Métodos de Agrupamento (Clustering) Aula 18

Prof. Dr. Paulo A. V. de Miranda
Instituto de Matemática e Estatística (IME),
Universidade de São Paulo (USP)
pmiranda@vision.ime.usp.br

Agrupamento hierárquico

Existem duas classes naturais de algoritmos para agrupamento:

- **Agrupamento aglomerativo (agglomerative clustering):** Cada item de dados é considerado como um grupo individual, e grupos são recursivamente fundidos até produzir um bom agrupamento final.
- **Agrupamento por divisão (divisive clustering):** Inicialmente, o conjunto de todos os dados é considerado como sendo um único grupo e, em seguida ele é recursivamente dividido para produzir um bom agrupamento final.

Agrupamento hierárquico

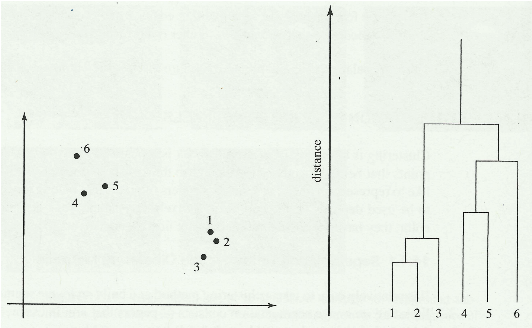
Algoritmo: Agrupamento aglomerativo (bottom-up)

Faça de cada ponto um grupo separado.
Até que o agrupamento seja satisfatório
    Junte os dois grupos com a menor distância inter-cluster.
fim

Algoritmo: Agrupamento por divisão (top-down)

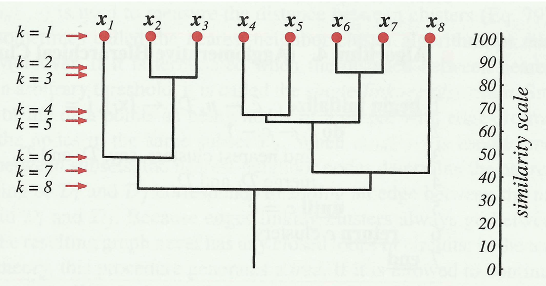
Construa um único grupo contendo todos os pontos.
Até que o agrupamento seja satisfatório
    Divida o grupo que produz os dois componentes com a maior distância inter-cluster.
fim

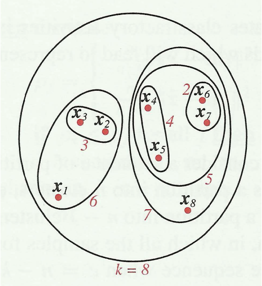
Exemplo:



Um dendrograma nos permite reconstruir o histórico de fusões, que resultou no agrupamento representado.

Exemplo:



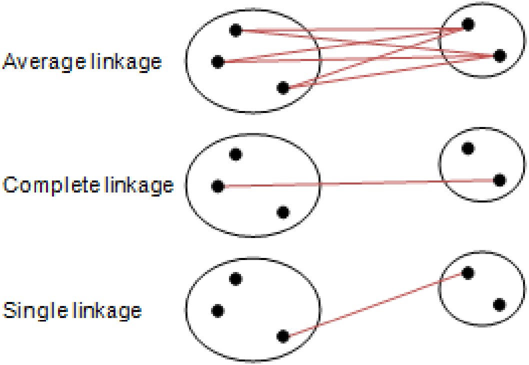


Uma outra representação é baseada em conjuntos por Diagrama de Venn.

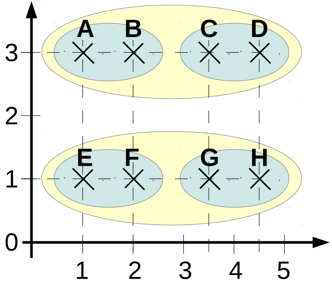
Exemplos de medidas de distância $D(X, Y)$ entre agrupamentos (clusters) X e Y:

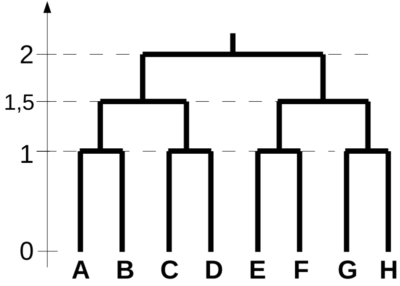
$$\begin{aligned}
D_{min}(X, Y) &= \min_{x \in X, y \in Y} d(x, y) \\
D_{max}(X, Y) &= \max_{x \in X, y \in Y} d(x, y) \\
D_{avg}(X, Y) &= \frac{1}{|X| \cdot |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)
\end{aligned}$$

Exemplo:



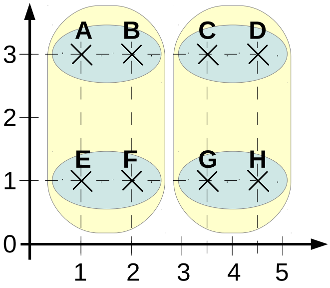
Exemplo: *single linkage* ou *nearest neighbour* - $D_{min}(X, Y)$

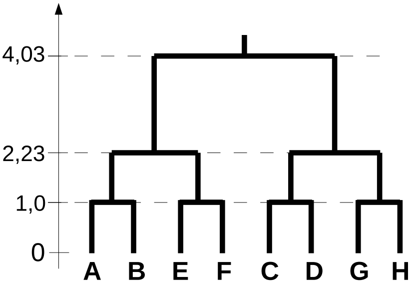




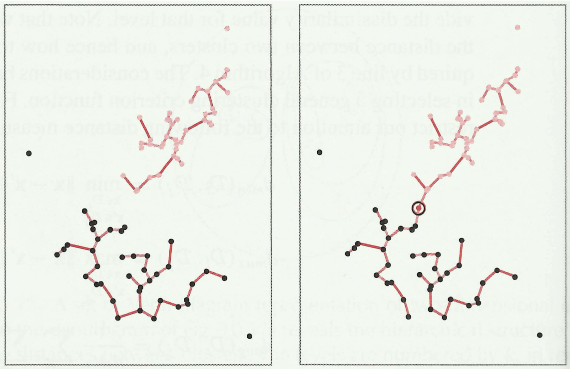
Agrupamento hierárquico

Exemplo: *complete linkage* ou *farthest neighbour* - $D_{max}(X, Y)$

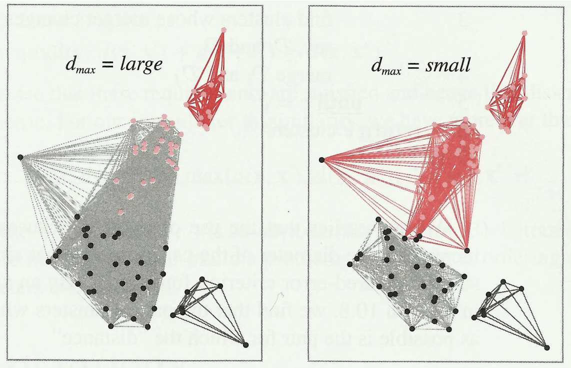




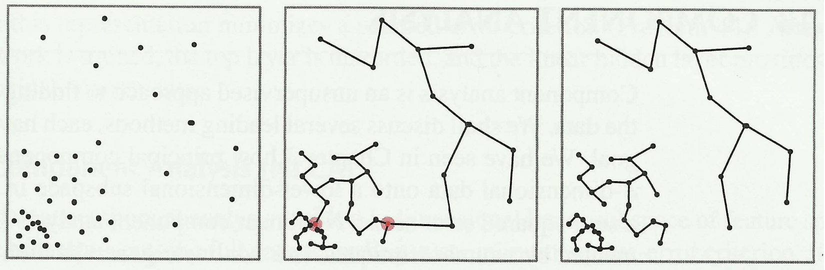
Exemplo: *single linkage* ou *nearest neighbour* - $D_{min}(X, Y)$



Exemplo: *complete linkage* ou *farthest neighbour* - $D_{max}(X, Y)$

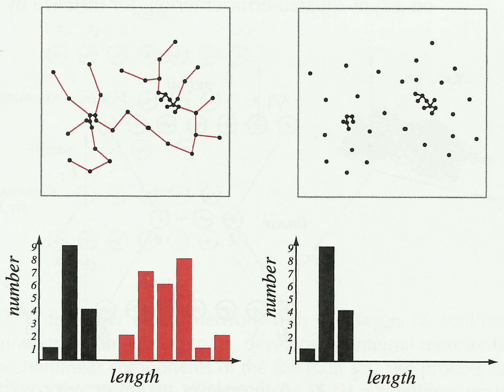


Outras maneiras de dividir o grafo da Árvore de Espalhamento Mínima em subgrafos:



Agrupamento pela remoção de arestas inconsistentes, cujo comprimento é significativamente maior do que a média de todas outras arestas incidentes nos seus nós.

Exemplo: Agrupamento pelo uso de dados estatísticos (e.g., distribuição do comprimento das arestas):



Ao apagar todas arestas com valores elevados, podemos então extrair os conjuntos densos como os maiores componentes conexos do grafo restante.

Bibliografia

- *L.M. Rocha, F.A.M. Cappabianco, A. X. Falcão,*
  Data clustering as an optimum-path forest problem with applications in image analysis
  International Journal of Imaging Systems and Technology, vol. 19, no. 2, pp. 50-68, 2009.
  http://onlinelibrary.wiley.com/doi/10.1002/ima.20191/abstract
- *L.M. Rocha, A.X. Falcão, L.G.P. Meloni,*
  A Robust Extension of the Mean Shift Algorithm using Optimum Path Forest,
  Proc. of the 12th Intl. Workshop on Combinatorial Image Analysis, pp. 29-38, 2008.
- *Richard O. Duda, Peter E. Hart, David G. Stork,*
  Pattern Classification, Segunda edição,
  Wiley-Interscience Publication, John Wiley & Sons, Inc., 2001.