SCC0173 – Mineração de Dados Biológicos

Agrupamento de Dados – Partes I & II:
Conceituação e Métodos Hierárquicos

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

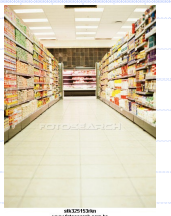
Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - gentilmente cedidos pelo Prof. Eduardo R. Hruschka
  - de (Tan et al., 2006)
  - de E. Keogh (SBBD 2003)
  - de G. Piatetsky-Shapiro (KDNuggets)
- Algumas figuras são de autoria e foram gentilmente cedidas por Lucas Vendramin

Conteúdo

- Agrupamento de Dados
- Algoritmos Hierárquicos
  - Métodos Aglomerativos
    - Single Linkage
    - Complete Linkage
    - Average Linkage
  - Métodos Divisivos

Motivação



Humanos se interessam por *categorizações*

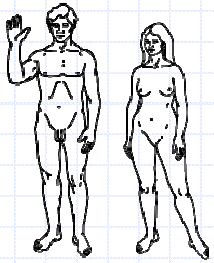
➢ Música: erudita, popular, religiosa, etc.

➢ Filmes: Animação, Aventura, Comédia, Drama, etc.

Diversas ciências se baseiam na *organização* de objetos de acordo com suas similaridades

- Biologia:

Reino: Animalia
Ramo: Chordata
Classe: Mammalia
Ordem: Primatas
Família: *Hominidae*
Gênero: *Homo* (homem moderno e parentes)
Espécie: *Homo sapiens*



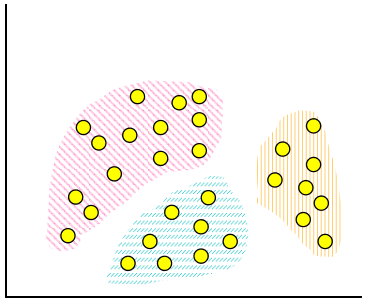
Prof. Eduardo R. Hruschka

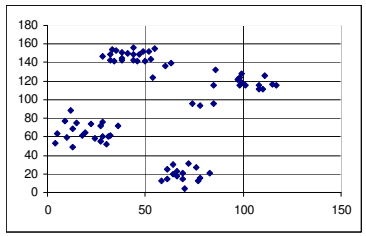
No entanto...

- Existem muitas situações nas quais não sabemos de antemão uma maneira apropriada de **agrupar** uma coleção de objetos de acordo com suas "similaridades"
  - massas de dados, possivelmente descritas por várias características (atributos) diferentes...
- Frequentemente não sabemos sequer se existe algum **agrupamento natural** dos objetos segundo um conjunto de características que descrevem esses objetos
  - que possa ser representativo de um ou mais fenômenos de interesse por trás dos dados em questão

Agrupamento de Dados (Clustering)

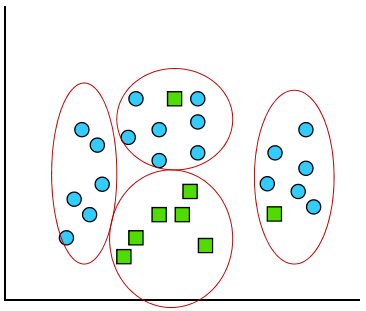
- Aprendizado não supervisionado
- Encontrar grupos "naturais" de objetos não rotulados





Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

Classificação X Clustering



Classificação:

Aprender um método para predizer as categorias (classes) de padrões não vistos a partir de exemplos pré-rotulados (classificados)

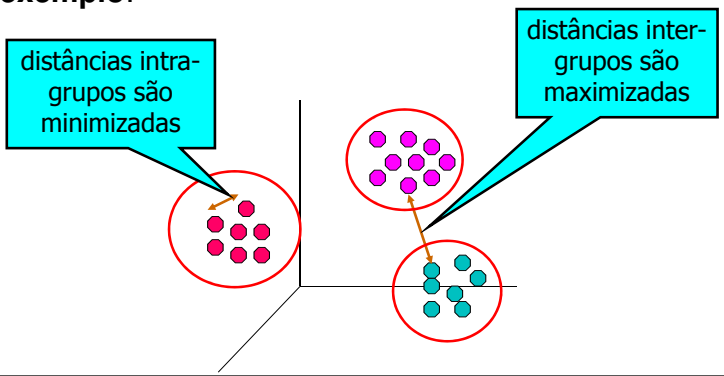
Agrupamento de Dados (Clustering):

Encontrar os rótulos das categorias (grupos ou **clusters**) e possivelmente o número de categorias diretamente a partir dos dados

Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

O que é Agrupamento de Dados

- Encontrar grupos de objetos tais que os objetos em um grupo sejam similares (ou relacionados) entre si e diferentes dos (ou não relacionados aos) objetos dos demais grupos
- **Por exemplo**:



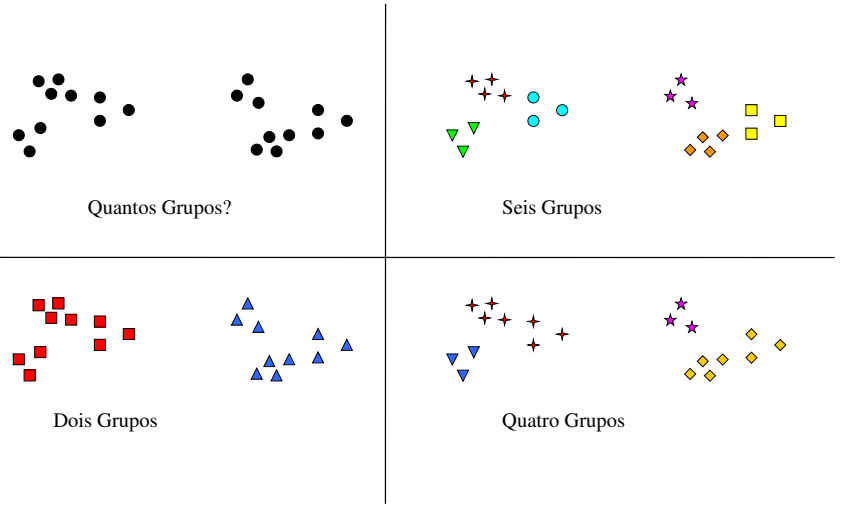
O que Não é Agrupamento de Dados

- Classificação Supervisionada
  - Utilizando rótulos das classes
- Simples Segmentação
  - Por exemplo: dividir estudantes em blocos de acordo com uma ordem alfabética de nome ou sobrenome
- Resultados de uma consulta (query)
  - Por exemplo: registros de empregados com idade de 35 a 45 anos e salário entre R$1000 e R$3000
- …

Algumas Aplicações de Clustering

- **Marketing:** descobrir grupos de clientes / nichos de mercado e usá-los para marketing direcionado
- **Bioinformática:** encontrar grupos de genes com expressões semelhantes, ...
- **Mineração de Textos:** categorizar documentos
- **Vários Outros:**
  - segmentação de imagens
  - detecção de anomalias
  - ...

Dificuldade: Noção de Grupo pode ser Ambigua



Visualizando Clusters

- Sistema visual humano é muito poderoso para reconhecer padrões
- Entretanto...
  - "*Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent*" (Carl Sagan)
- Everitt et al., Cluster Analysis, Chapter 2 (Visualizing Clusters), Fourth Edition, Arnold, 2001

Definindo o que é um **Cluster**

- Conceitualmente, definições são subjetivas:
  - Homogeneidade (coesão interna)...
  - Heterogeneidade (separação entre grupos)...
  - Conectividade, Densidade, ...
- É preciso formalizar matematicamente
- Existem diversas medidas
  - Cada uma induz (impõe) uma estrutura aos dados...
  - Em geral, baseadas em algum tipo de **(dis)similaridade**

Como Definir (Dis)Similaridade ?

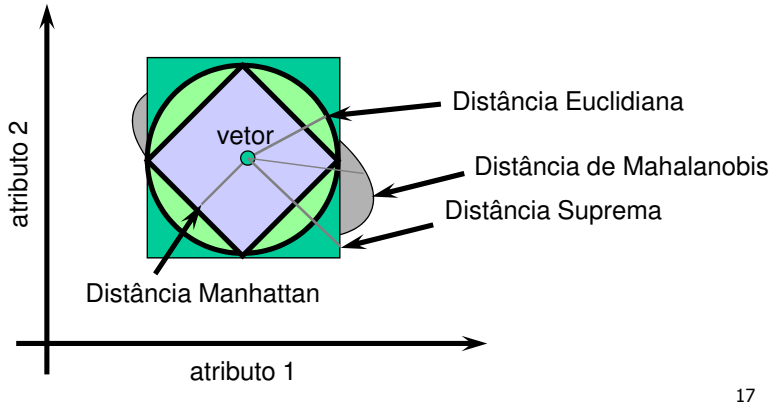
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Medidas de (Dis)Similaridade

- Já conhecemos diversas medidas, por exemplo:
  - Minkowski (Manhatan, Euclideana, Suprema, ...)
  - Pearson
  - Casamento Simples (Simple Matching)
  - Jaccard
  - Cosseno
  - ...

Relembrando (Medidas de Distância)

- Onde se situam os pontos eqüidistantes de um vetor



Notação

- **Matriz de Dados X:**
  - *N* linhas (objetos) e *n* colunas (atributos):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

- Cada padrão (linha da matriz) é denotado por um vetor $\mathbf{x}_i$
  - Exemplo:

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} & \cdots & x_{1n} \end{bmatrix}$$

Notação

- **Matriz de (Dis)similaridade:**
  - *N* linhas e *N* colunas:

$$\mathbf{D} = \begin{bmatrix} d(\mathbf{x}_1,\mathbf{x}_1) & d(\mathbf{x}_1,\mathbf{x}_2) & \cdots & d(\mathbf{x}_1,\mathbf{x}_N) \\ d(\mathbf{x}_2,\mathbf{x}_1) & d(\mathbf{x}_2,\mathbf{x}_2) & \cdots & d(\mathbf{x}_2,\mathbf{x}_N) \\ \vdots & & \ddots & \vdots \\ d(\mathbf{x}_N,\mathbf{x}_1) & d(\mathbf{x}_N,\mathbf{x}_2) & \cdots & d(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix}$$

- Simétrica se *d* possuir propriedade de simetria

Abordagens de Clustering

- Muitos métodos / algoritmos diferentes:
  - Para dados numéricos e/ou categóricos
  - Para dados **relacionais** ou **não relacionais**
  - Para obter **partições** ou **hierarquias** de partições
  - Grupos **mutuamente exclusivos** ou **sobrepostos**
  - ...

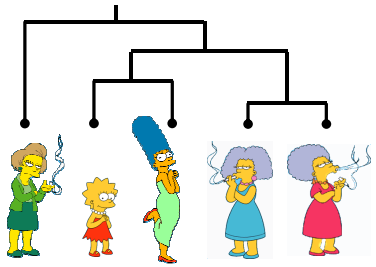
Métodos Relacionais *vs* Não Relacionais

- Métodos Não Relacionais:
  - Demandam a matriz de dados **X**
- Métodos Relacionais:
  - Operam exclusivamente sobre a matriz **D**
  - Vantagens:
    - Abordagem unificada para quaisquer tipos de atributos
    - Dados sigilosos
    - Domínios de aplicação subjetivos (e.g. ciências sociais)
  - Desvantagem: Custo computacional em geral mais elevado

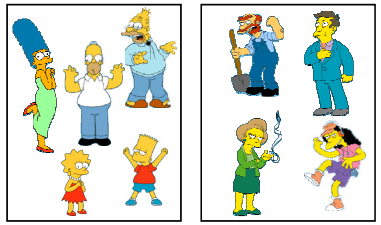
Métodos Particionais *vs* Hierárquicos

- **Métodos Particionais:** constroem uma partição dos dados
- **Métodos Hierárquicos:** constroem uma hierarquia de partições

Hierárquicos



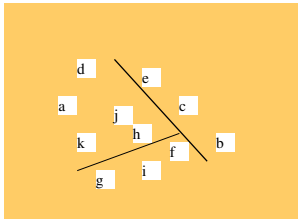
Particionais



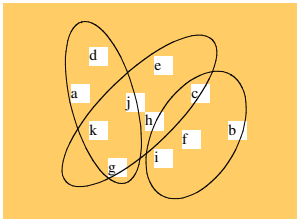
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Partições com e sem Sobreposição

Sem sobreposição



Com sobreposição



Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

Métodos Hierárquicos

- Hierarquia é um tipo usual de organização
- Exemplo: Árvores Filogenéticas em Biologia

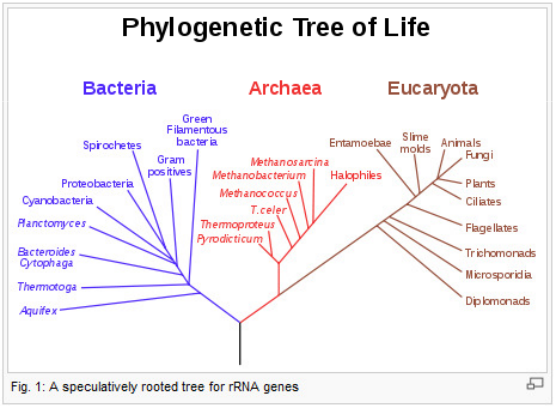
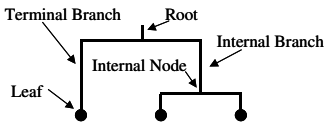


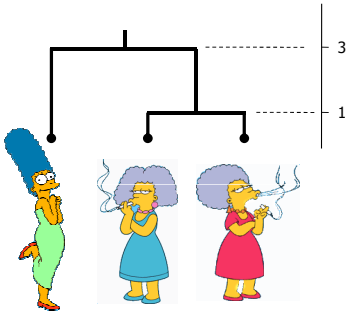
Fig. 1: A speculatively rooted tree for rRNA genes

http://en.wikipedia.org/wiki/Phylogenetic_tree

Métodos Hierárquicos

Dendrograma: Hierarquia + Dissimilaridades entre Clusters





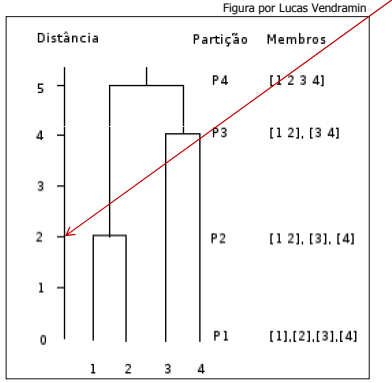
* A dissimilaridade entre dois clusters (possivelmente **atômicos**) é representada como a altura do nó interno mais baixo compartilhado

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

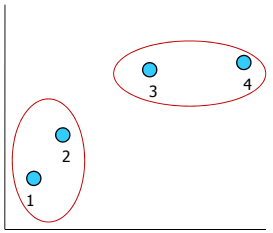
Exemplo de Dendrograma

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 7 & 13 \\ 2 & 0 & 5 & 10 \\ 7 & 5 & 0 & 4 \\ 13 & 10 & 4 & 0 \end{bmatrix}$$

Figura por Lucas Vendramin



Dendrograma



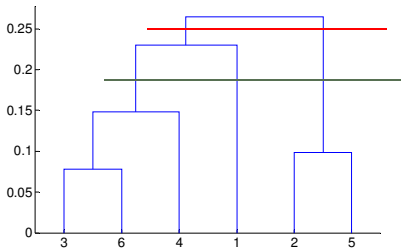
uma das partições aninhadas

Dendrogramas e Partições

- Partições são obtidas via **cortes** no dendrograma
  - cortes horizontais
  - no. de grupos da partição = no. de interseções
- **Exemplos:**

$P_2 = \{ (\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_6), (\mathbf{x}_2,\mathbf{x}_5) \}$

$P_1 = \{ (\mathbf{x}_1), (\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_6), (\mathbf{x}_2,\mathbf{x}_5) \}$



Baseado no original do Prof. Eduardo R. Hruschka

Outro Exemplo de Dendrograma

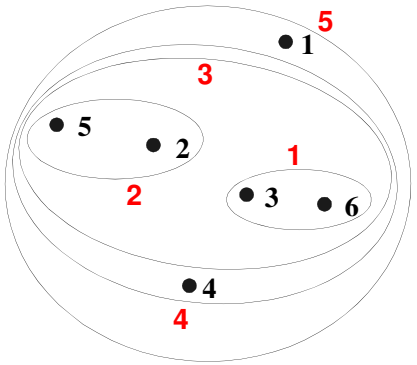
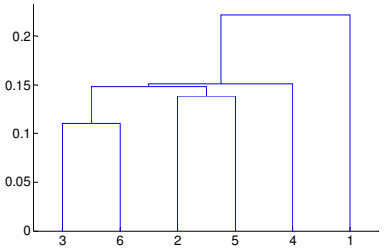
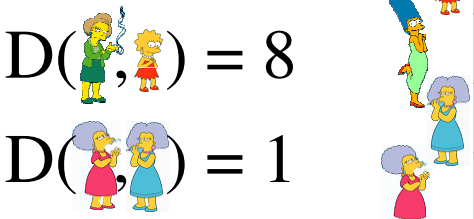


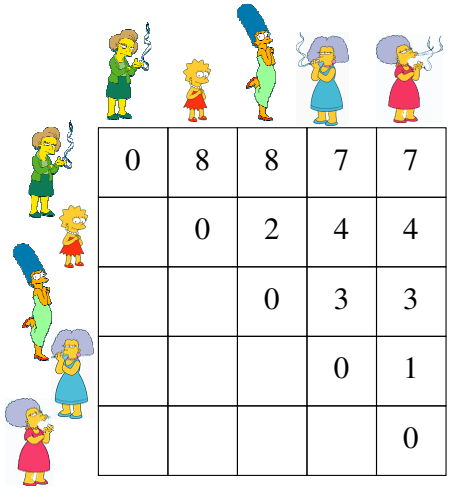
Diagrama de Venn



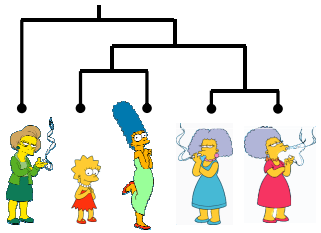
Dendrograma

Algoritmos hierárquicos podem operar sobre uma matriz de dissimilaridades: relacionais !





Métodos Clássicos para Agrupamento Hierárquico

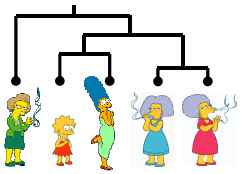


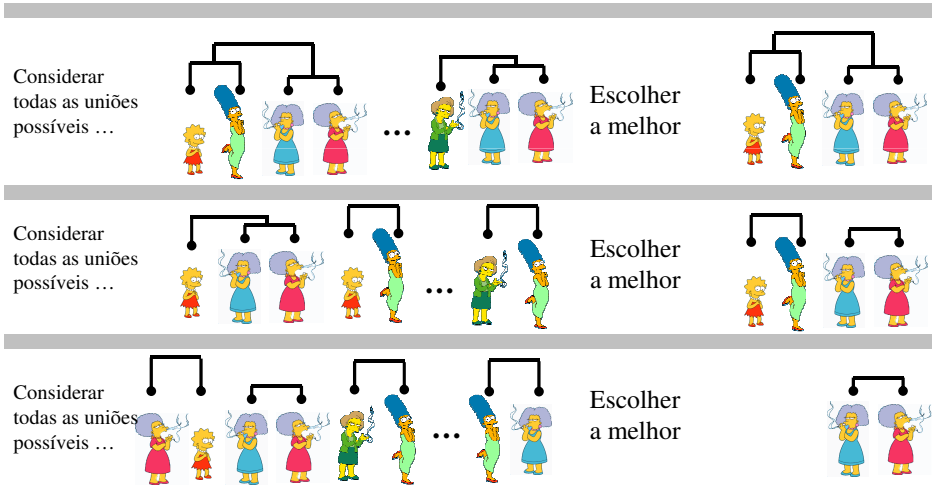
Bottom-Up (aglomerativos):
- Iniciar colocando cada objeto em um *cluster*
- Encontrar o melhor par de *clusters* para unir
- Unir o par de clusters escolhido
- Repetir até que todos os objetos estejam reunidos em um só *cluster*

Top-Down (divisivos):
- Iniciar com todos objetos em um único *cluster*
- Sub-dividir o *cluster* em dois novos *clusters*
- Aplicar o algoritmo recursivamente em ambos, até que cada objeto forme um *cluster* por si só

Bottom-Up (aglomerativo):

Iniciando com cada objeto em seu próprio cluster, encontrar o melhor par de *clusters* para unir em um novo *cluster*. Repetir até que todos os *clusters* sejam fundidos em um único *cluster*.





O que é a "Melhor União Possível" ?

- Na maioria dos algoritmos, trata-se da união entre os dois grupos (clusters) mais similares
  - diferentes algoritmos utilizam diferentes estratégias para avaliar a (dis)similaridade entre pares de clusters
  - tudo que precisamos para compreender as diferenças entre eles é entender **como comparar dois clusters**

Como Comparar os Clusters?

- **Single Linkage** (ou Min):
  - Dissimilaridade entre *clusters* é dada pela menor dissimilaridade entre dois objetos (um de cada *cluster*)

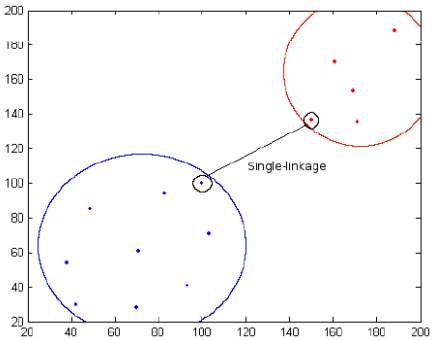
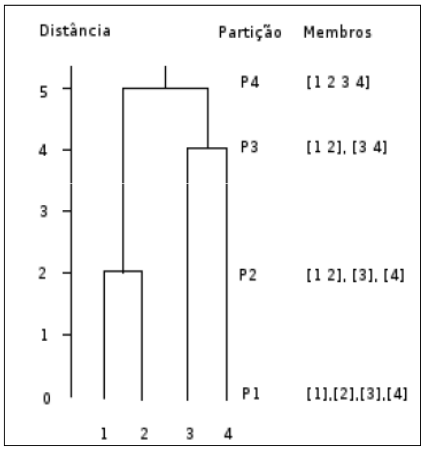


Figura por Lucas Vendramin

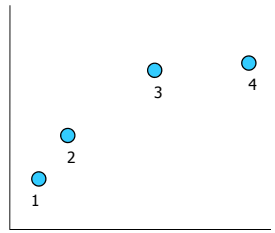
Exemplo: no quadro...

Figura por Lucas Vendramin



dendrograma

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & 2 & 7 & 13 \\ 2 & 0 & 5 & 10 \\ 7 & 5 & 0 & 4 \\ 13 & 10 & 4 & 0 \end{bmatrix}$$



objetos

Propriedade Útil

- Propriedade da Função Mínimo (min):
  - $\min\{\mathbf{D}\} = \min\{ \min\{\mathbf{D}_1\} , \min\{\mathbf{D}_2\} \}$
    - $\mathbf{D}$, $\mathbf{D}_1$ e $\mathbf{D}_2$ são conjuntos de valores reais tais que $\mathbf{D}_1 \cup \mathbf{D}_2 = \mathbf{D}$
  - Exemplo:
    - $\min\{10, -3, 0, 100\} = \min\{ \min\{10, -3\}, \min\{0, 100\} \} = -3$
  - Propriedade vale recursivamente (para $\min\{\mathbf{D}_1\}$ e $\min\{\mathbf{D}_2\}$)
- Utilidade para Single Linkage
  - Dada a distância entre os grupos **A** e **B** e entre **A** e **C**
    - É trivial calcular a distância entre **A** e (**B** $\cup$ **C**) !

Exemplo (Everitt et al., 2001):

- Consideremos a seguinte matriz de distâncias iniciais ($\mathbf{D}_1$) entre 5 objetos {1,2,3,4,5}. Qual par de objetos será escolhido para formar o 1º *cluster*?

$$\mathbf{D}_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ \boxed{2} & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

- A menor distância entre objetos é $d_{12}=d_{21}=2$, indicando que estes dois objetos serão unidos em um *cluster.* Na seqüência, calcula-se:

$d_{(12)3}=min\{d_{13},d_{23}\}=d_{23}=5;$

$d_{(12)4}=min\{d_{14},d_{24}\}=d_{24}=9;$

$d_{(12)5}=min\{d_{15},d_{25}\}=d_{25}=8;$

- Desta forma, obtém-se uma nova matriz de distâncias ($\mathbf{D}_2$), que será usada na próxima etapa do agrupamento hierárquico:

$$\mathbf{D}_2 = \begin{array}{c} \\ 12 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & \boxed{3} & 0 \end{bmatrix}$$

- Qual o novo *cluster* a ser formado?
- Unindo os objetos **5** e **4** obtemos três clusters: {1,2}, {4,5}, {3}
- Como $d_{(12)3}$ já está calculado, calculamos na seqüência:

  $d_{(12)(45)} = min\{d_{(12)(4)}, d_{(12)(5)}\} = d_{(12)(5)} = 8$ ➡ propriedade anterior !

  $d_{(45)3} = min\{d_{43}, d_{53}\} = d_{43} = 4$

  obtendo a seguinte matriz:

$$\mathbf{D}_3 = \begin{array}{c} \\ 12 \\ 3 \\ 45 \end{array} \begin{bmatrix} 0 & & \\ 5 & 0 & \\ 8 & \boxed{4} & 0 \end{bmatrix}$$

* Unir *cluster* {3} com {4,5};

* Finalmente, unir todos os *clusters* em um único *cluster*

Exercício:

- Obtenha o dendrograma completo para o exemplo visto de execução do single linkage (matriz de distâncias abaixo)

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

Exercícios:

- Converta a matriz de similaridades abaixo em uma matriz de dissimilaridades através da transformação d = 1 – s e em seguida execute o algoritmo single linkage sobre ela:

$$\mathbf{S} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{|ccccc|} \multicolumn{5}{c}{\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array}} \\ \hline 1{,}00 & 0{,}90 & 0{,}10 & 0{,}65 & 0{,}20 \\ 0{,}90 & 1{,}00 & 0{,}70 & 0{,}60 & 0{,}50 \\ 0{,}10 & 0{,}70 & 1{,}00 & 0{,}40 & 0{,}30 \\ 0{,}65 & 0{,}60 & 0{,}40 & 1{,}00 & 0{,}80 \\ 0{,}20 & 0{,}50 & 0{,}30 & 0{,}80 & 1{,}00 \\ \hline \end{array}$$

Mostre o dendrograma completo resultante

- Execute o agora o algoritmo sobre a **matriz S original**, <u>modificando o que for necessário</u>. Converta similaridades em dissimilaridades apenas para mostrar o dendrograma

Características de Single Linkage

- **Pode encontrar grupos com formas complexas**

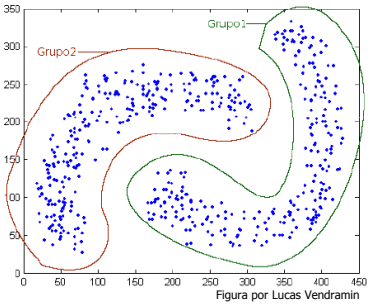


Figura por Lucas Vendramin

- **Mas é muito sensível a ruído...**

Como Comparar os Clusters?

- **Complete Linkage** (ou Max):
  - Dissimilaridade entre *clusters* é dada pela maior dissimilaridade entre dois objetos (um de cada *cluster*)

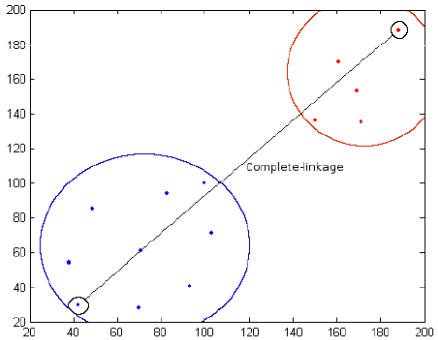


Figura por Lucas Vendramin

Propriedade Útil

- Propriedade da Função Máximo (max):
  - $\max\{\mathbf{D}\} = \max\{ \max\{\mathbf{D}_1\} , \max\{\mathbf{D}_2\} \}$
    - $\mathbf{D}$, $\mathbf{D}_1$ e $\mathbf{D}_2$ são conjuntos de valores reais tais que $\mathbf{D}_1 \cup \mathbf{D}_2 = \mathbf{D}$
  - Exemplo:
    - $\max\{10, -3, 0, 100\} = \max\{ \max\{10, -3\}, \max\{0, 100\} \} = 100$
  - Propriedade vale recursivamente (para $\max\{\mathbf{D}_1\}$ e $\max\{\mathbf{D}_2\}$)
- Utilidade para Complete Linkage
  - Dada a distância entre os grupos **A** e **B** e entre **A** e **C**
    - É trivial calcular a distância entre **A** e (**B** $\cup$ **C**) !

Exemplo de Complete Linkage:

- Seja a seguinte matriz de distâncias iniciais ($\mathbf{D}_1$) entre 5 objetos :

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} & & & & \\ 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

- Execução de complete linkage através de sucessivas atualizações da matriz de distâncias (uso da propriedade anterior):
  - No quadro...

Exercícios:

- Converta a matriz de similaridades abaixo em uma matriz de dissimilaridades através da transformação d = 1 – s e em seguida execute complete linkage sobre esta matriz:

S =

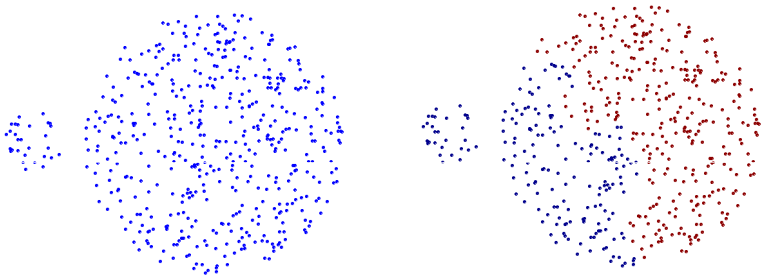
	1	2	3	4	5
1	1,00	0,90	0,10	0,65	0,20
2	0,90	1,00	0,70	0,60	0,50
3	0,10	0,70	1,00	0,40	0,30
4	0,65	0,60	0,40	1,00	0,80
5	0,20	0,50	0,30	0,80	1,00

Mostre o dendrograma completo resultante

- Execute o agora o algoritmo sobre a **matriz S original**, modificando o que for necessário. Converta similaridades em dissimilaridades apenas para mostrar o dendrograma

Características do Complete Linkage

• Menos sensível a ruído, mas…



objetos **corte em 2 clusters**

•Tende a "quebrar" clusters grandes (vide figura), é sensível a outliers e tende a formar clusters globulares

Como Comparar os Clusters?

- **Average Linkage** (ou Group Average)
  - Dissimilaridade entre *clusters* é dada pela dissimilaridade média entre cada par de objetos (um de cada *cluster*)

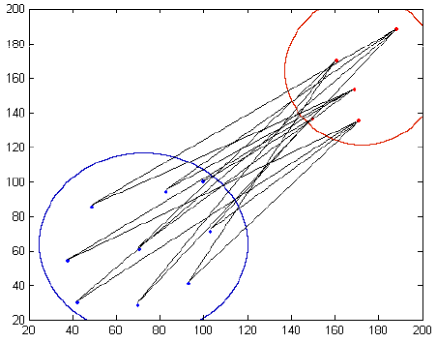


Figura por Lucas Vendramin

Nota Importante

- Para fins de atualização da matriz de (dis)similaridade em **average linkage**, o cálculo da (dis)similaridade entre um novo cluster (dado pela união de outros dois) e os demais deve considerar o no. de objetos em cada cluster envolvido
  - já que average linkage calcula uma média !
- Especificamente, sendo $|\mathbf{C}_i|$ o número de objetos em um cluster $\mathbf{C}_i$ e $d(\mathbf{C}_i,\mathbf{C}_j)$ a (dis)similaridade entre dois clusters $\mathbf{C}_i$ e $\mathbf{C}_j$, é simples deduzir que:

$$d(\mathbf{C}_i, \mathbf{C}_j \cup \mathbf{C}_k) = \frac{|\mathbf{C}_j|}{|\mathbf{C}_j|+|\mathbf{C}_k|} d(\mathbf{C}_i, \mathbf{C}_j) + \frac{|\mathbf{C}_k|}{|\mathbf{C}_j|+|\mathbf{C}_k|} d(\mathbf{C}_i, \mathbf{C}_k)$$

Exemplo de Average Linkage:

- Seja a seguinte matriz de distâncias iniciais ($\mathbf{D}_1$) entre 5 objetos :

$$\mathbf{D}_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

- Execução de average linkage através de sucessivas atualizações da matriz de distâncias (uso da propriedade anterior):
  - No quadro...

Exercício:

- Converta a matriz de similaridades abaixo em uma matriz de dissimilaridades através da transformação d = 1 – s e em seguida execute average linkage sobre esta matriz:

S =

	1	2	3	4	5
1	1,00	0,90	0,10	0,65	0,20
2	0,90	1,00	0,70	0,60	0,50
3	0,10	0,70	1,00	0,40	0,30
4	0,65	0,60	0,40	1,00	0,80
5	0,20	0,50	0,30	0,80	1,00

Mostre o dendrograma completo resultante

Comparação

- **Single** x **Complete** x **Average**:
  - Single Linkage:
    - Capaz de detectar clusters de formas complexas
    - No entanto, muito sensível a ruído nos dados (e.g. "pontes")

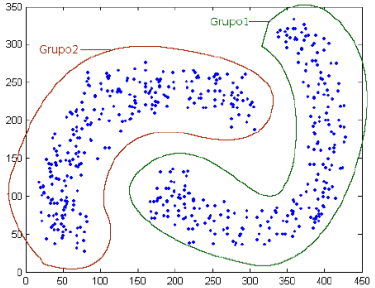


Figura por Lucas Vendramin

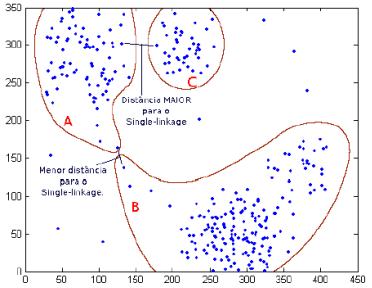


Figura por Lucas Vendramin

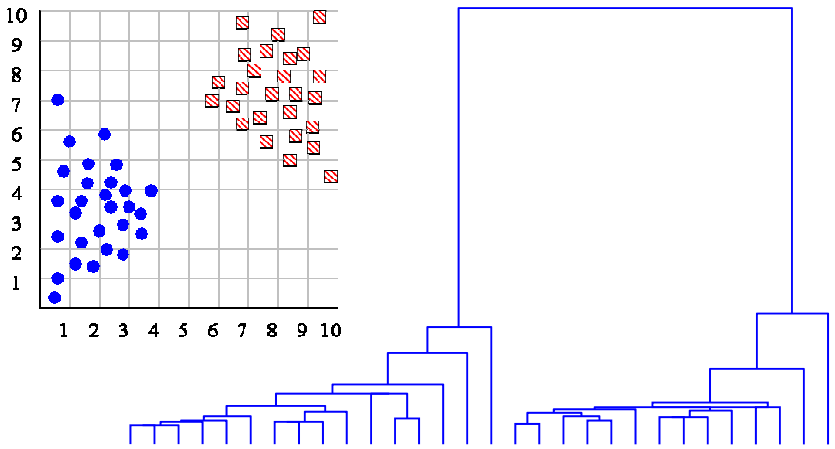
Comparação

- **Single** x **Complete** x **Average**:
  - Complete Linkage:
    - Reduz sensibilidade a ruído (e.g. pontes entre clusters)
    - No entanto:
      - aumenta risco de separar clusters grandes
      - perde capacidade de detecção de formas complexas
        - favorece clusters globulares
  - Average Linkage:
    - Também favorece clusters bem comportados (globulares)
    - Mas é menos sensível (mais robusto) a ruído e outliers !

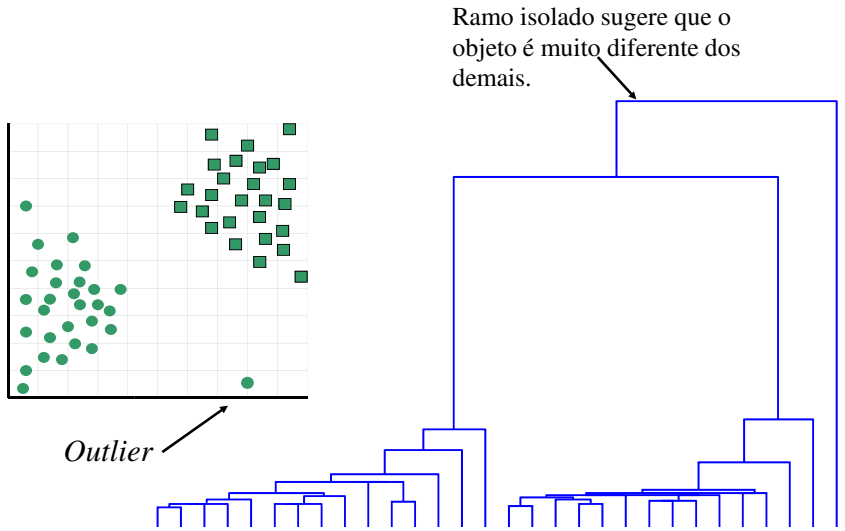
Importância do Dendrograma

- Dendrogramas constituem por si sós uma técnica de **visualização de dados multi-dimensionais**
- Podem ser úteis para análise exploratória de dados
  - estimativa do número natural de grupos
  - detecção de outliers
  - combinação com outras técnicas de visualização
  - ...

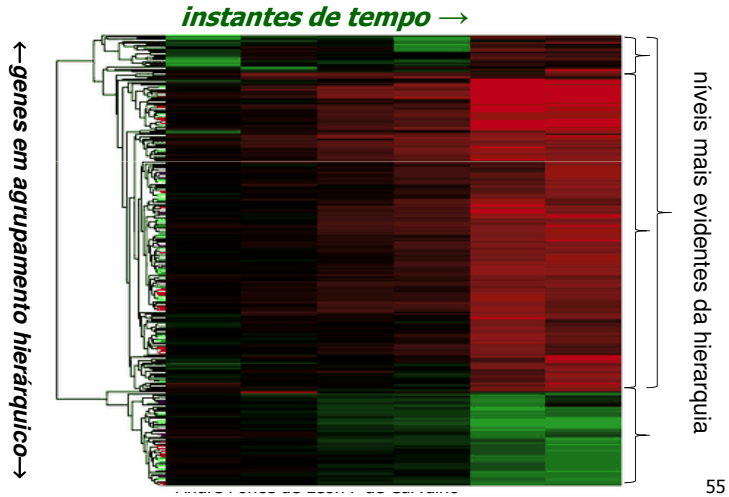
Pode-se examinar o dendrograma para tentar estimar o número *mais natural* de clusters. No caso abaixo, existem duas sub-árvores bem separadas, sugerindo dois grupos de dados. Na prática, porém, as distinções nem sempre são tão simples…



Pode-se usar o dendrograma para tentar detectar *outliers:*



Dendrogramas e Heatmaps



Como Plotar o Dendrograma ?

- Para muitos objetos, ilustração manual é inviável
- Gerar o gráfico automaticamente demanda ordenar de forma apropriada os objetos no eixo horizontal
- Algoritmo Recursivo Simples:
  - Iniciar com o topo da hierarquia (grupo único)
  - Dividir o eixo horizontal em 2 subintervalos e colocar em cada um os objetos de cada um dos 2 grupos que derivam do grupo único
  - Executar recursivamente o passo anterior para cada subintervalo

Métodos Divisivos

- Iniciam com um único *cluster*, que é sub-dividido em 2
- Recursivamente sub-divide cada um dos 2 *clusters*
  - Até que cada objeto constitua um **singleton**
- Em geral, são menos usados do que os aglomerativos:
  - É mais simples unir 2 *clusters* do que dividir...
    - número de modos para dividir N objetos em 2 *clusters* é $(2^{N-1} - 1)$. Por exemplo, para N=50 existem 5.63×10^{14} maneiras de se obter dois *clusters* !
- Questão:
  - *Como dividir um cluster ?*

- Heurística de MacNaughton-Smith et al. (1964):
  - Para um dado *cluster*, escolher o objeto mais distante dos demais
    - Este formará o *novo cluster*
  - Para cada objeto, calculam-se as distâncias médias deste aos objetos do *cluster* original e aos objetos do *novo cluster,*
  - O objeto mais próximo do *novo cluster* e mais distante do *cluster* original é transferido para o *novo cluster* ; e repete-se o processo
- Exemplo (Everitt et al., 2001):

$$\mathbf{D} = \begin{array}{c|ccccccc} 1 & 0 & & & & & & \\ 2 & 10 & 0 & & & & & \\ 3 & 7 & 7 & 0 & & & & \\ 4 & 30 & 23 & 21 & 0 & & & \\ 5 & 29 & 25 & 22 & 7 & 0 & & \\ 6 & 38 & 34 & 31 & 10 & 11 & 0 & \\ 7 & 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{array}$$

- Para este exemplo, objeto "1" é o mais distante (*novo cluster* – A)
- Demais objetos permanecem no *cluster principal* (*cluster* – B)
- *Clusters* obtidos: A={1} e B={2,3,4,5,6,7}
- Sejam D_A e D_B as distâncias médias de um objeto de B em relação aos objetos de A e B, respectivamente:

Mais próximos de A do que de B

Objetos B	D_A	D_B	D_B-D_A
2	10	25	15,0
3	7	23,4	16,4
4	30	14,8	-15,2
5	29	16,4	-12,6
6	38	19,0	-19,0
7	42	22,2	-19,8

Objeto escolhido para mudar de *cluster*

Desta forma, obtemos os *clusters* {1,3} e {2,4,5,6,7}

Repetindo o processo temos ...

Objetos B	D_A	D_B	D_B-D_A
2	8,5	29,5	12,0
4	25,5	13,2	-12,3
5	25,5	15,0	-10,5
6	34,5	16,0	-18,5
7	39,0	18,7	-20,3

Mudar para A

Novos *clusters*: {1,3,2} e {4,5,6,7}.

Próximo passo: todos (D_B-D_A) negativos;

Pode-se então repetir o processo em cada *cluster* acima, separadamente...

Métodos Divisivos

- **Exercício**:
  - Aplicar o algoritmo hierárquico divisivo com heurística de MacNaughton-Smith et al. Na seguinte base de dados:

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

Sumário dos Métodos Hierárquicos

• **No. de Clusters**: não necessitam especificar o número de clusters *a priori*, mas de qualquer forma é necessário selecionar *a posteriori* ...

• **Procedimento Guloso**: não se pode reparar o que foi feito num passo anterior – não necessariamente leva à solução ótima

• **Escalabilidade**: complexidade computacional é, no mínimo, quadrática em função do número de objetos

• **Interpretabilidade**: Produz uma hierarquia, que é aquilo desejado em muitas aplicações (e.g. taxonomia), e permite análise de outliers

• **Cálculo Relacional**: Não demandam matriz de dados original