Aula 10 – Mineração de Dados
Agrupamento: Algoritmos Hierárquicos

Profa. Elaine Faria

UFU

Agrupamento Hierárquico

- É um procedimento para transformar uma matriz de proximidade em uma sequência de partições aninhadas
- Produz uma sequência (hierarquia) de agrupamentos
- Usado em áreas que utilizam estrutura de agrupamento hierárquica
  - Ex: biologia e arqueologia

Agrupamento Hierárquico

- Seja B= $\{G_1, G_2, ..., G_m\}$ uma partição dos dados $X = \{x_1, x_2, ..., x_n\}$

- Uma partição B está aninhada em C ($B \subset C$)
  - Se cada componente de B é um subconjunto de um componente de C
    - C é formado unindo componentes de B

Exemplo: B está aninhando em C

$C=\{(x_1,x_3,x_5,x_7),(x_2,x_4,x_6,x_8),(x_9,x_{10})\}$

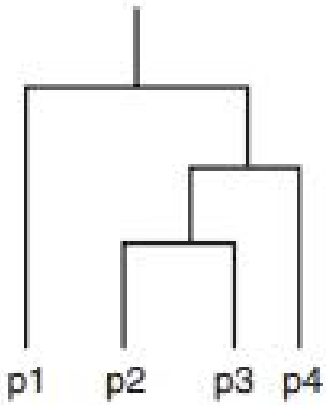
$B=\{(x_1,x_3),(x_5,x_7),(x_2),(x_4,x_6,x_8),(x_9,x_{10})\}$

Agrupamento Hierárquico

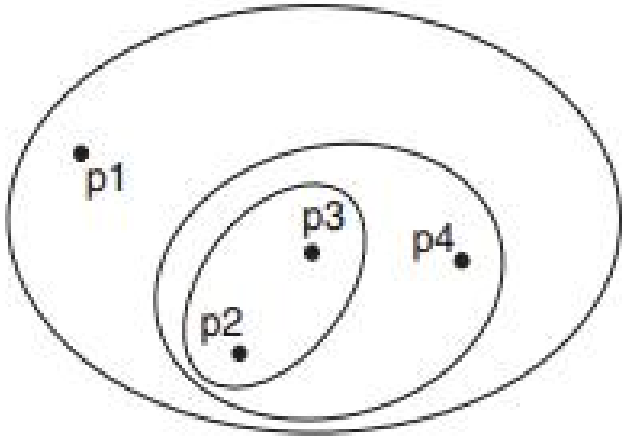
- É sempre exibido graficamente usando um diagrama chamado **dendograma**
  - Mostra os grupos-subgrupos e a ordem na qual os grupos foram unidos (visão aglomerativa) ou divididos (visão divisiva)

- Para dados com 2 dimensões, o agrupamento hierárquico pode ser visto usando o diagrama de grupos aninhados

Agrupamento Hierárquico



Dendograma

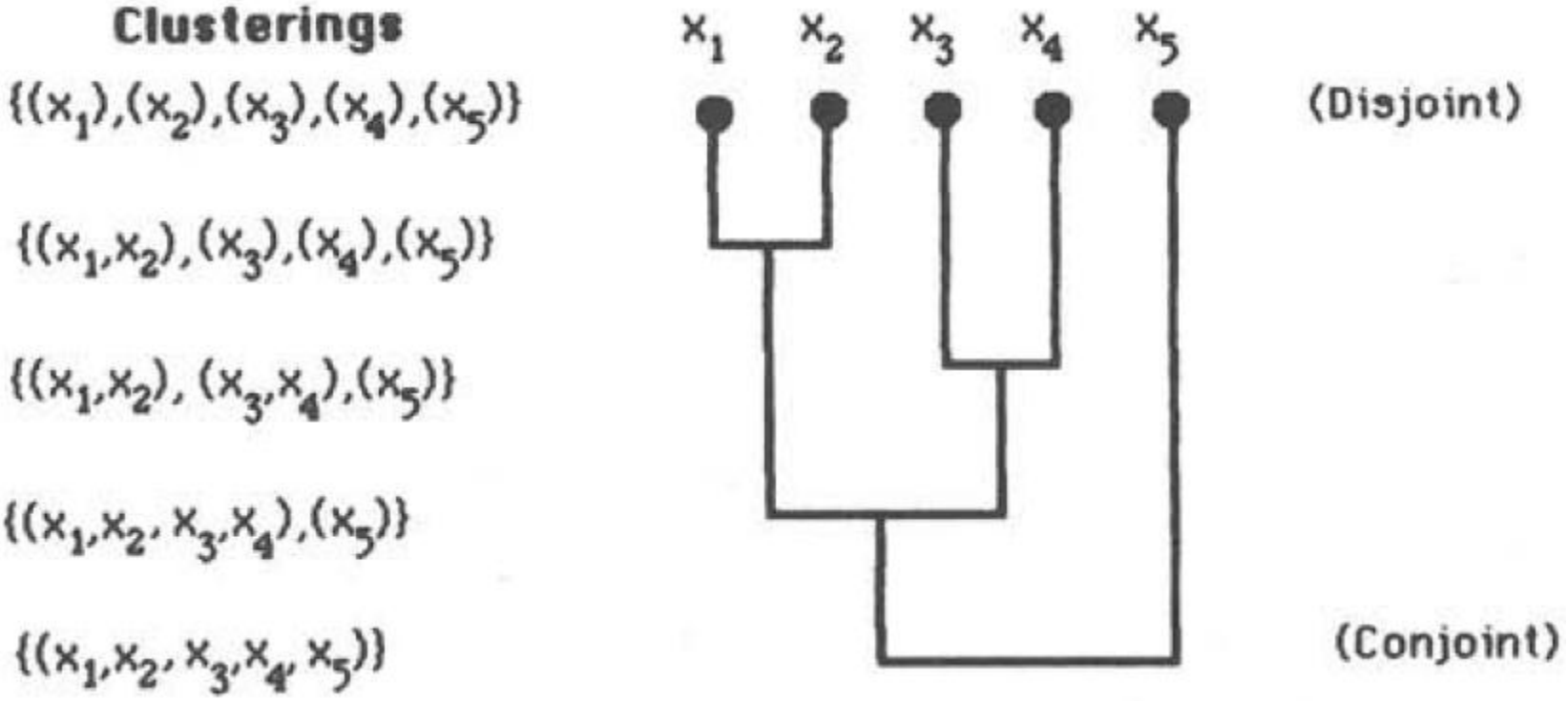


Conjunto de diagrama aninhado

Representação de agrupamento hierárquico

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Agrupamento Hierárquico



Exemplo de Dendograma

Exemplo retirado de Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.

Agrupamento Hierárquico

- **Aglomerativo**
  - Inicia com os elementos como grupos individuais
  - A cada passo, o par de elementos mais próximos é unido
  - Exige a definição de uma noção de proximidade
- **Divisivo**
  - Inicia com um grupo, contendo todos os elementos
  - A cada passo, dividir o grupo até que grupos com um único elemento sejam obtido
  - É necessário decidir qual grupo dividir a cada passo e como fazer essa divisão

Agrupamento Hierárquico

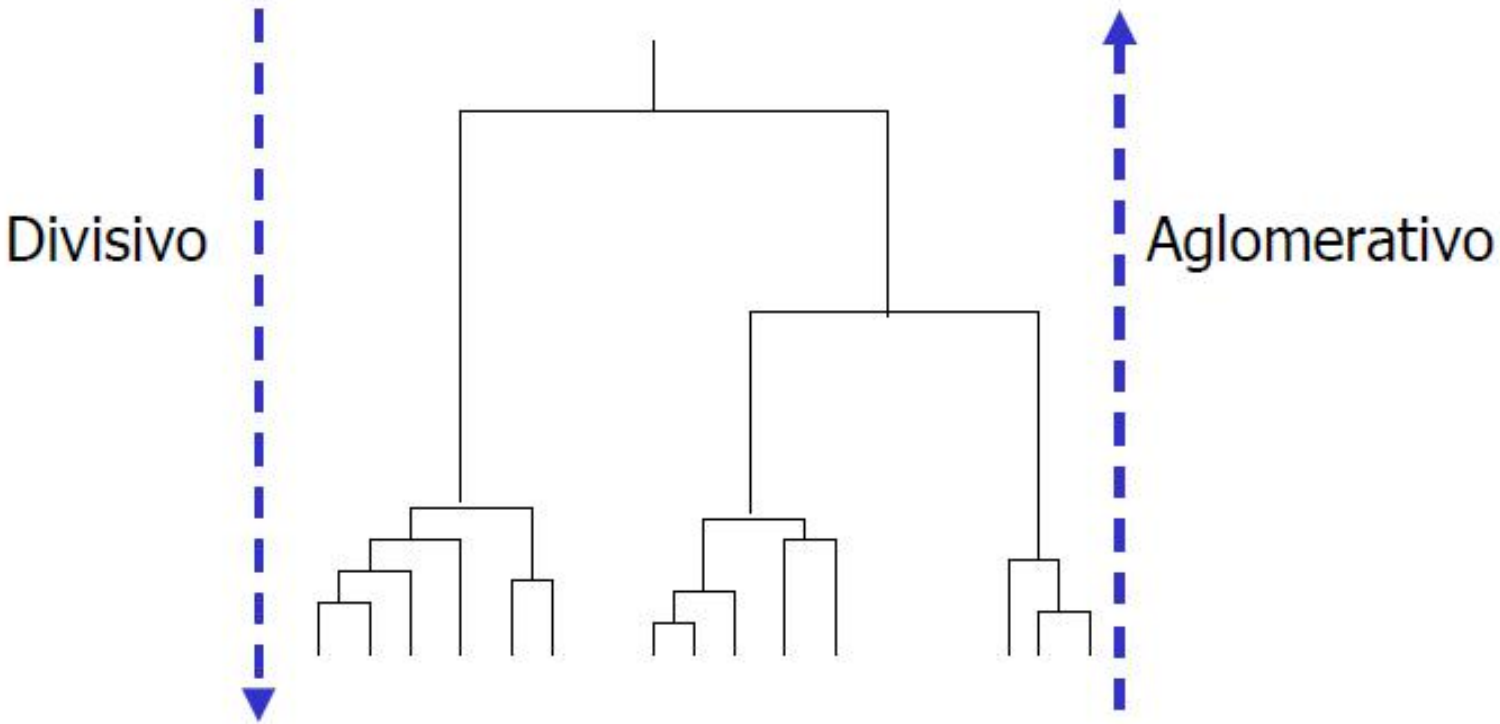


Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Algoritmo Básico para Agrupamento Hierárquico Aglomerativo

Compute a matriz de proximidade, se necessário.

repita

unir os dois grupos mais próximos

atualize a matriz de proximidade para refletir a proximidade entre o novo grupo e os grupos originais

Até que somente um grupo seja obtido

Agrupamento Hierárquico

- Como escolher uma partição?
  - Partição com *n* clusters
    - Selecionando partição com *n* clusters na seqüência de agrupamentos da hierarquia
  - Partição que melhor se encaixa nos dados
    - Procurar no dendograma grandes mudanças em níveis adjacentes
      - Nesse caso, uma mudança de j para j-1 grupos pode indicar que j é o melhor número de grupos
      - Existem outros procedimentos, alguns mais objetivos

Agrupamento Hierárquico

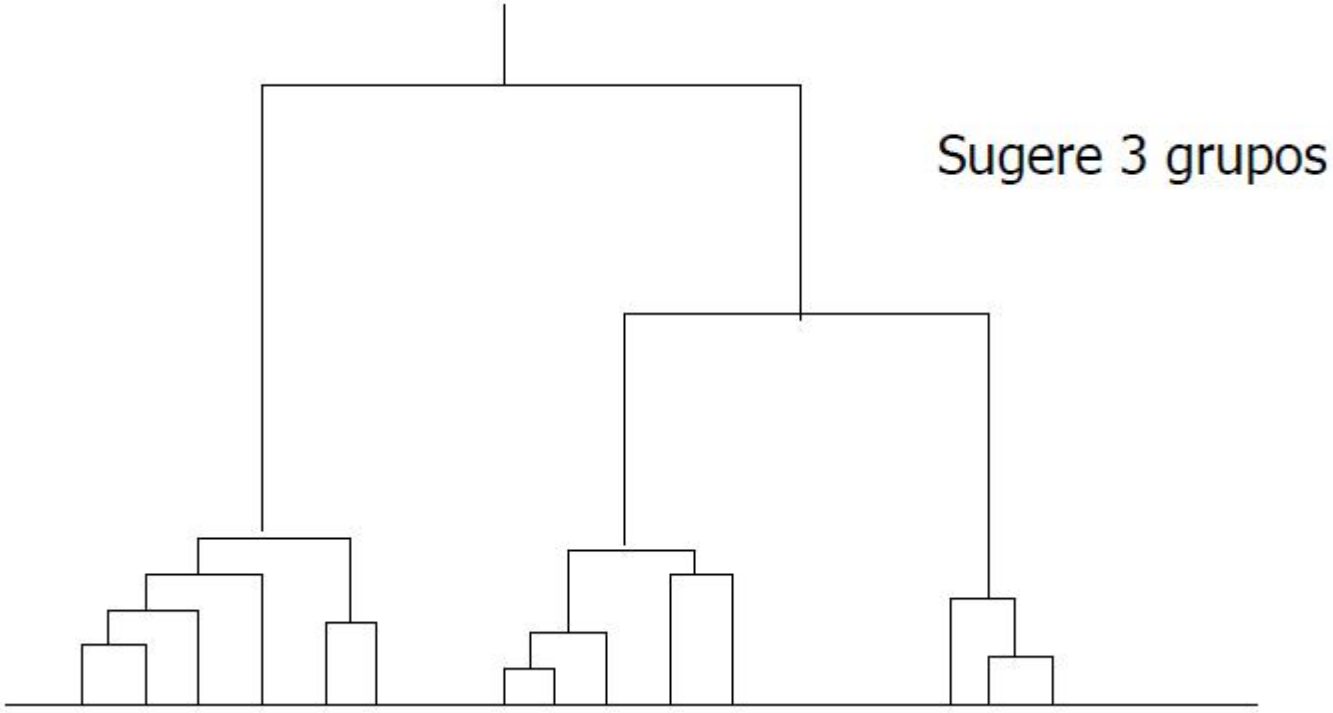


Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Agrupamento Hierárquico

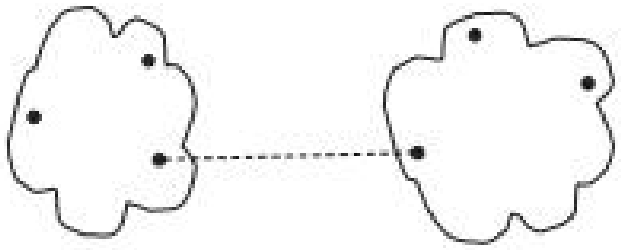
- Outra alternativa
  - Usar medida de auto-similaridade de um cluster C_t
    - Interromper processo quando a distância entre os objetos em algum dos clusters for maior que um valor θ

Definindo Proximidade entre Grupos

- Min
- Max
- Média do grupo
- Centróide
- Técnica alternativa
  - Método Ward's

Definindo Proximidade entre Grupos

- Min (*Single Link*)
  - Define a proximidade dos grupos como a proximidade entre os dois elementos mais próximos que estão em diferentes grupos
    - Usando grafos: a menor aresta entre dois nós em diferentes conjuntos de nós

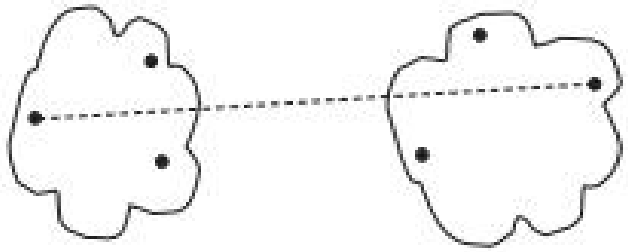


Min – *Single Link*

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Definindo Proximidade entre Grupos

- Max (*Complete Link*)
  - A proximidade dos grupos é calculada como a maior distância entre dois elementos em grupos diferentes
    - Usando grafos: a maior aresta entre dois nós em diferentes conjuntos de nós

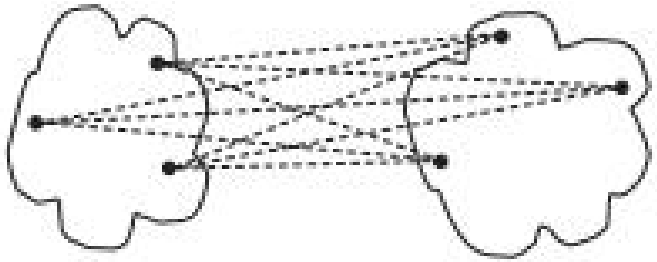


Max – *Complete Link*

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Definindo Proximidade entre Grupos

- Média do Grupo
  - Define a proximidade dos grupos como a média das proximidades entre os pares de elementos dentre todos os pares de diferentes grupos



AVG – Média do Grupo

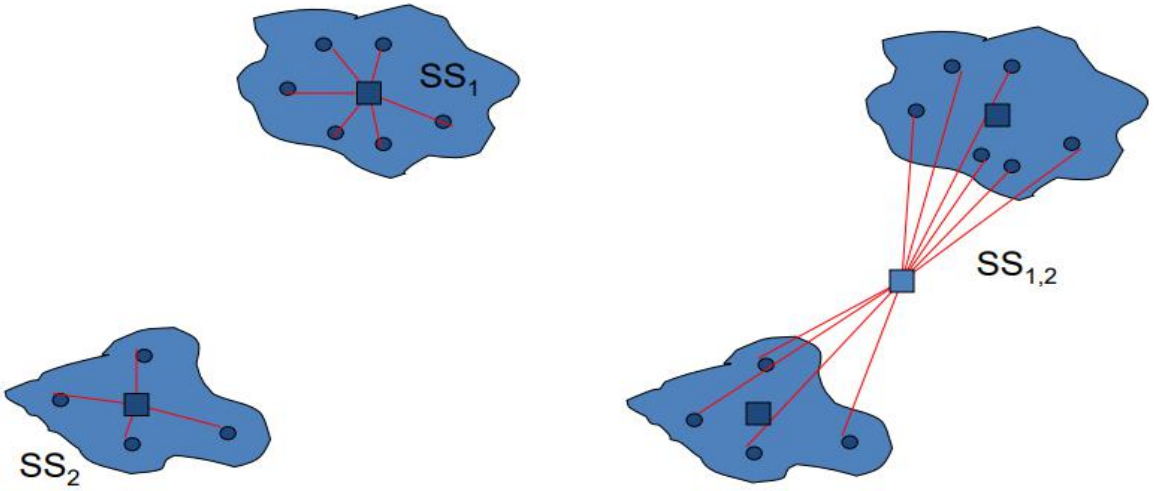
Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Definindo Proximidade entre Grupos

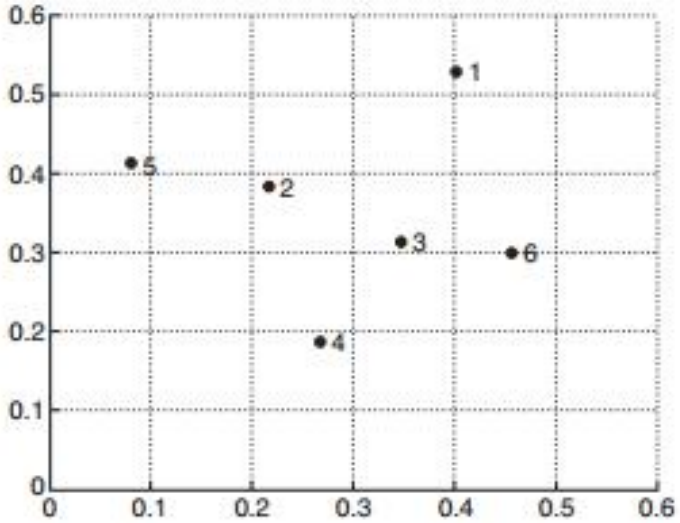
- Proximidade definida como centróides
  - Cada cluster é representado por um centróide
  - A medida de proximidade é definida com a proximidade entre os centróides dos grupos

Definindo Proximidade entre Grupos

- Método Ward's
  - Assume que um grupo é representado pelo seu centróide
  - Mede a proximidade entre dois grupos em termos do aumento no SSE que resulta da união de dois grupos
  - Minimizar a soma dos quadrados das distâncias dos elementos ao centróide dos grupos



Exemplo 1



Conjunto de seis elementos com 2-dimensões

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Coordenadas xy dos 6 elementos

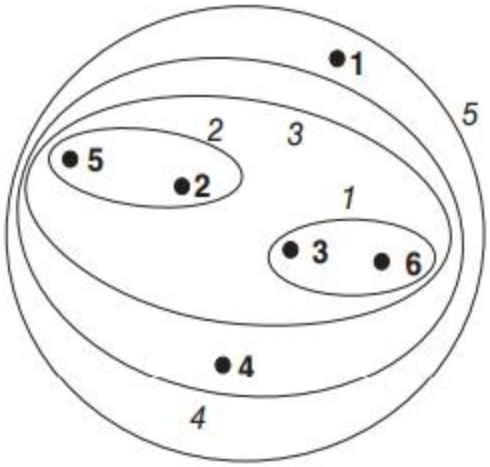
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Matriz de distância euclidiana para os 6 elementos

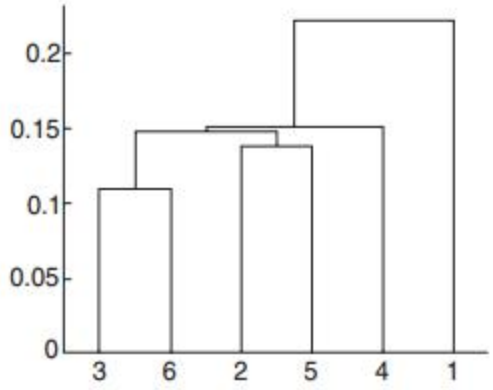
Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Exemplo 1 – *Single Link ou MIN*

- *Single Link*



Agrupamento *Single Link*



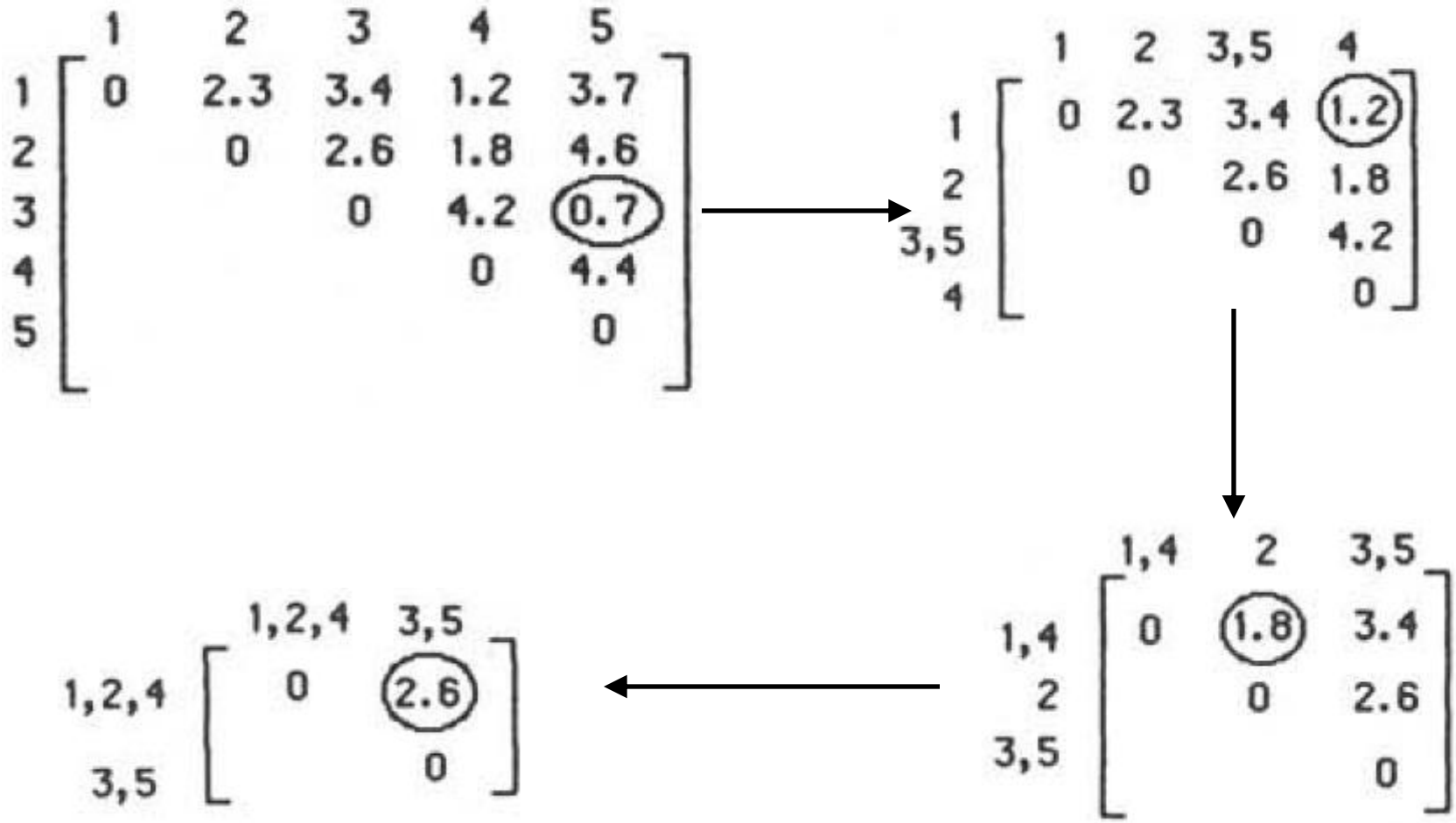
Dendograma *Single Link*

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

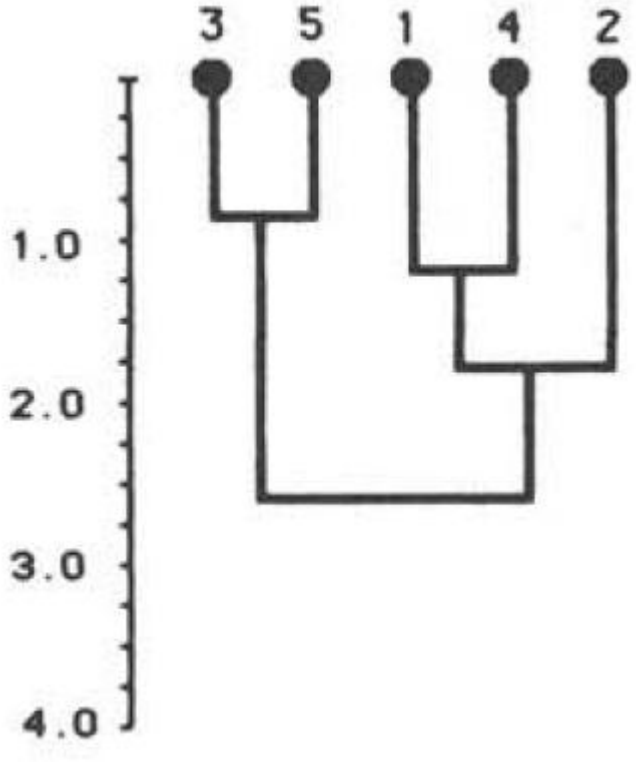
Single Link

- Propriedade da Função Mínimo (min):
  - min{**D**} = min{ min{**D**1} , min{**D**2} }
    - **D**, **D**1 e **D**2 são conjuntos de valores reais tais que **D**1 U **D**2 = **D**
  - Exemplo:
    - min{10, -3, 0, 100} = min{ min{10, -3}, min{0, 100} } = -3

Exemplo 2 - *Single Link*



Exemplo 2 - *Single Link*



Dendograma para o exemplo 2

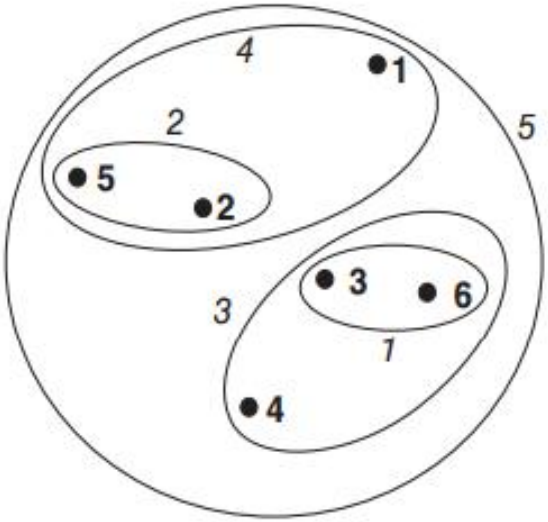
Exercício – *Single Link*

- Dada a matriz de distâncias entre 5 objetos {1,2,3,4,5}, executar o agrupamento hierárquico *single link* e mostrar a sequência de partições obtidas. Desenhe o dendograma

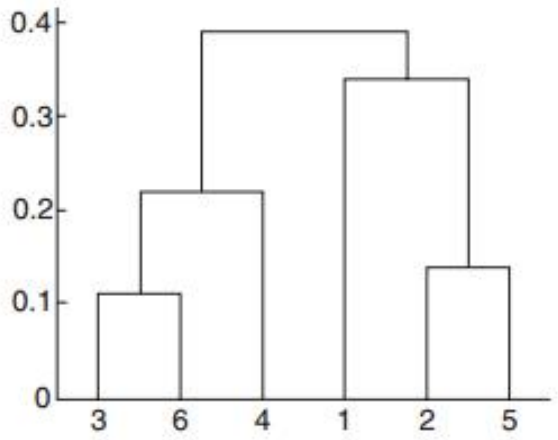
$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}\begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

Exemplo 1 – *Complete Link ou MAX ou CLIQUE*

- *Complete Link*



Agrupamento *Complete Link*



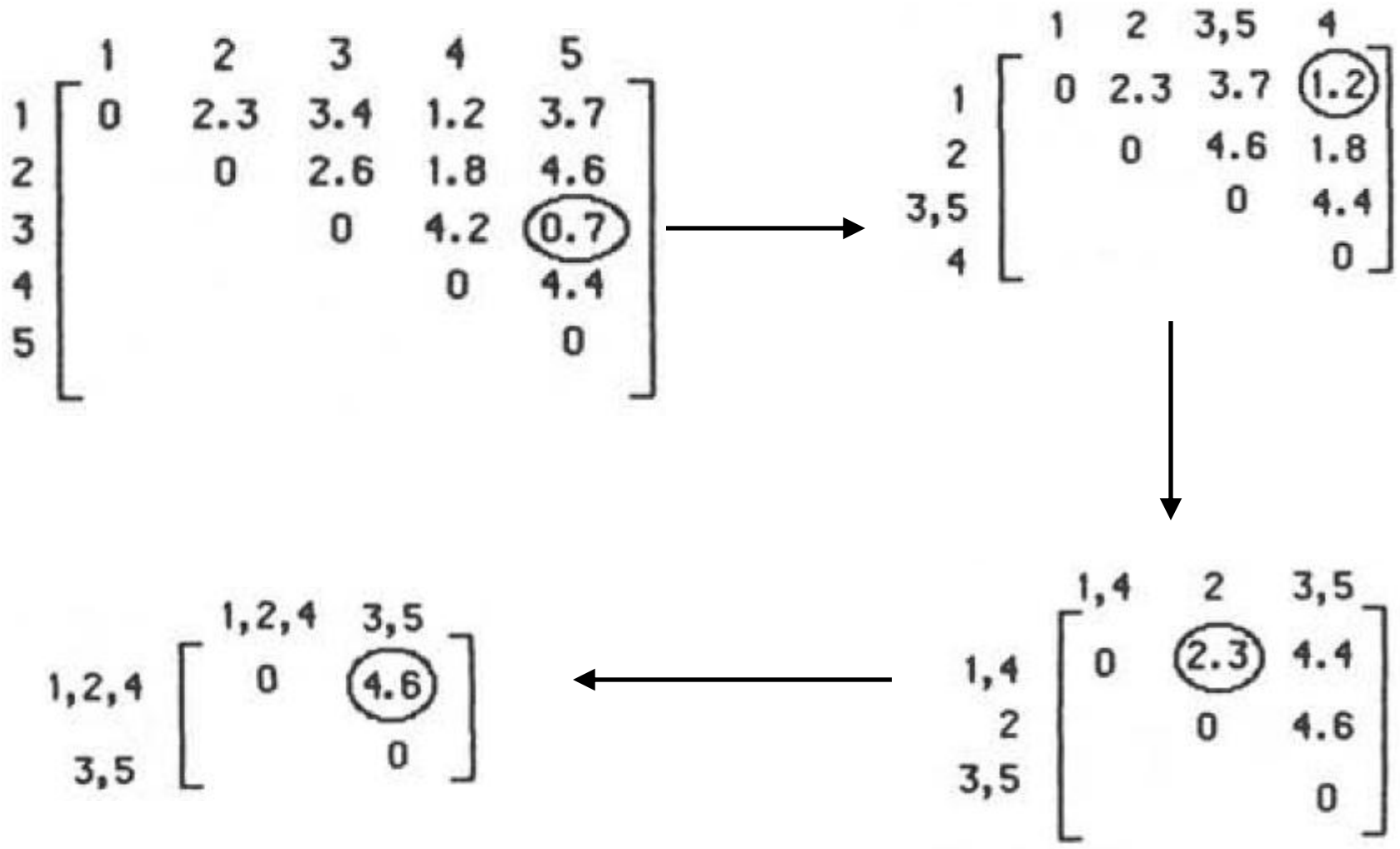
Dendograma *Complete Link*

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Complete Link

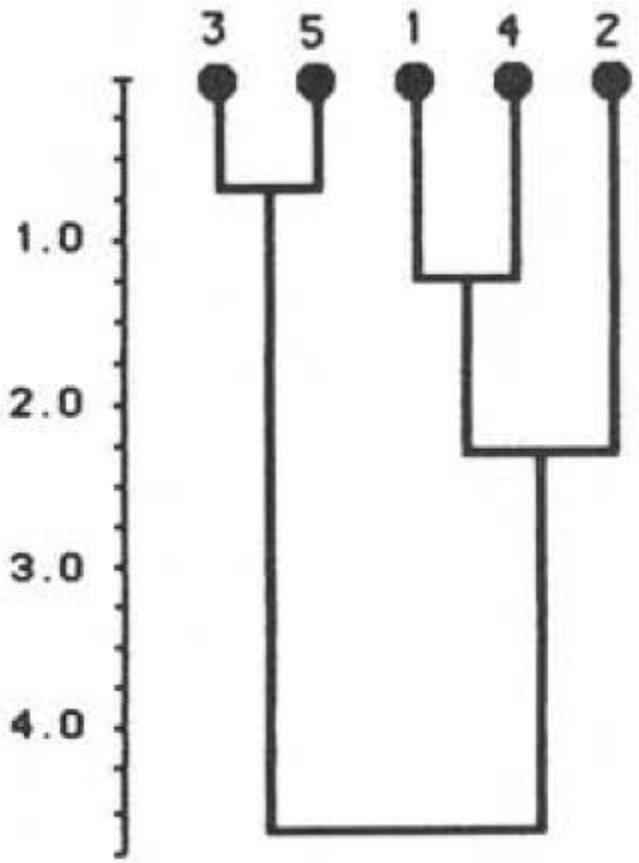
- Propriedade da Função Máximo (max):
  - max{**D**} = max{ max{**D**1} , max{**D**2} }
    - **D**, **D**1 e **D**2 são conjuntos de valores reais tais que **D**1 U **D**2 = **D**
  - Exemplo:
    - max{10, -3, 0, 100} = max{ max{10, -3}, max{0, 100} } = 100

Exemplo 2– *Complete Link*



Exemplo retirado de Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.

Exemplo 2– *Complete Link*



Dendograma para o exemplo 2

Exemplo retirado de Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.

Exercício – *Complete Link*

- Dada a matriz de distâncias entre 5 objetos {1,2,3,4,5}, executar o agrupamento hierárquico *complete link* e mostrar a sequência de partições obtidas. Desenhe o dendograma

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

Complexidade de Tempo e Espaço

- Algoritmo básico algomerativo
  - Usa matriz de similaridade (n x n)
  - Armazenamento das proximidade: ½ n^2, sendo *n* o nro de elementos; a matriz é simétrica
  - Armazenamento do *track* dos grupos: *n-1*, excluindo os grupos com 1 elemento
  - Complexidade de tempo:
    - $O(n^2)$ para calcular a matriz de proximidade
    - $O(n^3)$ para calcular o agrupamento (sem usar estruturas eficientes)
      - $O(m^2 \log m)$ se usar uma lista ordenada ou heap -> custo de ordenar

Principais questões em agrupamento hierárquico

- Falta de uma função objetivo geral
  - Usam vários critérios para decidir localmente, a cada passo, quais grupos devem ser unidos (ou divididos)
  - Evitam a dificuldade de tentar resolver um problema de otimização combinatorial

- Habilidade para tratar grupos de diferentes tamanhos
  - Como tratar os diferentes tamanhos dos pares de grupos a serem unidos?
    - Propostas ponderadas e não ponderadas

Principais questões em agrupamento hierárquico

- Decisões de união entre grupos não podem ser desfeitas
  - Uma vez que dois grupos são unidos, com base em uma decisão local, isso não pode ser desfeito futuramente

Leitura Recomendada

- Leitura do
  - Capítulo 8 (seção 8.3) do livro Tan et al, 2006.
    - Está disponível em: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

- Leitura do Capítulo 3 (seções 3.2.1 e 3.2.2) do livro do Jain e Dubes, 1999.

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
- Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.
- Ng, R.T., Han, J., Efficient and Effective Clustering Methods for Spatial Data Mining, VLDB, 1994.