência de duas variáveis categóricas por ($\phi^2 = \chi^2/n$, $n = a + b + c + d$, $\chi^2$ com 1 grau de liberdade). Para n fixo, uma grande similaridade (ou correlação) é consistente com a falta de independência entre as variáveis. Uma outra importante observação que pode ser feita é que para agrupamento de variáveis os coeficientes de similaridade e de distâncias podem ser usadas, apenas tomando-se o cuidado de substituir p (número de variáveis) por n (número de objetos).

## 8.3. Agrupamentos

Muitos algoritmos existem para formar os agrupamentos, devido a existência de vários critérios existentes para conceituar os grupos que nem sempre são aceitos universalmente. Uma outra razão para isso, é que raramente pode-se examinar todas as possibilidades de agrupamento, mesmos com os mais rápidos e possantes computadores.

Serão divididas e apresentadas neste material as técnicas de agrupamentos denominadas hierárquicas das não hierárquicas.

### 8.3.1. Agrupamentos hierárquicos

Os agrupamentos hierárquicos são realizados por sucessivas fusões ou por sucessivas divisões. Os métodos hierárquicos aglomerativos iniciam com tantos grupos quanto aos objetos, ou seja, cada objeto forma um agrupamento. Inicialmente, os objetos mais similares são agrupados e fundidos formando um único grupo. Eventualmente o processo é repetido, e com o decréscimo da similaridade, todos os subgrupos são fundidos, formando um único grupo com todos os objetos.

Os métodos hierárquicos divisivos trabalham na direção oposta. Um único subgrupo inicial existe com todos os objetos e estes são subdivididos em dois subgrupos de tal forma que exista o máximo de semelhança entre os objetos dos mesmos subgrupos e a máxima dissimilaridade entre elementos de subgrupos distintos. Estes subgrupos são posteriormente subdivididos em outros subgrupos dissimilares. O processo é repetido até que haja tantos subgrupos quanto objetos.

Os resultados finais destes agrupamentos podem ser apresentados por gráficos denominados dendrogramas. Os dendrogramas apresentam os elementos e os respectivos pontos de fusão ou divisão dos grupos formados em cada estágio.

Os esforços deste capítulo serão concentrados nos métodos hierárquicos aglomerativos (“Linkage Methods”). Serão discutidos os métodos de ligação simples (mínima distância ou vizinho mais próximo), ligação completa (máxima distância ou vizinho mais distante) e ligação média (distância média). As idéias para estes três processos estão, esquematicamente, apresentados na Figura 8.2.

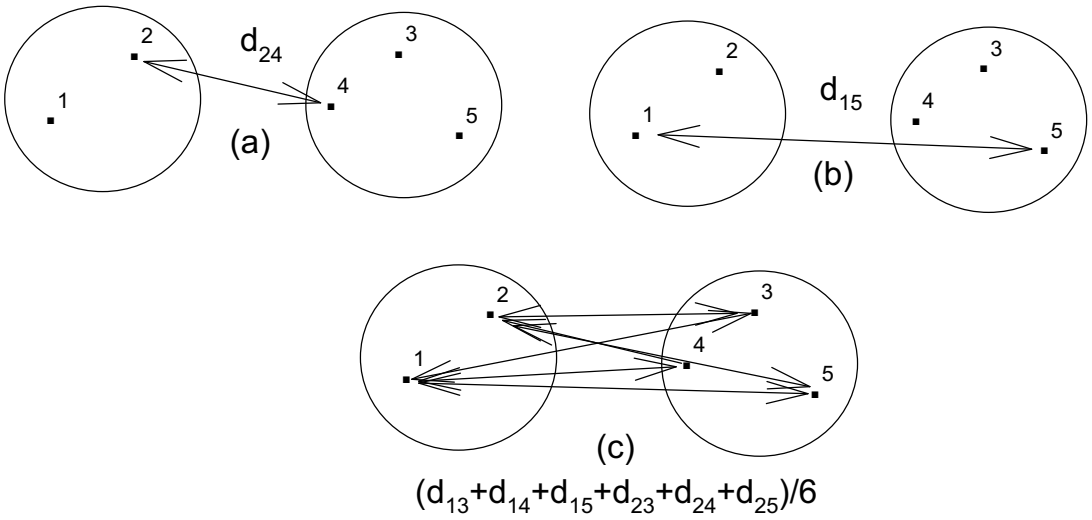


Figura 8.2. Distâncias entre os grupos para os métodos da (a) ligação simples, (b) ligação completa e (c) ligação média.

A seguir está apresentado um algoritmo geral para os agrupamentos hierárquicos aglomerativos com n objetos (itens ou variáveis).

1. Iniciar com n grupos, cada um com um único elemento e com uma matriz simétrica n x n de dissimilaridades (distâncias) $D=\{d_{ik}\}$.
2. Buscar na matriz D o par de grupos mais similar (menor distância) e faça a distância entre os grupos mais similares U e V igual a $d_{uv}$.
3. Fundir os grupos U e V e nomeá-lo por (UV). Recalcular e rearranjar as distâncias na matriz D (a) eliminando as linhas e colunas correspondentes a U e V e (b) acrescentando uma linha e coluna com as distâncias entre o grupo (UV) e os demais grupos.

4. Repetir os passos 2 e 3 num total de (n-1) vezes (todos os objetos estarão em único grupo). Anotar a identidade dos grupos que vão sendo fundidos e os respectivos níveis (distâncias) nas quais isto ocorre.

(a) Ligação simples (vizinho mais próximo)

Para exemplificar é considerado um exemplo, no qual destacam-se 4 objetos (A, B, C, D), e para o qual a matriz de distâncias entre os objetos é apresentada a seguir.

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Para ilustrar o método da ligação simples, os objetos menos distantes devem, inicialmente, ser fundidos. Então, $\min_{i,k}\{d_{ik}\}$ =$d_{AB}$=3. O próximo passo é fundir A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes. As distâncias dos vizinhos mais próximos são,

$$d_{(AB),C} = \min\{d_{AC}, d_{BC}\} = \min\{7, 9\} = 7$$

$$d_{(AB),D} = \min\{d_{AD}, d_{BD}\} = \min\{8, 6\} = 6$$

A nova matriz D para o próximo passo é:

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

A menor distância é entre D e C, com $d_{DC}$=5, os quais foram fundidos formando o subgrupo DC, no nível 5. Recalculando as distâncias têm-se,

$$d_{(DC),(AB)} = \min\{d_{D(AB)}, d_{C(AB)}\} = \min\{6, 7\} = 6$$

A nova matriz D fica,

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

Conseqüentemente o grupo DC é fundido com AB na distância 6. Na Figura 8.3, foi apresentado o dendrograma, com os resultados alcançados.

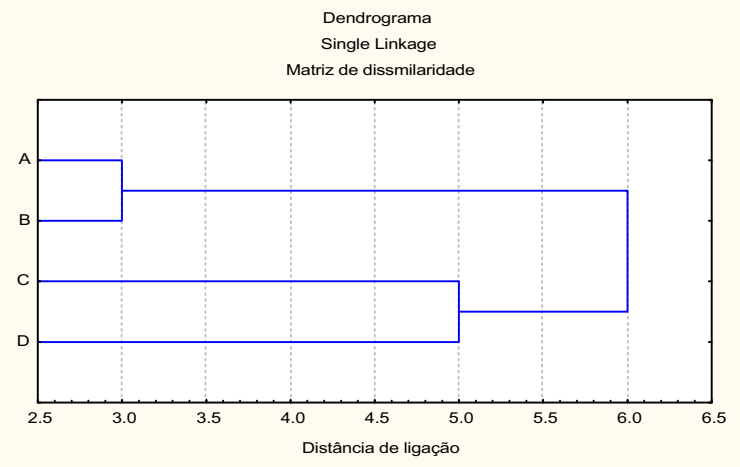


Figura 8.3. Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação simples (vizinho mais próximo).

(b) Ligação completa (vizinho mais distante)

O método da ligação completa é realizado da mesma forma que o do vizinho mais próximo, com exceção de que a distância entre grupos é tomada como a "máxima" distância entre dois elementos de cada grupo. Para ilustrar, será usado o mesmo exemplo. Assim no passo inicial, a matriz de dissimilaridade D é a mesma anterior. Assim, para ilustrar o método da ligação completa, inicialmente, são fundidos os objetos menos distantes. Então, $\min_{i,k}\{d_{ik}\}$ =$d_{AB}$=3. O próximo passo é fundir A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes. As distâncias dos vizinhos mais distantes são,

$$d_{(AB),C} = \max\{d_{AC}, d_{BC}\} = \max\{7, 9\} = 9$$

$$d_{(AB),D} = \max\{d_{AD}, d_{BD}\} = \max\{8, 6\} = 8$$

A nova matriz D para o próximo passo é:

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 9 & 0 & \\ 8 & 5 & 0 \end{bmatrix} \end{array}$$

A menor distância é entre D e C, com $d_{DC}$=5, os quais foram fundidos formando o subgrupo DC, no nível 5. Recalculando as distâncias têm-se,

$$d_{(DC),(AB)} = \max\{d_{D(AB)}, d_{C(AB)}\} = \max\{8, 9\} = 9$$

A nova matriz D fica,

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 9 & 0 \end{bmatrix} \end{array}$$

Conseqüentemente o grupo DC é fundido com AB na distância 9. Na Figura 8.4, foi apresentado o dendrograma, com os resultados alcançados.

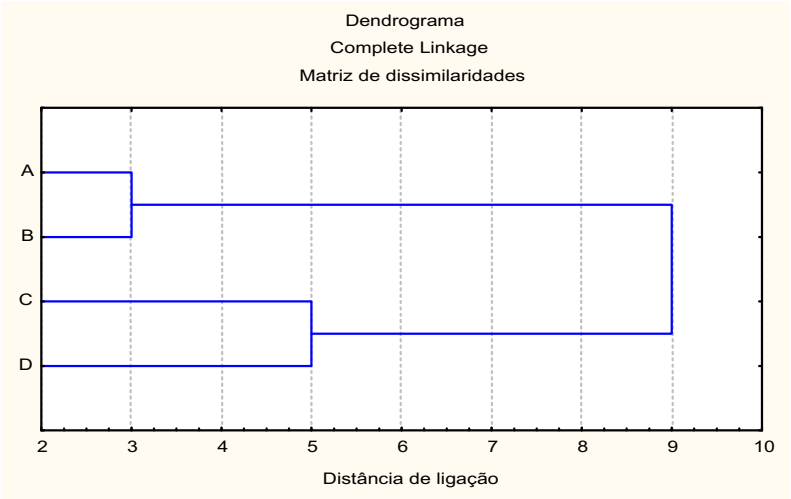


Figura 8.4. Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação completa (vizinho mais distante).

Comparando-se os resultados alcançados nas Figuras 8.3 e 8.4, pode-se notar que os dendrogramas para o vizinho mais próximo não diferem na alocação dos objetos e sim na magnitude da fusão dos grupos CD com AB.

(c) Ligação média (método do centróide)

O método da ligação média é realizado da mesma forma que o do vizinho mais próximo e mais distante, com exceção de que a distância entre grupos é tomada como a média da distância entre dois elementos de cada grupo. Para ilustrar, será usado o mesmo exemplo. Assim no passo inicial, a matriz de dissimilaridade D é a mesma anterior. Assim, para ilustrar o método da ligação completa, inicialmente, são fundidos os objetos menos distantes. Então, $\min_{i,k}\left\{d_{ik}\right\}$ =$d_{AB}$=3. O próximo passo é fundir A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes. As distâncias dos vizinhos mais distantes são,

$$d_{(AB),C} = (d_{AC} + d_{BC}) / 2 = (7 + 9) / 2 = 8$$

$$d_{(AB),D} = (d_{AD} + d_{BD}) / 2 = (8 + 6) = 7$$

A nova matriz D para o próximo passo é:

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 8 & 0 & \\ 7 & 5 & 0 \end{bmatrix} \end{array}$$

A menor distância é entre D e C, com $d_{DC}$=5, os quais foram fundidos formando o subgrupo DC, no nível 5. Recalculando as distâncias têm-se,

$$d_{(DC),(AB)} = (d_{D(AB)} + d_{C(AB)}) = (7+8)/2 = 7{,}5$$

A nova matriz D fica,

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 7{,}5 & 0 \end{bmatrix} \end{array}$$

Conseqüentemente o grupo DC é fundido com AB na distância 7,5. Na Figura 8.5, foi apresentado o dendrograma, com os resultados alcançados.

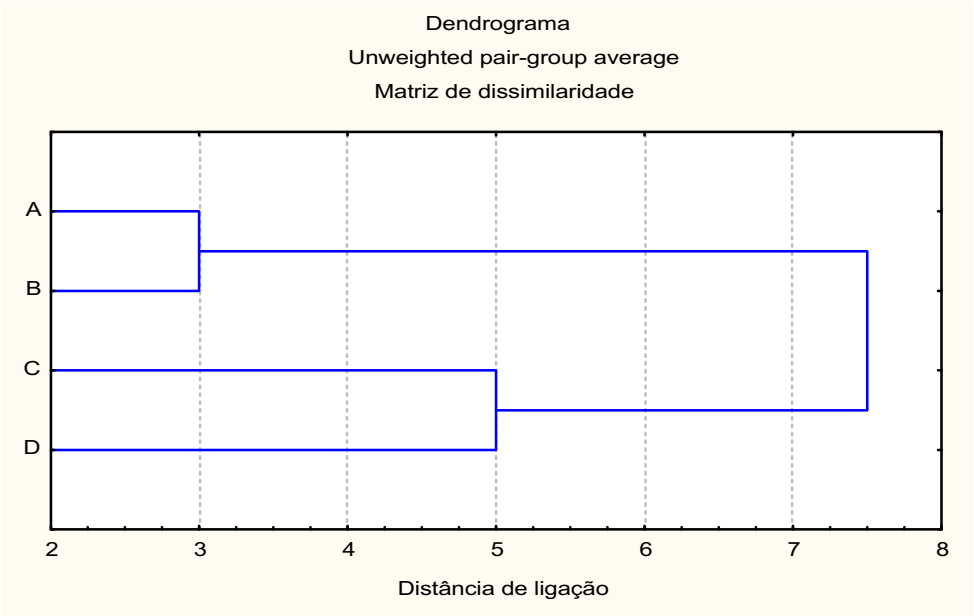


Figura 8.5. Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação média (centróide).

## 8.3.2. Agrupamentos não hierárquicos

Os agrupamentos não hierárquicos procuram a partição de n objetos em k grupos. Os métodos exigem a pré-fixação de critérios que produzam medidas sobre a qualidade da partição produzida. Um dos mais populares métodos é o das k-médias.

O algoritmo das k-médias, de uma forma bastante simplificada, é dividido em três passos:

1. Particionar os itens em k grupos iniciais arbitrariamente;
2. Percorrer a lista de itens e calcular as distâncias de cada um deles para o centróide (médias) dos grupos. Fazer a realocação do item para o grupo em que ele apresentar mínima distância, obviamente se não for o grupo ao qual este pertença. Recalcular os centróides dos grupos que ganharam e perderam o item.
3. Repetir o passo 2 até que nenhuma alteração seja feita.

**Exemplo 8.1**

Utilizando 4 itens (A, B, C e D) e 2 variáveis ($X_1$ e $X_2$) dividir em k=2 grupos, pelo método das k-médias.

| | Observação | |
|---|---|---|
| Objeto | $x_1$ | $x_2$ |
| A | 2 | 0 |
| B | 5 | 2 |
| C | 1 | 4 |
| D | 8 | 4 |

i) particionar os itens arbitrariamente em 2 grupos, como por exemplo AD e BC. Calculara a média de cada grupo.

| | Centróide | |
|---|---|---|
| Objeto | $\bar{x}_1$ | $\bar{x}_2$ |
| AD | (2+8)/2=5 | (0+4)/2=2 |
| BC | (1+5)/2=3 | (2+4)/2=3 |

ii) Neste passo a distância de cada item será computada em relação ao centróide de cada grupo e se necessário, os objetos serão realocados para o grupo mais próximo.

$$d^2_{A(AD)} = (2-5)^2 + (0-2)^2 = 13$$
$$d^2_{A(BC)} = (2-3)^2 + (0-3)^2 = 10$$

Neste caso há necessidade de realocação de A para o grupo BC, sendo que os centróides dos grupos devem ser recalculados.

| Objeto | Centróide | |
|---|---|---|
| | $\bar{x}_1$ | $\bar{x}_2$ |
| D | 8 | 4 |
| ABC | 2,667 | 2 |

Recalculando as distâncias dos objetos para o centróide dos grupos e checando a possibilidade de realocação, tem-se:

$d^2_{A,D} = 52$ $\quad d^2_{B,D} = 13$ $\quad d^2_{C,D} = 49$

$d^2_{A,(ABC)} = 4{,}44$ $\quad d^2_{B,(ABC)} = 5{,}44$ $\quad d^2_{C,(ABC)} = 6{,}77$

| Grupo | Item (distância quadrática p/ centróide) | | | |
|---|---|---|---|---|
| | A | B | C | D |
| D | 52,0 | 13,0 | 49,0 | 0,0 |
| ABC | 4,4 | 5,4 | 6,8 | 32,4 |

Nenhuma realocação deve ser realizada, pois os objetos têm menor distância para os respectivos grupos aos quais eles pertencem. Para realizar uma

checagem da estabilidade de a partição alcançada é recomendável executar novamente o algoritmo com uma nova partição inicial.

## 8.4. Exercícios

8.4.1. Agrupar os 4 objetos cuja matriz de dissimilaridades está apresentada a seguir, utilizando todos os métodos apresentados nesse material.

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 9 & 0 & & \\ 25 & 36 & 0 & \\ 49 & 100 & 16 & 0 \end{bmatrix} \end{array}$$

## 8.5. Referências

ANDERSON, T.W. **An introduction to multivariate statistical analysis**. 2nd edition. New York, John Wiley, 1984, 675p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4th edition. Prentice Hall, New Jersey, 1998. 816p.