SCC5895 – Análise de Agrupamento de Dados

Validação de Agrupamento: Parte I

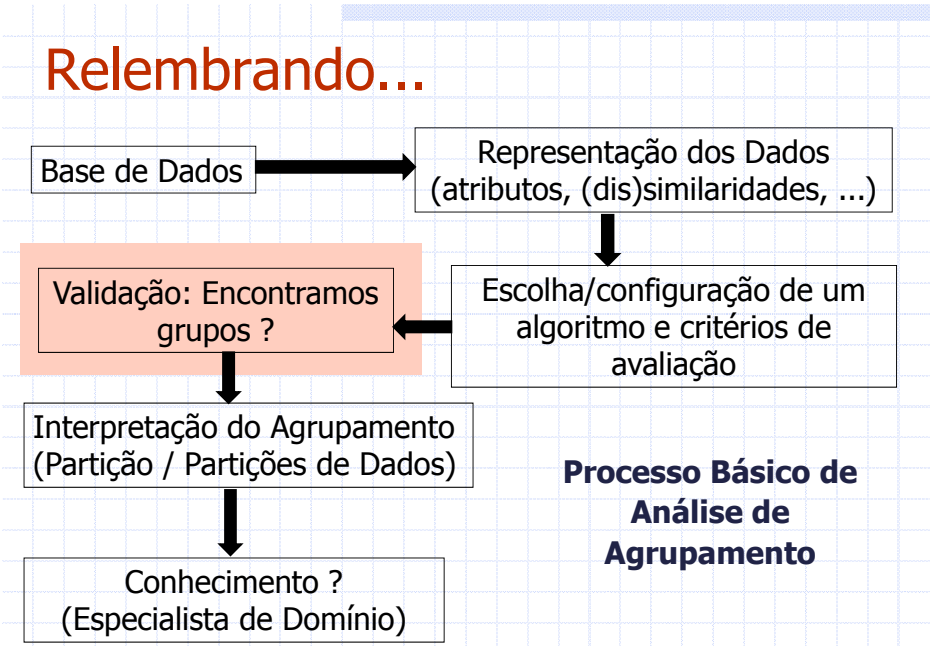
Prof. Ricardo J. G. B. Campello

PPG-CCMC / ICMC / USP

Aula de Hoje

- Validação de Agrupamento
- Critérios de Validade de Agrupamento
  - Critérios Externos
  - Critérios Internos e Relativos
    - Critérios de Otimização e Stopping Rules
  - Avaliação de Partições e de Hierarquias
  - Avaliação de Partições com Sobreposição

Relembrando...



Processo Básico de Análise de Agrupamento

Baseado no original do Prof. Eduardo R. Hruschka

Comment on Cluster Validity

"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:

próxima aula
- Encontramos grupos de fato ?
  - grupos são pouco usuais ou facilmente encontrados ao acaso ?

aula de hoje
- 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** de agrupamentos

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
- Índices que medem o nível de compatibilidade entre uma partição obtida e uma partição de referência dos mesmos dados são denominados **critérios de validade externos**

Critérios de Validade Externos

- Existem vários critérios externos na literatura:
  - Rand Index
  - Jaccard
  - Rand Index Ajustado
  - Fowlkes-Mallows
  - Estatística Γ
  - Normalized Mutual Information
  - ...
- Discutiremos a seguir aqueles mais conhecidos e utilizados
  - iniciando com a avaliação de partições, depois de hierarquias

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) → "**classes**"
  - grupos da **partição sob avaliação** → **clusters**
- 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

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

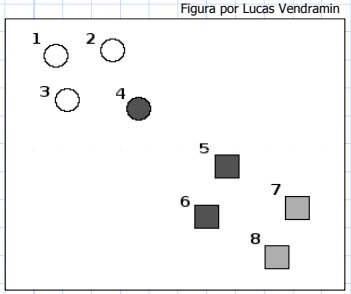
a: No. de pares que pertencem à mesma classe e ao mesmo cluster

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 distintas e 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 sérias...
- 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...

Jaccard

Elimina o termo **d** sob a ótica de que um agrupamento é uma coleção de agregações de pares de objetos, separações sendo apenas uma consequência

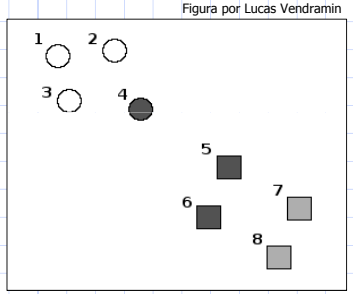
$$Jc = \frac{a}{a+b+c}$$

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

Figura por Lucas Vendramin



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

a = 5; b = 7; c = 2

Jc = 5/(5+7+2) = 0.3571

Rand Index Ajustado

- Limitação de Jaccard e também do Rand Index original:
  - **Não** são "adjusted for chance", i.e., o valor esperado não é nulo para 2 partições completamente aleatórias de um conj. de dados
  - Correção a ser feita:

$$\text{Critério\_Ajustado} = \frac{\text{Critério} - E\{\text{Critério}\}}{\text{Max\_Critério} - E\{\text{Critério}\}}$$

    - Problema é estimar o valor esperado teórico do critério, E{Critério}
  - Hubert & Arabie (1985) determinaram o valor esperado do índice de Rand, dando origem ao **Adjusted Rand Index** (ARI)

Rand Index Ajustado

- ARI pode ser escrito como:

$$ARI = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}}$$

onde M = N(N – 1)/2 = a + b + c + d

- **Exercício**:
  - calcular o valor do ARI referente ao mesmo par de partições anterior, usado como exemplo para ilustrar Rand e Jaccard

Rand Index Ajustado

- É evidente que ARI possui valor máximo 1
  - mas ao contrário de Rand e Jaccard, zero não é o valor mínimo, é o valor esperado ao comparar duas partições aleatórias
- Experimento (Jain & Dubes, 1988):
  - Quatro conjuntos de 100 pontos em 5 dimensões
    - dados estruturados (misturas de Gaussianas) e aleatórios (distrib. uniforme)
    - dados com 2 e 8 grupos (puramente arbitrários no caso de distrib. uniforme)
  - Dados agrupados com single- e complete-linkage (SL e CL)
  - Cortes realizados no número correto de grupos (2 ou 8)
  - Partições obtidas comparadas com os rótulos
  - Experimento repetido 100 vezes (simulação de Monte Carlo)

Experimento ARI x Jaccard
(Jain & Dubes, 1988)

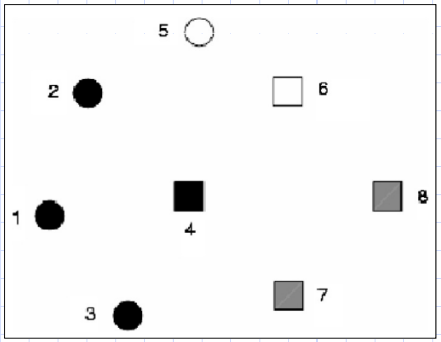
TABLE 4.8 Comparison of External Indices for Partitions

	Jaccard		Corrected Rand	
k^a	Mean	Std. Dev.	Mean	Std. Dev.
Clustered data				
SL				
2	0.934	0.169	0.870	0.336
8	0.597	0.172	0.683	0.169
CL				
2	0.989	0.043	0.988	0.050
8	0.859	0.116	0.908	0.082
Random data				
SL				
2	0.496	0.007	0.00027	0.005
8	0.118	0.004	0.00053	0.004
CL				
2	0.354	0.029	−0.00069	0.017
8	0.068	0.008	−0.00102	0.015

Note: [a] k is the true number of clusters for clustered data and the number of pseudo clusters for random data.

Exercício

Calcule o valor dos critérios de Rand, Rand Ajustado e Jaccard entre as duas partições (dos mesmos 8 objetos) ilustradas na figura abaixo. Uma das partições é representada por 3 cores, enquanto a outra por duas formas geométricas:



Validação Externa de Hierarquias

- Uma hierarquia de referência dificilmente está disponível, mas, se estiver, temos basicamente duas alternativas
- **Método Direto**
  - Aplica-se um critério externo N – 2 vezes, uma vez para cada nível intermediário das hierarquias em questão
  - Compõe-se de alguma forma o conjunto de valores obtidos (por exemplo, soma ou média)
- **Método Indireto**
  - Avalia-se a correlação entre duas matrizes que representam, de alguma forma, as hierarquias sendo comparadas

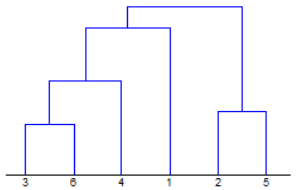
Método Indireto (Exemplo de Configuração)

- Correlação: **Estatística Γ de Hubert Normalizada**

$$\Gamma = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}[X(i,j)-\mu_X][Y(i,j)-\mu_Y]}{\sigma_X\sigma_Y} = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}[X(i,j)Y(i,j)] - \mu_X\mu_Y}{\left[\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}[X(i,j)]^2-\mu_X\right]^{\frac{1}{2}}\left[\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}[Y(i,j)]^2-\mu_Y\right]^{\frac{1}{2}}}$$

(Pearson aplicada a matrizes no contexto de agrupamento)

- Matrizes: elemento (i,j) é igual a "c" se os objetos i e j aparecem unidos pela 1a vez no c-ésimo nível hierárquico
- Exemplo:



$$\mathbf{Y} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 6 & 5 & 5 & 6 & 5 \\ & 6 & 6 & 2 & 6 \\ & & 4 & 6 & 2 \\ & & & 6 & 4 \\ & & & & 6 \\ & & & & \end{bmatrix}$$

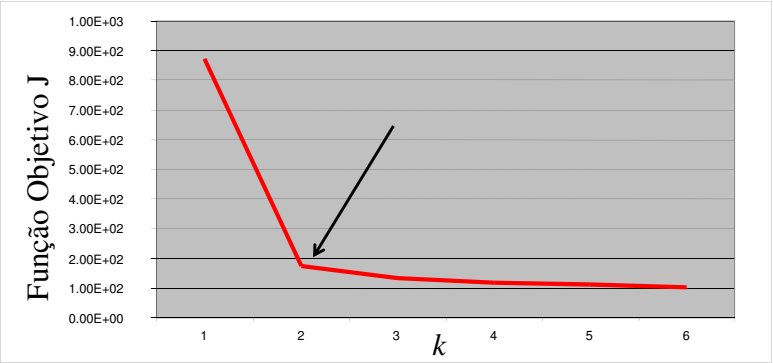
Critérios de Validade Internos

- De maneira geral, em análise de agrupamento prática, normalmente não se dispõe de uma partição ou hierarquia de referência para validar a estrutura de grupos obtida
  - 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 !**

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

Critérios de Validade Internos

- Já vimos que o SSE pode ser usado para auxiliar a responder uma das questões fundamentais em validação: qual é o número de grupos 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:

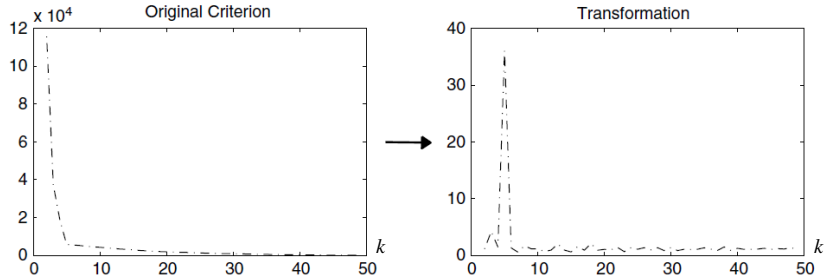


Figura por Lucas Vendramin

Critérios de Validade Internos

- **Avaliação de Hierarquias**:
  - Note que o critério pode também ser aplicado às sucessivas partições produzidas por um algoritmo hierárquico
  - O resultado pode indicar o ponto de corte do dendrograma
  - Nesse caso, o critério é denominado de **regra de parada**
    - **Stopping Rule**
- Existem diversas regras de parada para interrupção precoce da construção ou corte a posteriori de hierarquias
- Veremos a seguir mais um exemplo

Critérios de Validade Internos

- **Regra de Parada (Stopping Rule) de Hartigan (1975):**

$$H(k) = \left(\frac{J(k)}{J(k+1)} - 1\right) \times (N - k - 1)$$

  - "**Rule of Thumb**":
    - Valores excedendo 10 justificam aumentar o no. de grupos de k para k + 1
    - Logo, hierarquia deve ser interrompida no menor k tal que $H(k) \leq 10$
- Trata-se de uma heurística com certa motivação estatística
  - Praticamente obsoleta atualmente
    - Entenderemos as razões ao estudar índices de validade internos **relativos** que também podem ser aplicados como stopping rules, de forma mais eficaz

Critérios de Validade Internos

- **Avaliação de Hierarquias Completas:**
  - Em alguns casos queremos avaliar toda uma hierarquia obtida, não apenas decidir o ponto de interrupção do dendrograma
  - Sabemos que podemos utilizar uma medida de correlação (e.g. Hubert / Pearson) se uma hierarquia esperada estiver disponível
  - Mas e se uma hierarquia externa não estiver disponível ?
  - Podemos medir a correlação entre a **matriz de dados** e a ***cophenetic matrix*** obtida do dendrograma sob avaliação
    - Qual a interpretação desta medida ?

Critérios de Validade Internos

- **Exercício:**
  - Execute o algoritmo single-linkage na matriz de distâncias abaixo, obtenha a *cophenetic matrix* correspondente e calcule a estatística de Hubert normalizada entre essas duas matrizes

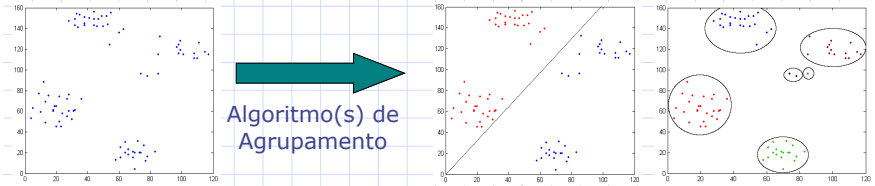
$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & 2 & 9 & 15 \\ 2 & 0 & 7 & 12 \\ 9 & 7 & 0 & 4 \\ 15 & 12 & 4 & 0 \end{bmatrix}$$

  - Substitua a *cophenetic matrix* por uma matriz da mesma natureza gerada aleatoriamente e repita o cálculo

Critérios de Validade Relativos

A aplicação prática de um ou mais algoritmos de agrupamento usualmente retorna múltiplas soluções distintas que precisam ser comparadas

- Algoritmos hierárquicos,
- Múltiplas execuções de k-means, …



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

- Note, porém, que a regra de parada (stopping rule) dada pelo "joelho" de SSE ou pelo pico da variação ΔJ podem ser considerados relativos sob certo aspecto
- De fato, tecnicamente, qualquer stopping rule para corte de hierarquias pode ser considerado como um critério "relativo" num contexto específico
  - no sentido que são capazes de sugerir que uma partição da hierarquia é mais apropriada que as demais
- No entanto, stopping rules não são critérios relativos num sentido amplo
  - não são capazes de comparar um par qualquer de partições

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

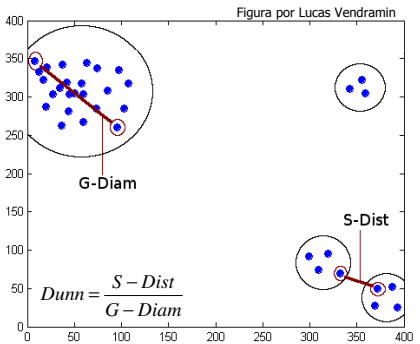
- Como consequência, tais critérios são capazes de produzir uma ordenação de um conjunto de partições de acordo com suas avaliações

Critérios de Validade Relativos

- Critérios relativos no contexto amplo definido anteriormente são mais flexíveis, pois:
  - Podem ser utilizados como **critérios de otimização**
  - Também podem ser utilizados como **stopping rules**
- Existem dezenas de tais critérios na literatura
- Estudos apontam alguns deles como superiores em algumas classes de problemas comuns na prática
  - Para problemas em geral, no entanto, não há qualquer garantia que um dado critério será o mais apropriado (**No free lunch !!!**)
- Veremos alguns critérios a seguir

Índice de Dunn

- Razão entre a menor distância inter-grupos (S-Dist) e a maior distância intra-grupo (G-Diam)
- Exemplo:



Índice de Dunn

- No índice original, as distâncias inter-grupos e intra-grupos são calculadas segundo os critérios de vinculação simples e diâmetro máximo
  - Muito sensível a ruído e outliers !
- Formalmente:

$$DN = \min_{\substack{p,q \in \{1,\ldots,k\} \\ p \neq q}} \left\{ \frac{\delta_{p,q}}{\max_{l \in \{1,\ldots,k\}} \Delta_l} \right\}$$

$$\delta_{p,q} = \min_{\mathbf{x}_i \in \mathbf{C}_p} \{ \min_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j|| \}$$

$$\Delta_l = \max_{\mathbf{x}_i \in \mathbf{C}_l} \{ \max_{\mathbf{x}_j \in \mathbf{C}_l} ||\mathbf{x}_i - \mathbf{x}_j|| \}$$

  - Complexidade $O(N^2)$

Variantes do Índice de Dunn

- Para reduzir a sensibilidade do índice, variantes para dist. inter- e intra grupos foram propostas
- Variantes para distância inter-grupos (separação):

$$\delta_{p,q} = \max_{\mathbf{x}_i \in \mathbf{C}_p, \mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j|| \longrightarrow \text{vinculação máxima (complete linkage)}$$

$$\delta_{p,q} = \frac{1}{N_p N_q} \sum_{\mathbf{x}_i \in \mathbf{C}_p} \sum_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j|| \longrightarrow \text{vinculação média (average linkage)}$$

$$\delta_{p,q} = ||\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_q|| \longrightarrow \text{distância entre centróides}$$

$$\delta_{p,q} = \frac{1}{N_p + N_q} \left(\sum_{\mathbf{x}_i \in \mathbf{C}_p} ||\mathbf{x}_i - \bar{\mathbf{x}}_q|| + \sum_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_j - \bar{\mathbf{x}}_p|| \right) \longrightarrow \text{mistura dos 2 anteriores}$$

$$\delta_{p,q} = \max \left\{ \max_{\mathbf{x}_i \in \mathbf{C}_p} \min_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j||, \max_{\mathbf{x}_j \in \mathbf{C}_q} \min_{\mathbf{x}_i \in \mathbf{C}_p} ||\mathbf{x}_i - \mathbf{x}_j|| \right\} \longrightarrow \text{métrica de Hausdorff}$$

Variantes do Índice de Dunn

- Variantes para distância intra-grupos (diâmetro):

$$\Delta_l = \frac{1}{N_l(N_l - 1)} \sum_{(\mathbf{x}_i \neq \mathbf{x}_j) \in \mathbf{C}_l} ||\mathbf{x}_i - \mathbf{x}_j|| \longrightarrow$$

Distância média entre todos os $N_l \times (N_l - 1) / 2$ objetos do l-ésimo grupo. Nota: constante 2 não aparece pois a somatória foi escrita de forma que trata os pares de objetos de maneira ordenada

$$\Delta_l = \frac{2}{N_l} \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \bar{\mathbf{x}}_l|| \longrightarrow$$

2 × raio do grupo, sendo raio definido como a distância média entre os objetos e o centróide

Critério Davies-Bouldin

$$DB = \frac{1}{k} \sum_{l=1}^{k} D_l$$

where $D_l = max_{l \neq m}\{D_{l,m}\}$. Term $D_{l,m}$ is the within-to-between cluster spread for the lth and mth clusters, i.e. $D_{l,m} = (\bar{d}_l + \bar{d}_m)/d_{l,m}$, where $\bar{d}_l$ and $\bar{d}_m$ are the average within-group distances for the lth and mth clusters, respectively, and $d_{l,m}$ is the inter-group distance between these clusters. These distances are defined as $\bar{d}_l = (1/N_l) \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \bar{\mathbf{x}}_l||$ and $d_{l,m} = ||\bar{\mathbf{x}}_l - \bar{\mathbf{x}}_m||$, where $\|\cdot\|$ is a norm (e.g. Euclidean).

Critério PBM

$$PBM = \left(\frac{1}{k} \frac{E_1}{E_K} D_K \right)^2$$

where E_1 denotes the sum of distances between the objects and the grand mean of the data, i.e. $E_1 = \sum_{i=1}^{N} ||\mathbf{x}_i - \overline{\mathbf{x}}||$, $E_K = \sum_{l=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \overline{\mathbf{x}}_l||$ represents the sum of within-group distances, and $D_K = \max_{l,m=1,\ldots,k} ||\overline{\mathbf{x}}_l - \overline{\mathbf{x}}_m||$ is the maximum distance between group centroids.

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 $O(N^2)$

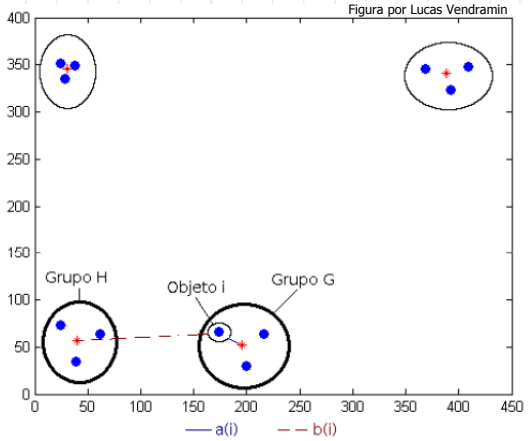
Propriedade Favorável: SWC ∈ [–1 ,+1]

Silhueta Simplificada (SSWC)

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

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

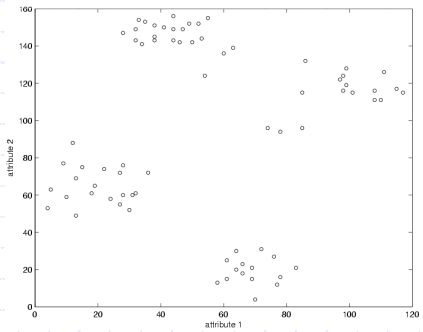
Figura por Lucas Vendramin

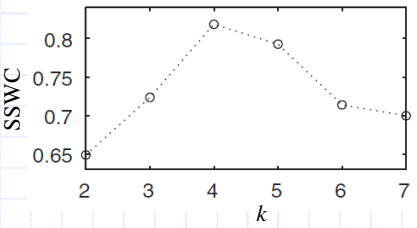


Silhueta Simplificada: $a(i)$ e $b(i)$ são calculados como a distância do i-ésimo objeto ao centróide do cluster em questão. Complexidade O(N).

Exemplo (SSWC)

- Relembrando a Subjetividade do Problema:
  - Quantos grupos abaixo...? Quatro? Cinco? Seis?
  - Sob a perspectiva **deste critério** (SSWC), são quatro!





Muitos Outros Critérios...

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

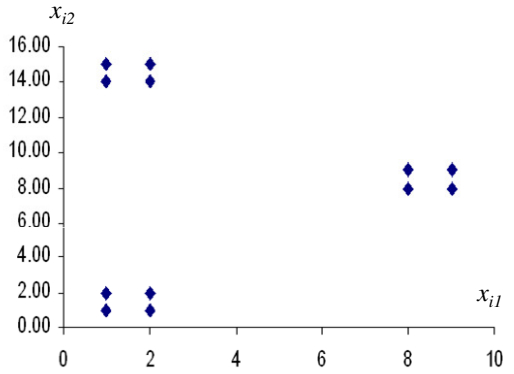
-	Criterion	Complexity
	Calinski-Harabasz (VRC)	$O(nN)$
	Davies-Bouldin (DB)	$O(n(k^2+N))$
	Dunn	$O(nN^2)$
	Silhouette Width Criterion (SWC)	$O(nN^2)$
	Alternative Silhouette (ASWC)	$O(nN^2)$
	Simplified Silhouette (SSWC)	$O(nNk)$
	Alternative Simplified Silhouette (ASSWC)	$O(nNk)$
	PBM	$O(n(k^2+N))$
	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

Exercícios

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

Prof. Eduardo R. Hruschka



- Calcule os índices de Dunn, DB, PBM e Silhuetas para a partição natural dos dados acima e também para outras partições. Compare os resultados.

Validação de Grupos

- Existem vários critérios estatísticos
- Nos restringiremos aqui a citar que alguns critérios definidos para partições podem ser decompostos em parcelas que se referem a grupos. Por exemplo:
  - Critérios Relativos
    - Davies-Bouldin: componentes D_l (pior caso da razão variâncias / separações)
    - Silhueta: média das silhuetas dos objetos do grupo (não da partição)
    - **Nota**: grupos grandes, esparsos, com sobreposição podem ser mal avaliados, mesmo que sejam grupos naturais. É preciso cautela...
  - Critérios Externos:
    - Rand, Jaccard, etc podem ser rescritos para grupos...

Problem

There are many relative clustering validity criteria

They behave differently in different application scenarios

It's a hard task for the user to choose a specific criterion when he or she faces such a variety of possibilities

A Milestone in this Area

Milligan & Cooper (1985):

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

- Compared the no. of clusters of the partition elected as the best one (among a set of candidates) by each criterion against the no. of clusters known to exist in a benchmark data set
  - for a number of synthetically generated data sets of a given class of interest
- Counted the number of times a criterion indicates as the best partition one with the "right" number of clusters

A Milestone in this Area

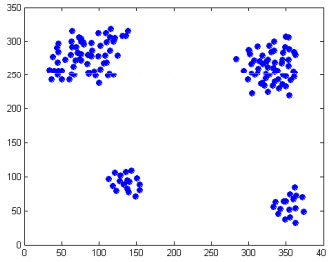
Milligan & Cooper (1985):

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

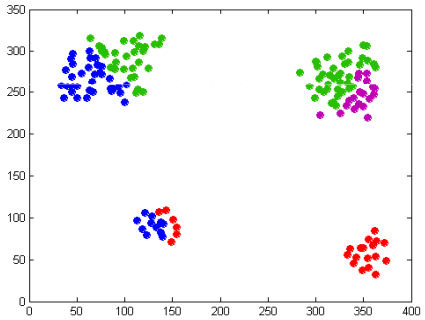
- Involved 30 criteria existing at that time
  - only a subset able to quantitatively assess the quality of partitions
  - the others are just stopping rules (for halting hierarchical algorithms)
- Results are still used by many authors to support their choices

Fine, but...

- There may exist various partitions of a data set into the *right* number of clusters that are very unnatural



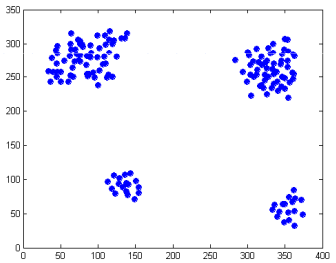
Data base – 4 clusters



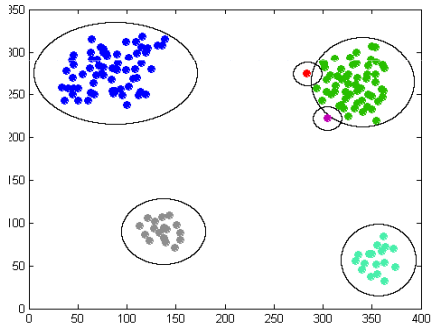
Partition – 4 clusters

Fine, but...

- There may exist numerous partitions into the *wrong* number of clusters, but clusters that exhibit a high degree of compatibility with the spatial distribution of the data



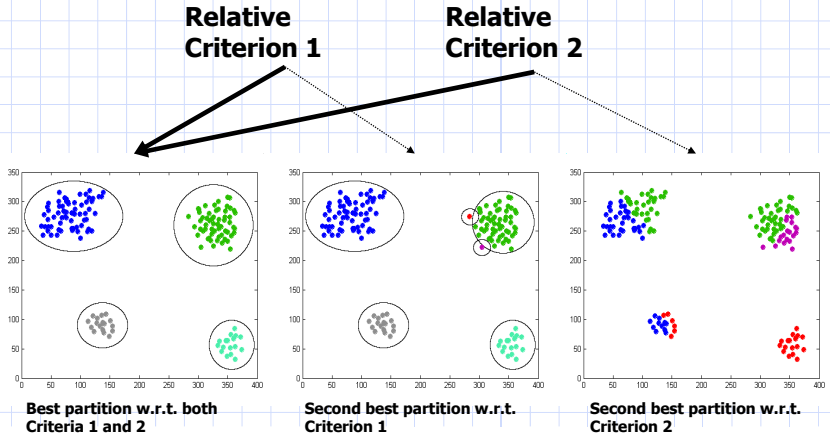
Data base – 4 clusters



Partition – 6 clusters

Fine, but...

- All partitions other than the "best one" are ignored...



Up-to-Date, Complementary Study:

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

Conceptual & Experimental Comparison of **40** Relative Validity Criteria:

- 962,928 partitions of a collection of 1080 data sets
- 2 statistical tests
- A number of evaluation scenarios
  - different no. clusters × no. attributes × clusters balances × ...
- Multiple evaluation aspects (**not only the number of clusters**)

Data sets representative of a particular class:

- volumetric clusters following multi-normal distributions

(Very) Summary of the Results

- Some results persistently hold with respect to all aspects and evaluation scenarios considered
- They strongly suggest that the best criteria for this class of data (Mixture of Gaussians) are:
  - The **Silhuettes**, **PBM**, **VRC**, and **Point-Biserial**
  - The Silhouettes showed the most robust results among them
    - virtually insensitive to any changes in the evaluation scenarios
    - simplified (prototype-based) version: comparable performance in $O(nkN)$
- Conclusions cannot be generalized to other classes of data !

Validação de Agrupamento com Sobreposição

- Existem diversos critérios externos e relativos para avaliação de partições soft, fuzzy e probabilísticas
- Uma revisão sobre critérios externos de validade de agrupamento com sobreposição é apresentada em:
  - Campello, R. J. G. B. "Generalized External Indexes for Comparing Data Partitions with Overlapping Categories" **Pattern Recognition Letters**, Vol. 31, p. 966-975, 2010
- Nos restringiremos aqui em discutir alguns critérios relativos comumente usados no contexto fuzzy

Relative Fuzzy Validity Criteria

❑ Idea:

- Taking into account the degree of membership of objects to clusters to evaluate the overall quality of a fuzzy partition in a quantitative way

❑ Requirements:

- Data set with *N* objects
  - Most times numerical objects are required $\mathbf{x}_j \in \Re^n \ \ (j = 1, ..., N)$
- Fuzzy partition matrix (c fuzzy clusters): $\mu_{ij} \in [0,1] \ \ (i = 1, ..., c)$
  - Most times prototypes are also required: $\mathbf{v}_i \in \Re^n \ \ (i = 1, ..., c)$

Fuzzy Hyper-Volume

$$FHV = \sum_{i=1}^{c} [\det(F_i)]^{1/2} \quad \Rightarrow \quad F_i = \frac{\sum_{j=1}^{N} \mu_{ij} (\mathbf{x}_j - \mathbf{v}_i)(\mathbf{x}_j - \mathbf{v}_i)^T}{\sum_{j=1}^{N} \mu_{ij}}$$

- The eigenvalues of F_i are directly related to the variances of the i-th fuzzy cluster
- $\det(F_i)$ provides a measure of the n-dimensional dispersion of the i-th cluster
- **The smaller the *FHV* value, the *better* the fuzzy partition of the data set**

Main Difficulties:

- Computation of $\det(F_i)$ is $O(n^3)$, where n is the number of attributes
- It subsumes data described by numerical attributes only

Average Partition Density

$$APD = \frac{1}{c} \sum_{i=1}^{c} \frac{R_i}{[\det(F_i)]^{1/2}} \quad \Rightarrow \quad \begin{cases} R_i = \sum_j \mu_{ij} \\ \forall j: (\mathbf{x}_j - \mathbf{v}_i)^T F_i^{-1} (\mathbf{x}_j - \mathbf{v}_i) < 1 \end{cases}$$

- R_i is the sum of memberships of the "central members" to the i-th cluster
- Compact fuzzy clusters provide large values of R_i and small values of $\det(F_i)$
- **The larger the *APD* value, the *better* the fuzzy partition**

Main Difficulties:

- Computation of $\det(F_i)$ and F_i^{-1} are both $O(n^3)$
- It subsumes data described by numerical attributes only

Xie-Beni Index

$$XB = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^m \|\mathbf{x}_j - \mathbf{v}_i\|_{\mathbf{A}}^2}{N \min_{p \neq q} \{\|\mathbf{v}_p - \mathbf{v}_q\|^2\}} \quad \Rightarrow \quad XB = \frac{J/N}{d}$$

total within-cluster fuzzy distance (J)

minimum distance between prototypes (d)

- The smaller J/N and greater d, the better the fuzzy partition with fixed c
- Both *compactness* (J/N) and *separation* (d) terms monotonically decrease with c
- **The smaller the *XB* value, the *better* the fuzzy partition**

Main Difficulties:

- *XB* may tend to decrease monotonically beyond *large* values of c
- It subsumes data described by numerical attributes only

Fuzzy Silhouette Width Criterion

Weighted Average of the Silhouettes computed over all the data objects:

$$FS = \frac{\sum_{j=1}^{N} (\mu_{pj} - \mu_{qj})^{\alpha} s_j}{\sum_{j=1}^{N} (\mu_{pj} - \mu_{qj})^{\alpha}}$$

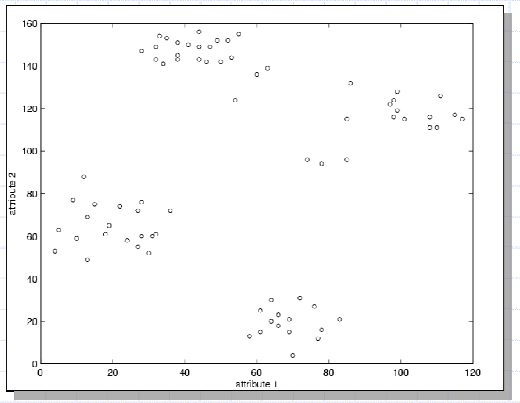
μ_{pj} : membership of the j-th object to its most compatible fuzzy cluster.

μ_{qj} : membership of the j-th object to its 2nd best matching fuzzy cluster.

- Weighted average instead of an arithmetic mean of the individual silhouettes
- Weight determined by the difference between the highest membership degrees
- Relative importance (weights) of objects located in overlapping areas is reduced
- **The larger the *FS* value, the *better* the fuzzy partition of the data set**
- $\alpha \rightarrow 0 \Rightarrow$ *FS* → *Non-Fuzzy Silhouette* (equivalent to each other for hard partitions)
- **Main Difficulties**:
  - may tend to favor partitions with *large* values of *c* (depending on α)
  - what is an appropriate value for α ?

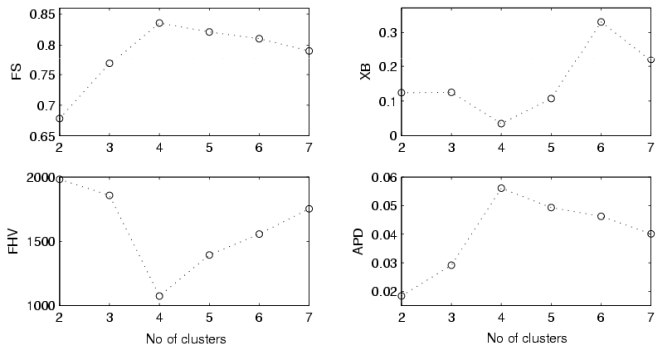
Example 1

- **Ruspini Data:**
  - 75 data objects
  - Visual inspection usually suggests 4 clusters



Example 1

- **FCM:**
  - Sequential execution over $c = 2, \ldots, 7$ (no. of clusters)
  - 10 runs starting from random prototypes (for each *c*)
  - The best values for each validity measure are stored

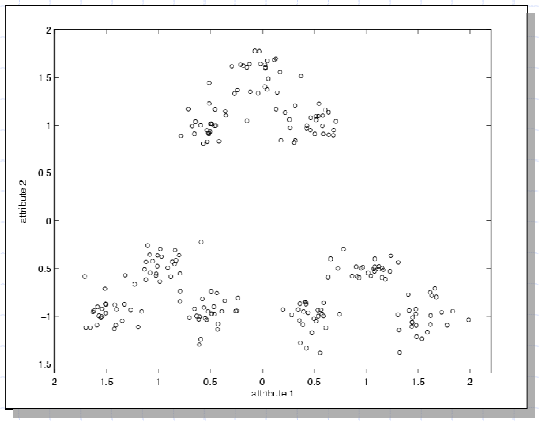


All criteria suggested 4 clusters (5 is the second choice)

Campello, R. J. G. B., Hruschka, E. R. "A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis", Fuzzy Sets and Systems, V. 157, p. 2858-2875, 2006

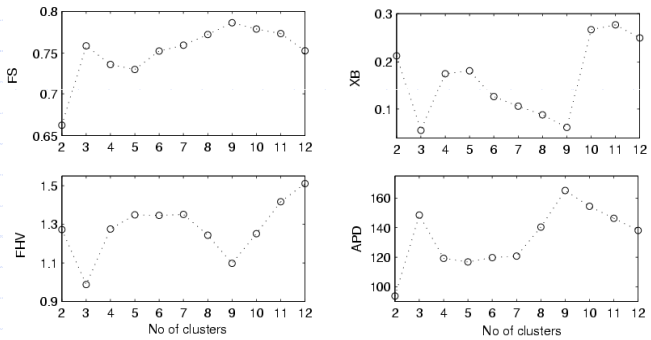
Example 2

- 225 data objects
- 9 normally distributed clusters with misleading spatial arrangement



Example 2

Multiple random restart sequential execution of FCM
(c = 2,…,12)

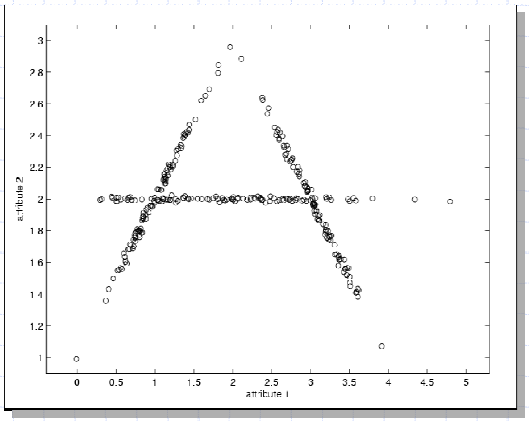


Misleading arrangement of clusters is captured by the indexes

Campello, R. J. G. B., Hruschka, E. R. "A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis", Fuzzy Sets and Systems, V. 157, p. 2858-2875, 2006

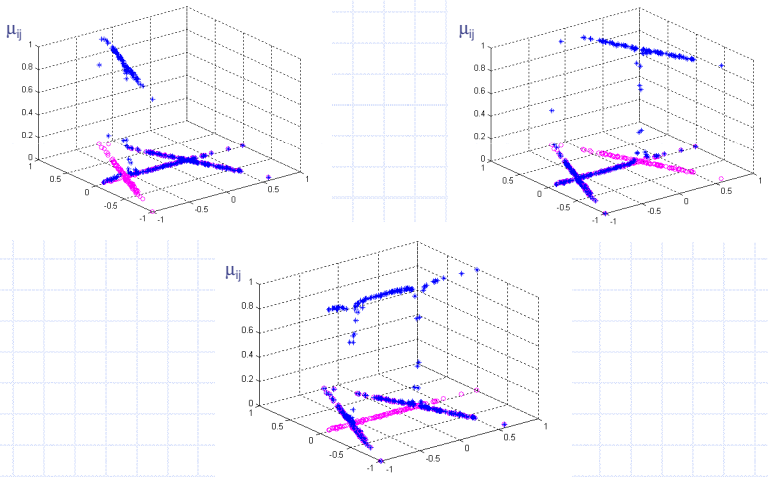
Example 3

- 300 data objects
- 3 ellipsoidal clusters that altogether resemble letter "A"



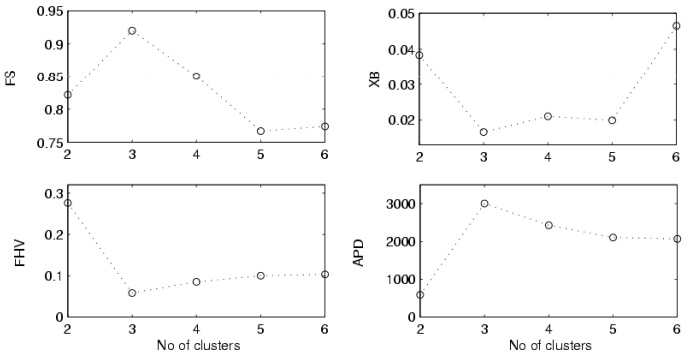
Example 3

Fuzzy Clusters: Typical scenario after running the Gustafson-Kessel algorithm with c = 3



Example 3

- **GK:**
  - Sequential execution over *c* = 2, …, 6 (no. of clusters)
  - 10 runs starting from random prototypes (for each *c*)
  - The best values for each validity measure are stored

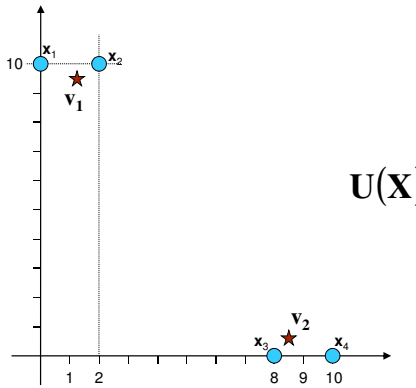


All criteria suggested 3 clusters

Campello, R. J. G. B., Hruschka, E. R. "A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis", Fuzzy Sets and Systems, V. 157, p. 2858-2875, 2006

Exercício

- Assuma hipoteticamente que os dados em azul na figura abaixo foram agrupados através de um algoritmo tipo FCM, com c = 2 clusters, fator m = 2 e uma inicialização tal que os protótipos (em vermelho) convergiram para as posições $\mathbf{v}_1 = [1{,}2 \quad 9{,}8]^T$ e $\mathbf{v}_2 = [8{,}8 \quad 0{,}2]^T$. Assuma ainda que a matriz de pertinência resultante, $\mathbf{U}(\mathbf{X}) = [\mu_{ij}]_{2\times4}$, é aquela indicada abaixo. Calcule os valores dos critérios FHV, APD, XB e FS para esta partição fuzzy dos dados.



$$\mathbf{U}(\mathbf{X}) = \begin{bmatrix} 0.85 & 0.9 & 0.1 & 0.15 \\ 0.15 & 0.1 & 0.9 & 0.85 \end{bmatrix}$$

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