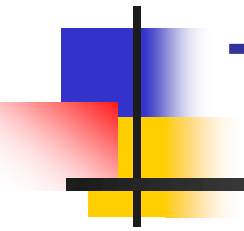
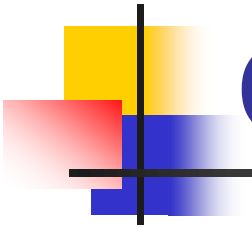


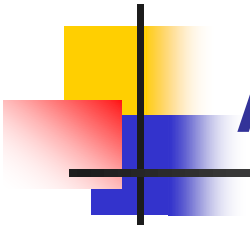
Tópicos Avançados em IA

Prof. Eduardo R. Hruschka



Créditos

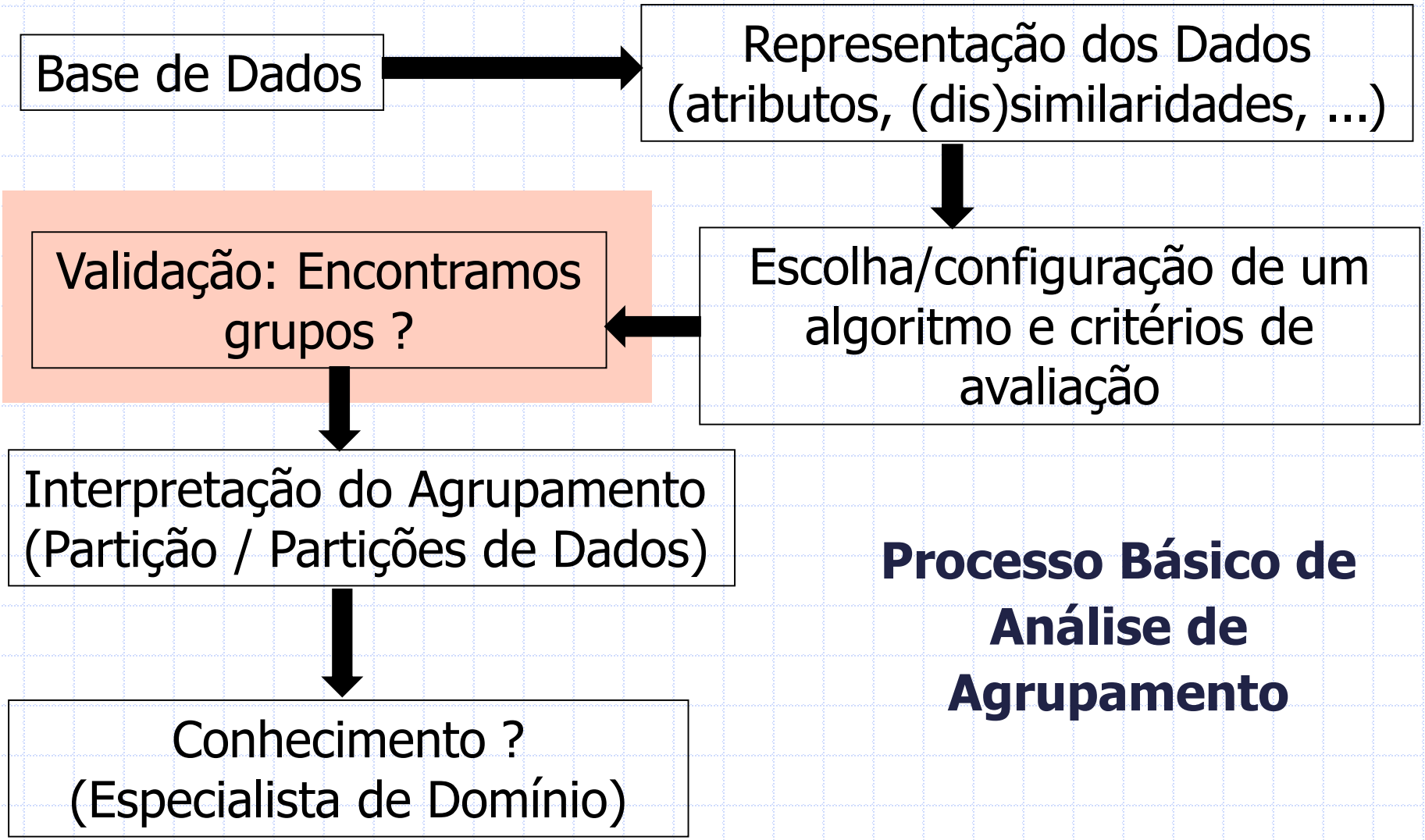
- Este material consiste de adaptações e extensões dos originais elaborados por Eduardo R. Hruschka e Ricardo J. G. B. Campello



Aula de Hoje

- Validação de Agrupamento
- Critérios de Validade de Agrupamento
  - Critérios Externos
  - Critérios Internos e Relativos
  - Avaliação de Partições

Relembrando...



Processo Básico de Análise de Agrupamento

Frase do dia :

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Jain and Dubes, *Algorithms for Clustering Data*, 1988

Validação de Agrupamento

- **Validação** é um termo que se refere de forma ampla aos diferentes procedimentos para avaliar de maneira objetiva e quantitativa os resultados de análise de agrupamento
- Cada um desses procedimentos pode nos ajudar a responder uma ou mais questões do tipo:
  - Encontramos grupos de fato ?
    - grupos são pouco usuais ou facilmente encontrados ao acaso ?
  - Qual a qualidade (relativa ou absoluta) dos grupos encontrados ?
  - Qual é o número natural / mais apropriado de grupos ?

Validação de Agrupamento

- A maneira quantitativa com que se dá um procedimento de validação é alcançada através de algum tipo de **índice**
  - **Índice ou Critério de Validade** (de agrupamento)
- Tais índices / critérios podem ser de três tipos
  - **Externos**: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento esperada ou conhecida
  - **Internos**: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados
  - **Relativos**: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa

Critérios de Validade Externos

- Embora o problema de *clustering* seja não supervisionado, em alguns cenários o resultado de agrupamento desejado pode ser conhecido. Por exemplo:
  - Reconhecimento visual dos clusters naturais (bases 2D, 3D)
  - Especialista de domínio
  - Bases geradas sinteticamente com distribuições conhecidas
    - *Benchmark data sets*
  - Bases de classificação sob a hipótese que classes são clusters
- Medem o nível de compatibilidade entre uma partição obtida e uma partição de referência dos mesmos dados

Critérios de Validade Externos

- Existem vários critérios externos na literatura:
  - Rand Index
  - Jaccard
  - Rand Index Ajustado
  - Fowlkes-Mallows
  - Estatística Γ
  - Normalized Mutual Information
  - ...
- Discutiremos apenas o Rand Index.

Critérios de Validade Externos

- Os critérios que veremos são baseados na comparação de pares de objetos das partições em questão
- Por conveniência, adotaremos a seguinte terminologia:
  - grupos da **partição de referência** (golden truth) $\rightarrow$ "**classes**"
  - grupos da **partição sob avaliação** $\rightarrow$ **clusters (grupos)**
- Podemos então definir as grandezas de interesse:

- **a**: No. de pares que pertencem à mesma classe e ao mesmo cluster
- **b**: No. de pares que pertencem à mesma classe e a clusters distintos
- **c**: No. de pares que pertencem a classes distintas e ao mesmo cluster
- **d**: No. de pares que pertencem a classes e clusters distintos

Rand Index

$$RI = \frac{a+d}{a+b+c+d}$$

Número de pares de objetos:

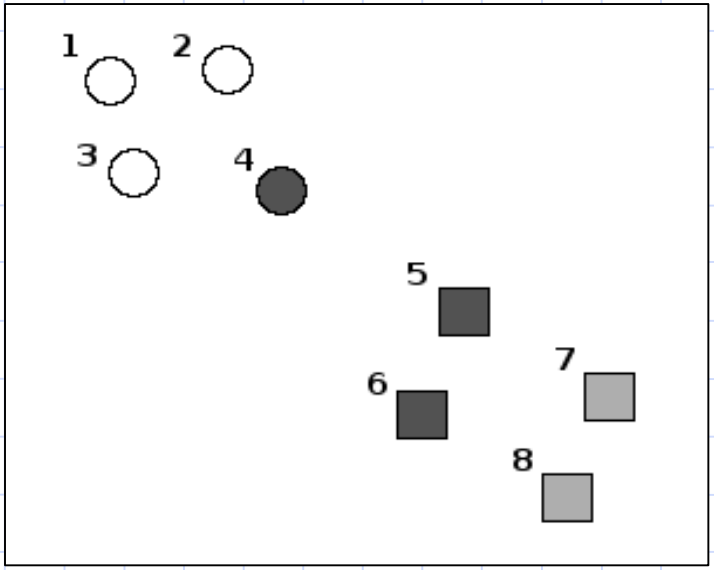
a: da mesma classe e do mesmo cluster (grupo)

b: da mesma classe e de clusters distintos

c: de classes distintas e do mesmo cluster

d: de classes distintas e de clusters distintos

Figura por Lucas Vendramin



2 Classes (Círculos e Quadrados)
3 Clusters (Preto, Branco e Cinza)

a = 5; b = 7; c = 2; d = 14

RI = 5+14/(5+7+2+14) = 0.6785

Rand Index

- O índice de Rand possui algumas limitações.
- A principal delas é o **viés** de favorecer a comparação de partições com níveis mais elevados de granularidade
  - Valores mais elevados ao comparar partições com mais grupos
- Razão Essencial:
  - mesmo peso para objetos agregados (termo **a**) ou separados (**d**)
  - termo **d** tende a dominar o índice
  - quanto mais grupos, mais pares pertencem a grupos distintos
    - isso é válido em qualquer uma das duas partições.
    - probabilidade / incidência de pares em comum é maior.

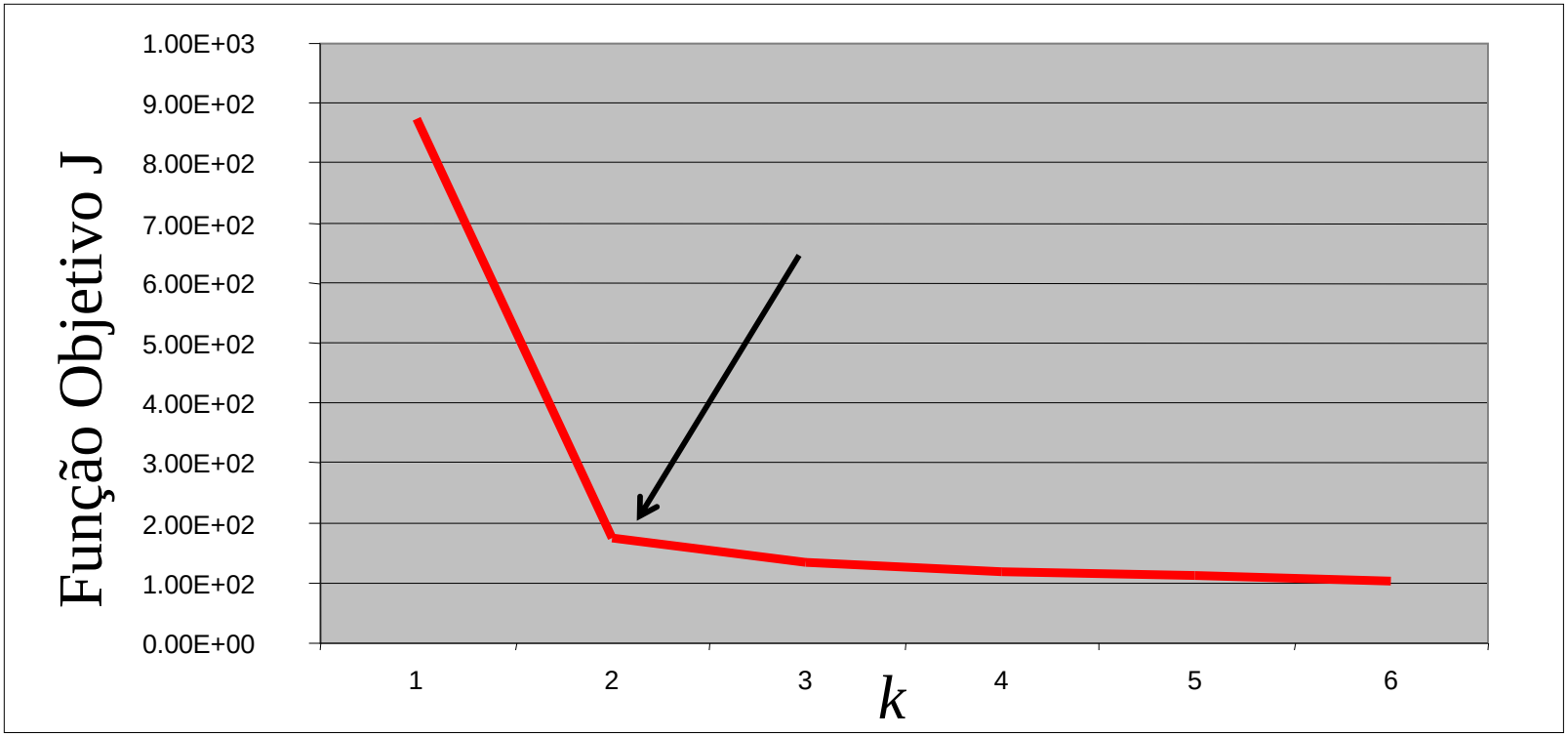
Critérios de Validade Internos

- Na prática, normalmente não se dispõe de uma partição ou hierarquia de referência.
  - temos apenas os dados e o resultado a ser avaliado...
- Critérios que avaliam a estrutura de grupos obtida utilizando apenas os próprios dados são denominados **critérios internos de validade** de agrupamento
- Já vimos um exemplo ao estudar o k-means – **SSE**:

$$J = \sum_{c=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_c} d\left(\mathbf{x}_j, \overline{\mathbf{x}}_c\right)^2$$

Critérios de Validade Internos

- Já vimos que o SSE pode também ser usado para auxiliar a responder uma das questões fundamentais em validação: k^* ?
- Por exemplo, via múltiplas execuções de um algoritmo particional



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

Critérios de Validade Internos

- A detecção automática do "joelho" pode se tornar mais simples se for possível transformá-lo em um pico:

$$\Delta J(k) = abs\left(\frac{J(k-1) - J(k)}{J(k) - J(k+1)}\right)$$

- Exemplo:

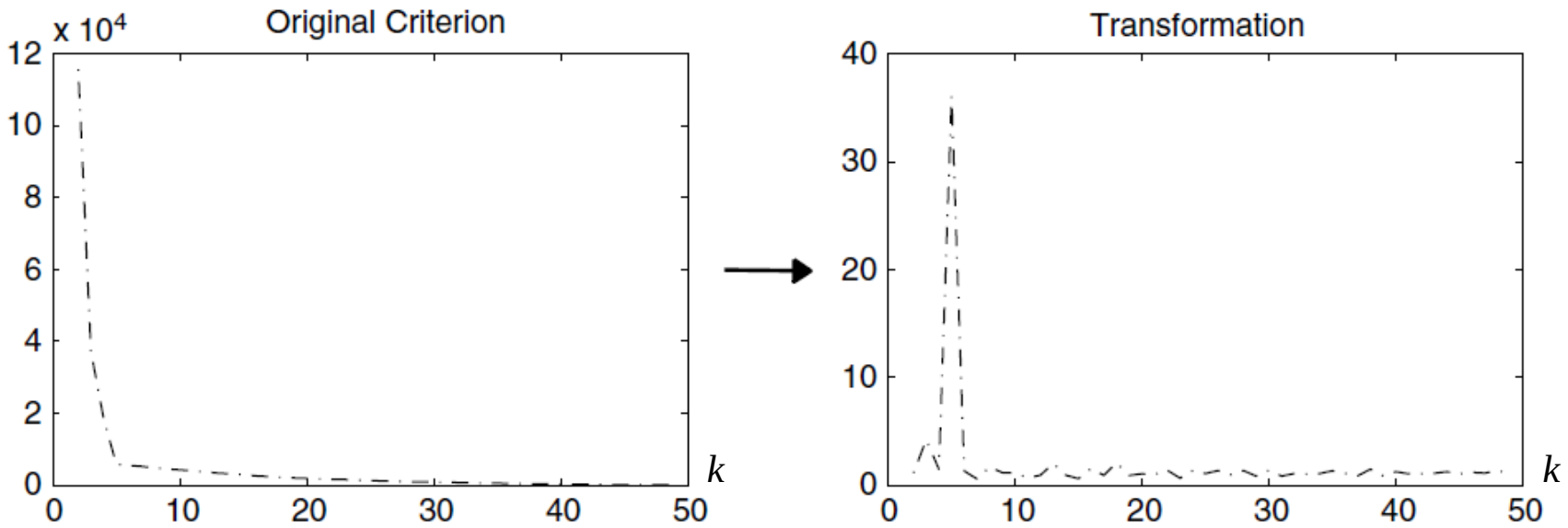


Figura por Lucas Vendramin

Critérios de Validade Relativos

- O termo **critério relativo** se refere a uma classe particular de critérios com habilidade para indicar qual a melhor dentre duas ou mais partições
  - O termo normalmente é associado a critérios internos
- A caracterização como relativo pode não depender apenas do critério, mas eventualmente do contexto
  - Por exemplo, o SSE é um critério relativo se as partições a serem comparadas possuem o mesmo no. de grupos
  - Para números de grupos distintos, os valores de SSE não são comensuráveis e o critério, portanto, não é relativo

Critérios de Validade Relativos

- Critérios relativos num contexto amplo são definidos aqui como aqueles capazes de:

  1. Avaliar individualmente uma única partição

  2. Quantificar esta avaliação através de um valor que possa ser comparado relativamente

- Como consequência, tais critérios são capazes de produzir uma ordenação de um conjunto de partições de acordo com suas avaliações

Critério da Largura de Silhueta

SWC = Silhueta média sobre todos os objetos: $$SWC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

Silhueta (i-ésimo objeto): $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$ (s(i) := 0 para singletons)

***a(i)*: dissimilaridade média do i-ésimo objeto ao seu cluster**

***b(i)*: dissimilaridade média do i-ésimo objeto ao cluster vizinho mais próximo**

Silhueta Original: $a(i)$ e $b(i)$ são calculados como a distância média (Euclidiana, Mahalanobis, etc) do i-ésimo objeto a todos os demais objetos do cluster em questão. Complexidade $O(N^2)$

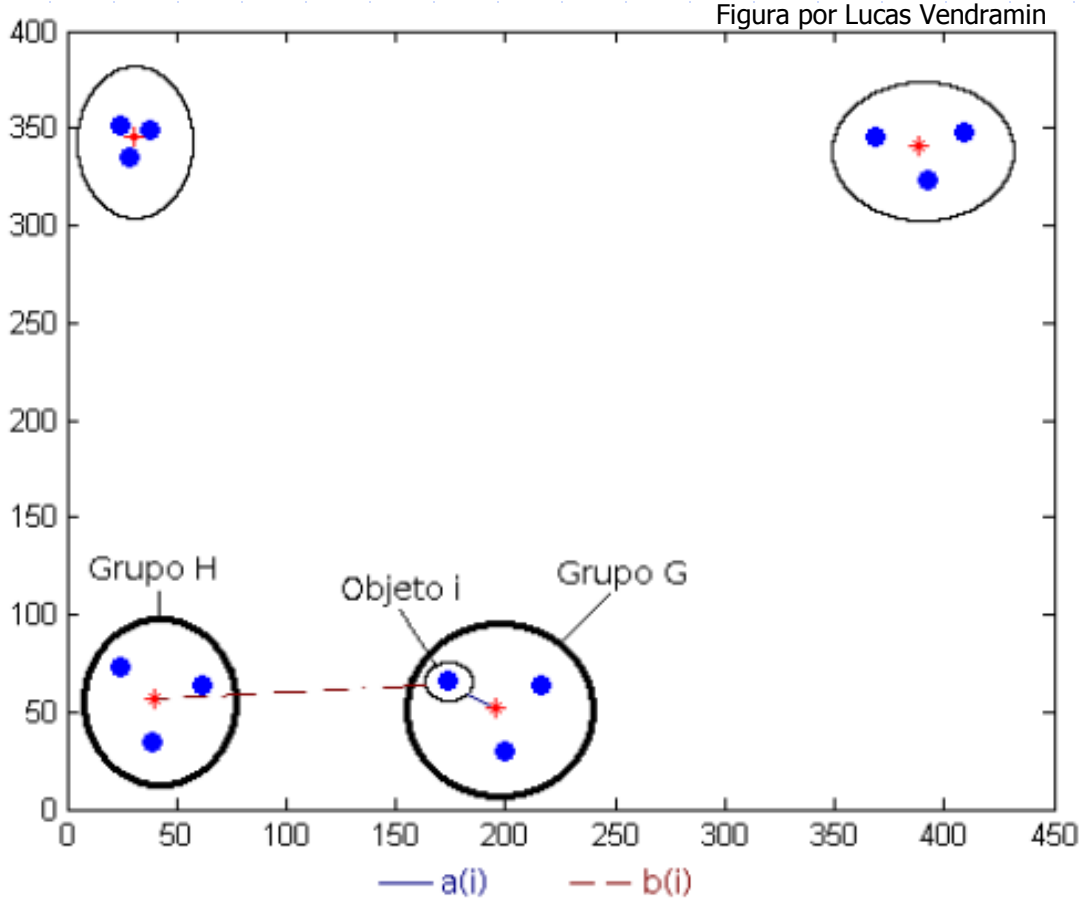
Propriedade Favorável: SWC $\in$ [–1 ,+1]

Silhueta Simplificada (SSWC)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$SWC = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

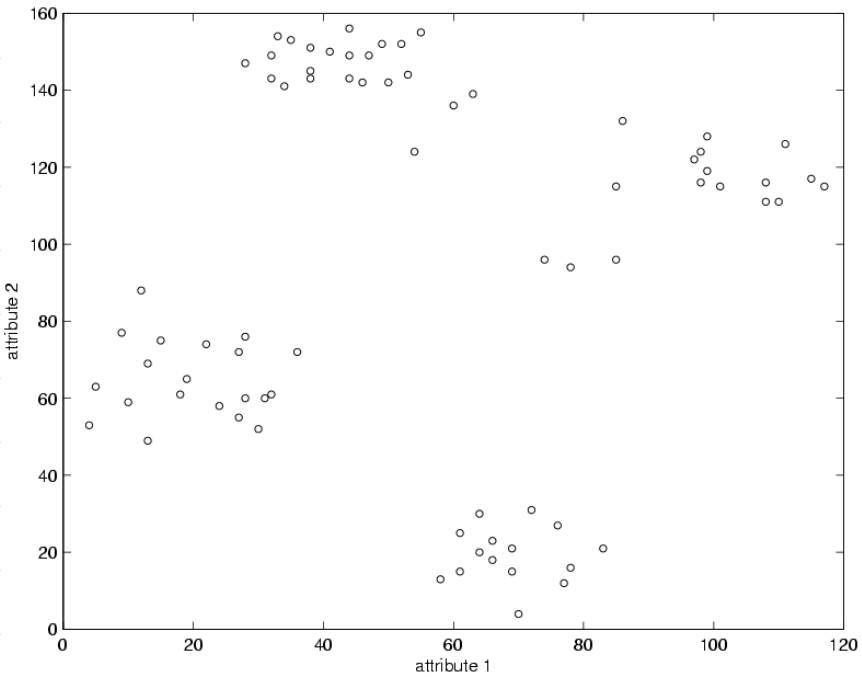
Figura por Lucas Vendramin

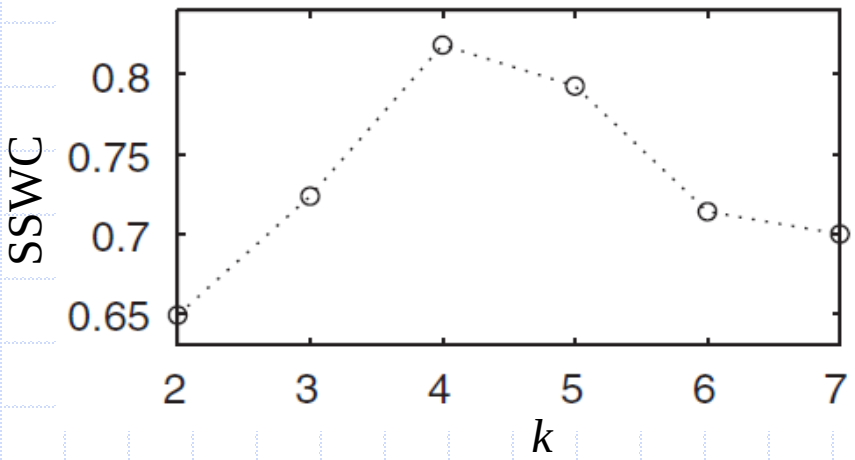


Silhueta Simplificada: $a(i)$ e $b(i)$ são calculados como a distância do i-ésimo objeto ao centróide do cluster em questão. Complexidade O(N).

Exemplo (SSWC)

- Relembrando a Subjetividade do Problema:
  - Quantos grupos abaixo...? Quatro? Cinco? Seis?
  - Sob a perspectiva **deste critério** (SSWC) temos k^*=4





Muitos Outros Critérios...

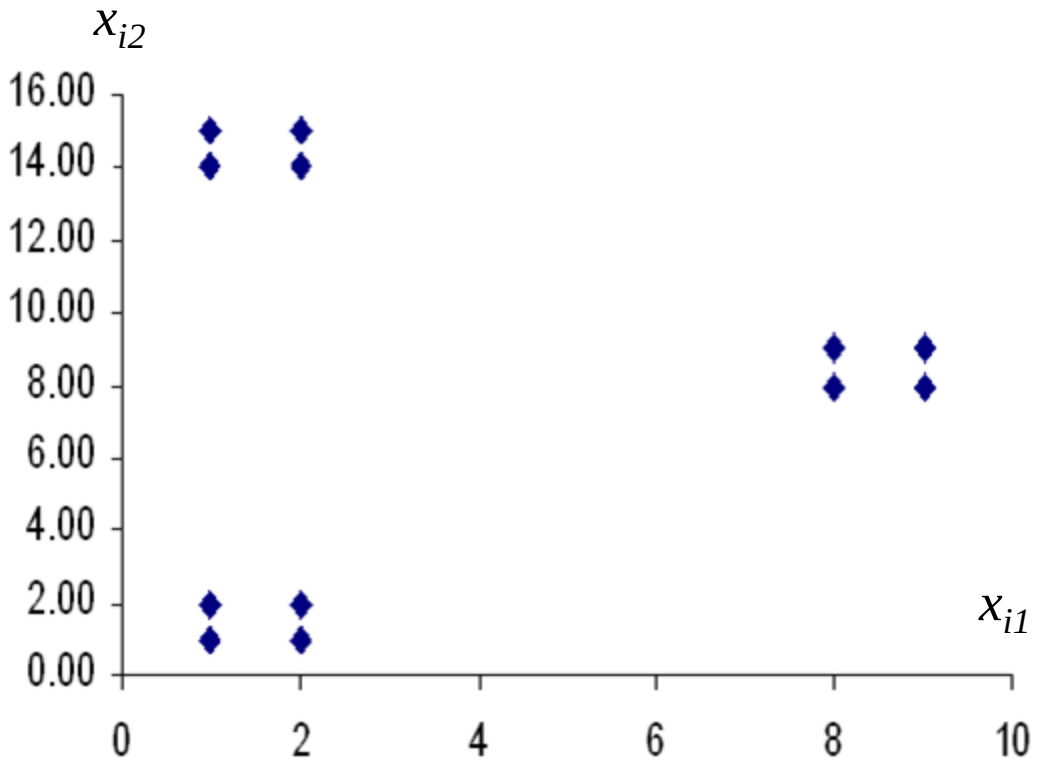
- **Variance Ratio Criterion** (VRC ou Calinski-Harabaz)
- **Point-Biserial**
- e muito mais...

-	Criterion	Complexity
	Calinski-Harabasz (VRC)	$O(nN)$
	Davies-Bouldin (DB)	$O(n(k^2+N))$
	Dunn	$O(nN^2)$
	Silhouette Width Criterion (SWC)	$O(nN^2)$
	Alternative Silhouette (ASWC)	$O(nN^2)$
	Simplified Silhouette (SSWC)	$O(nNk)$
	Alternative Simplified Silhouette (ASSWC)	$O(nNk)$
	PBM	$O(n(k^2+N))$
	C-Index	$O(N^2(n+log_2N))$
	Gamma	$O(nN^2+N^4/k])$
	G(+)	$O(nN^2+N^4/k])$
	Tau	$O(nN^2+N^4/k])$
	Point-Biserial	$O(nN^2)$
	$C/\sqrt{k}$	$O(nN)$
*	Trace(W)	$O(nN)$
*	Trace(CovW)	$O(nN)$
*	Trace($W^{-1}B$)	$O(n^2N+n^3)$
*	|T|/|W|	$O(n^2N+n^3)$
*	Nlog(|T|/|W|)	$O(n^2N+n^3)$
*	k^2W	$O(n^2N+n^3)$
*	log(SSB/SSW)	$O(n(k^2+N))$
*	Ball-Hall	$O(nN)$
*	McClain-Rao	$O(nN^2)$

Vendramin, L., Campello, R. J. G. B. & Hruschka, E. R. "*Relative Clustering Validity Criteria: A Comparative Overview*" **Statistical Analysis and Data Mining**, Vol. 3, p. 209-235, 2010

Exercício

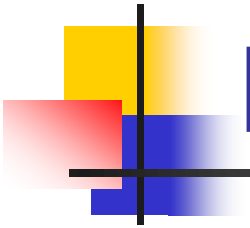
Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



- Calcular o valor da silhueta para a partição natural dos dados acima e também para outras partições (aleatórias). Compare os resultados.

Leitura Recomendada

- Vendramin, L. , Campello, R. J. G. B. , Hruschka, E. R. "Relative Clustering Validity Criteria: A Comparative Overview" *Statistical Analysis and Data Mining*, Wiley, Vol. 3, p. 209-235, 2010



Referências

- Jain, A. K. & Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
- Höppner, F., Klawonn, F., Kruse, R., Runkler, T., Fuzzy Cluster Analysis, 1999
- Milligan, G. W. & Cooper, M. C. "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, Vol. 50, No. 2, 159-179, 1985
- Vendramin, L. , Campello, R. J. G. B. , Hruschka, E. R. "Relative Clustering Validity Criteria: A Comparative Overview" *Statistical Analysis and Data Mining*, Wiley, Vol. 3, p. 209-235, 2010