AVALIAÇÃO DE RESULTADOS DE AGRUPAMENTO

MARGARIDA G. M. S. CARDOSO

AGENDA

INTRODUÇÃO

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO

Classes "a priori"

Índices de concordância

Limiares para índices de concordância

Visualizar a concordância entre partições

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO

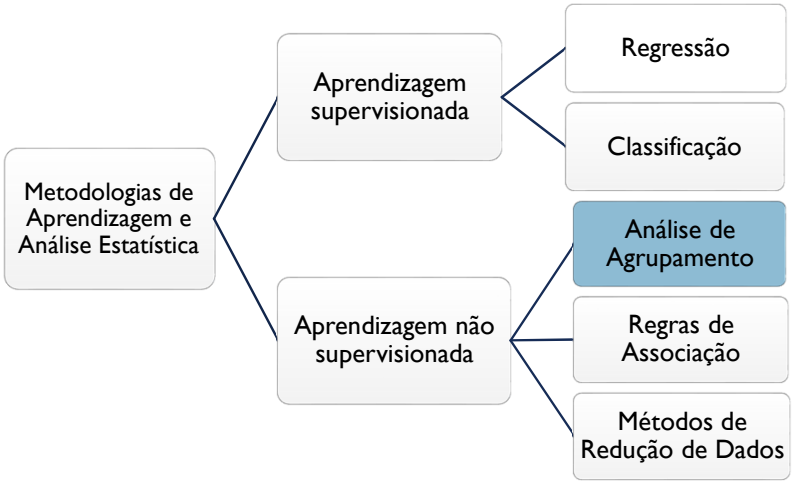
A estatística Gama

Índices de coesão-separação

Limiares para índices de coesão–separação

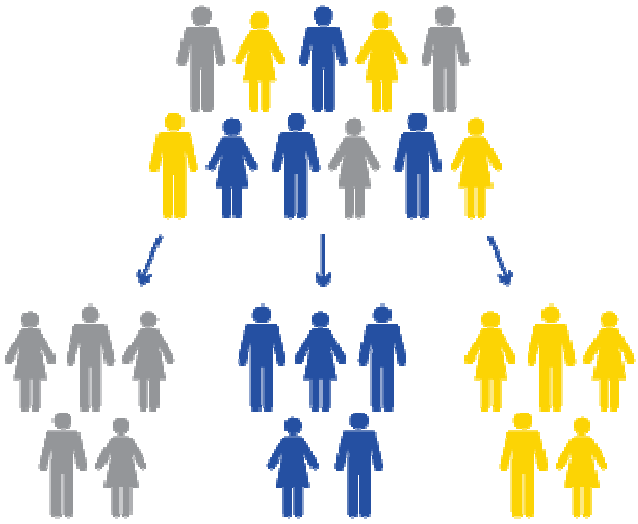
Estabilidade e validação cruzada

NOTAS FINAIS



INTRODUÇÃO

- A análise de agrupamento tem como objetivo genérico constituir grupos de elementos que sejam, de algum modo, homogéneos entre si e distintos dos que se integram noutros grupos
- As estruturas obtidas podem ser difusas, probabilísticas, partições..
- São inúmeros os processos de agrupamento: algoritmos hierárquicos, K-Médias diversos, variantes do algoritmo EM, baseados na teoria dos grafos, agrupamento conceptual…- e. g. (Jain, 2010):" Data clustering: 50 years beyond K-means"
- As aplicações são muito diversas – segmentação de indivíduos, produtos, experimentos…



1. INTRODUÇÃO

Decisões básicas:

- Seleção de entidades a agrupar
- Seleção de variáveis base de agrupamento
- Pré-avaliação de existência de um agrupamento (rejeição de hipótese de homogeneidade)
- Escolha de um modelo e/ou uma função objetivo (f.o.) assim como do tipo de estrutura de grupos pretendida
- Eleição de um algoritmo de agrupamento
- Determinação do número de grupos

1. INTRODUÇÃO

A obtenção de um ótimo para f.o. de agrupamento poderia finalizar o processo de avaliação dos resultados de agrupamento…

No entanto, é raro garantir-se a obtenção de um ótimo global ou poder reconhecer como tal uma solução obtida. O analista depara-se, então, com a necessidade de avaliar a qualidade de uma solução de agrupamento recorrendo a propriedades desejáveis do mesmo.

2. AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO

- 2.1 Classes "a priori"
- 2.2 Índices de Concordância
- 2.3 Limiares de Índices de Concordância
- 2.4 Visualizar a Concordância entre Partições

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

- em dados sintéticos – e.g. geração a partir de modelos de misturas finitas (Maitra e Melnykov 2010)

- em dados reais – e.g. no UCI Machine Learning Repository (Bache et al. 2013)

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

Iris (dados no UCI Machine Learning Repository)

Medidas (cm) de:
COMPRIMENTO DE SÉPALA
LARGURA DE SÉPALA
COMPRIMENTO DE PÉTALA
LARGURA DE PÉTALA

CLASSES:
IRIS SETOSA: 50
IRIS VERSICOLOUR: 50
IRIS VIRGINICA: 50

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

Wholesale customers (Dados no UCI Machine Learning Repository)

Gastos anuais (u.m.) em produtos:
FRESCOS
LÁCTEOS
MERCEARIA
CONGELADOS
DETERGENTES E PAPEL
CHARCUTARIA

"CLASSES":
HORECA: 298
RETALHO: 142

HoReCa

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

No agrupamento dos dados Iris e Wholesale usa-se uma variante do algoritmo EM, sendo usadas as estimativas obtidas via ML

Para determinar o número de grupos são usados critérios da Teoria de Informação que atendem à ML mas também à complexidade do modelo (BIC-Bayesian Information Criterion, por exemplo)

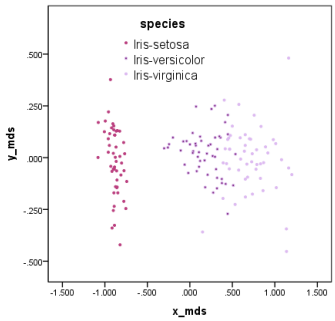
Notar que: As estimativas MAP ou MML poderiam também ser utilizadas.

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

Iris

Para obter agrupamento de dados estima-se um modelo de mistura finita mediante uma variante do algoritmo Expectation- Maximization

Obtém-se uma solução com 2 grupos para BIC e ICL-Integrated Completed Likelihood; 3 grupos para SICL-Supported Integrated Completed Likelihood (Baudry et al, 2015)



ICL	G1	G2
Setosa	50	0
Versicolor	0	50
Virginica	0	50

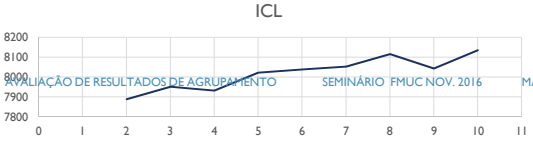
SICL	G1	G2	G3
Setosa	0	50	0
Versicolor	45	0	5
Virginica	0	0	50

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.1 CLASSES "A PRIORI"

Wholesale customers

Para obter agrupamento de dados estima-se um modelo de mistura finita mediante um algoritmo Expectation- Maximization (usa-se logaritmo de variáveis base)

Soluções com 2 grupos para ICL-Integrated Completed Likelihood e 4 para SICL-Supported Integrated Completed Likelihood



ICL	G1	G2
Horeca	246	52
Retalho	122	20

SICL	G1	G2	G3	G4
Horeca	14	45	18	221
Retalho	62	3	70	7

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

Índice de concordância simples

- Concordância percentual
- Informação Mútua Normalizada
- Variação de Informação
- …

Índices de concordância pareada

- Rand
- Jaccard
- Kulczynski
- Czekanwski
- …

(Cardoso, 2007)

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

	G1	G2	G3
Setosa	0	50	0
Versicolor	45	0	5
Virginica	0	0	50

Depois de permutação (um problema de afetação …)

	G1	G2	G3	total
Setosa	50	0	0	50
Versicolor	0	45	5	50
Virginica	0	0	50	50
Total	50	45	55	150

$$Perc\left(P_a^K, P_b^K\right) = \frac{50 + 45 + 50}{150} = 96{,}7\%$$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

	G1	G2	G3	total
Setosa	50	0	0	50
Versicolor	0	45	5	50
Virginica	0	0	50	50
Total	50	45	55	150

Atendendo aos valores esperados das contagens sob hipótese H0 de independência: $\frac{50\times50}{150}$=16,7…(Cohen 1960):

	G1	G2	G3
Setosa	16,7		
Versicolor		20,7	
Virginica			12,7

$$Kappa\left(P_a^K, P_b^K\right) = 0{,}95$$

$$= Perc_{\_}Corr\left(P_a^K, P_b^K\right) = \frac{\sum_{k=1}^{K}\frac{n_{kk}}{n} - \sum_{k=1}^{K}\frac{n_{k+}n_{+k}}{n^2}}{1-\sum_{k=1}^{K}\frac{n_{k+}n_{+k}}{n^2}}$$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

Índices de concordância simples (IC)

Concordância percentual

Informação Mútua Normalizada (Strehl e Ghosh 2002) e ajustada (Vinh et al, 2009)

….

$$Perc\left(P_a^K, P_b^K\right) = \sum_{k=1}^{K}\sum_{q=1}^{K}\frac{n_{kk}}{n}$$

$$I(P^K, P^Q)/\sqrt{H(P^K)H(P^Q)} = \sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n} log\left(\frac{n_{kq}}{n_{k+}n_{+q}}\right)/\sqrt{H(P^K)H(P^Q)}$$

$$H(P^K) = -\sum_{k=1}^{K}\frac{n_{k+}}{n} log\left(\frac{n_{k+}}{n}\right)$$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

Concordância pareada — P^K agrupa o par?

		Sim (1)	Não (0)
P^Q agrupa o par?	Sim (1)	a_{11}	a_{10}
	Não (0)	a_{01}	a_{00}

$$a_{11} = \frac{1}{2}\sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2 - \frac{n}{2}$$
$$a_{10} = \frac{1}{2}\left(\sum_{k=1}^{K} n_{k+}^2 - \sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2\right)$$
$$a_{01} = \frac{1}{2}\left(\sum_{q=1}^{Q} n_{+q}^2 - \sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2\right)$$
$$a_{00} = \frac{1}{2}\left[n^2 + \sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2 - \left(\sum_{q=1}^{Q} n_{+q}^2 + \sum_{k=1}^{K} n_{k+}^2\right)\right]$$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.2 ÍNDICES DE CONCORDÂNCIA

Índices de concordância pareada (ICp)

Rand (Rand 1971) ajustado (Hubert e Arabie, 1985)

$$Rand(P^K, P^Q) = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{00}}$$

Um índice destacado por Milligan and Cooper (1985)…

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.3 LIMIARES DE ÍNDICES DE CONCORDÂNCIA

Em geral, para obter um qualquer índice ajustado, $IC_a(P^K, P^Q)$:

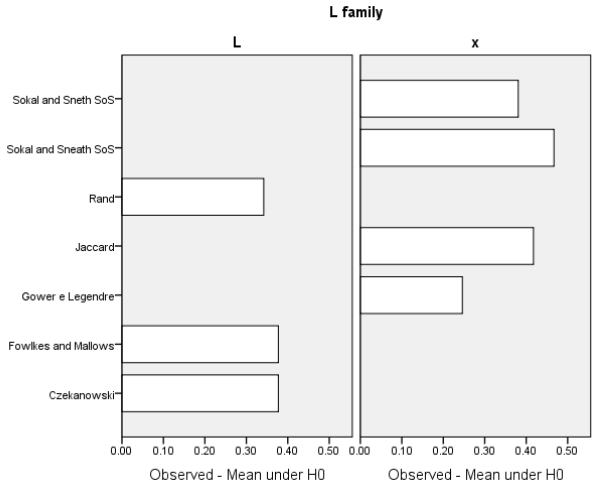
$$IC_a(P^K, P^Q) = \frac{IC(P^K, P^Q) - E_{H0}[IC(P^K, P^Q)]}{Max[IC(P^K, P^Q)] - E_{H0}[IC(P^K, P^Q)]}$$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.3 LIMIARES DE ÍNDICES DE CONCORDÂNCIA

- A família $\mathcal{L}$ de índices (incluindo Rand e Fowlkes–Mallows) proporciona um enquadramento geral para o ajustamento de vários ICp (pela média sob H0) - (Albatineh 2010)
- Para ajustar o índice de Jaccard pode usar-se uma estratégia aproximativa, partindo de uma relação com o índice C de (Czekanowski 1932) da família $\mathcal{L}$ - (Albatineh e Niewiadomska-Bugaj 2011)
- Para tornar viável o ajustamento, com precisão, de um qualquer índice de concordância pode utilizar-se uma estratégia que passa pela simulação de tabelas de classificação cruzada sob hipótese de concordância por acaso (H_0) (Amorim e Cardoso 2012, 2015)

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.3 LIMIARES DE ÍNDICES DE CONCORDÂNCIA

- Exemplo: Num cenário de grupos moderadamente separados e não equilibrados…



AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.3 LIMIARES DE ÍNDICES DE CONCORDÂNCIA

- Resultados para Iris (**espécies setosa, virginica e versicolor** como classes *a priori*)

		SIMPLES	AJUSTADOS
K=2	nMI	0.7612	0.5794
	Rand	0.7763	0.5681
K=3	nMI	0.8997	0.8983
	Rand	0.9575	0.9039

Nota: apenas os índices ajustados estão em escala comparável

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.3 LIMIARES DE ÍNDICES DE CONCORDÂNCIA

- Resultados para Wholesale
  (canais de distribuição **Horeca e Retalho** como proxy de classes *a priori*)

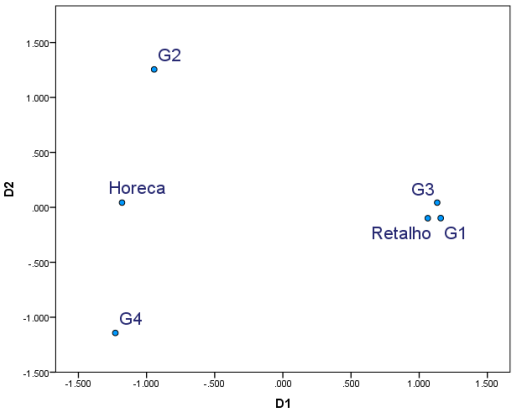
K=2		SIMPLES	AJUSTADOS
	nMI	0.0017	0.0015
	Rand	0.5207	-0.0151
K=4	nMI	0.3998	0.2886
	Rand	0.7080	0.4371

Nota: apenas os índices ajustados estão em escala comparável

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
2.4 VISUALIZAR A CONCORDÂNCIA ENTRE PARTIÇÕES

- Para visualizar a concordância entre grupos e classes, propõe-se o uso do algoritmo PREFSCAL desenvolvido, para MDU–Multidimensional Unfolding (Busing, Groenen e Heiser 2005)
- Como alvo de representação consideram-se medidas que deduzem a concordância "por acaso" (Martins and Cardoso, submetido)

	G1	G2	G3	G4
Horeca	14	45	18	221
Retalho	62	3	70	7



$R^2([\hat{d}_{kq}], [d_{kq}]) \cong 1)$

AVALIAÇÃO EXTERNA DE UM AGRUPAMENTO
SUMÁRIO

- A avaliação externa baseia-se em classes conhecidas "a priori"
- A partir da tabela de contingência entre as partições que se constituem e as classes "a priori" determinam-se índices de concordância (simples e pareada)
- Para viabilizar a correta interpretação dos valores dos índices, estes são ajustados a partir de limiares que se estabelecem sob hipótese de concordância por acaso (via analítica, aproximativa ou com recurso à simulação)
- A representação da concordância entre partições via MDU pode substituir a resolução de um problema de afetação (necessário ao cálculo da concordância percentual) e ilustrar a proximidade entre os grupos das duas partições que se comparam

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO

- 3.1 A Estatística Gama
- 3.2 Índices de Coesão-Separação
- 3.3 Limiares para Índices de Coesão-Separação
- 3.4 Estabilidade e Validação Cruzada
- 3.5 Outros modos de avaliação

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.1 ESTATÍSTICA GAMA

Considere-se a estatística:

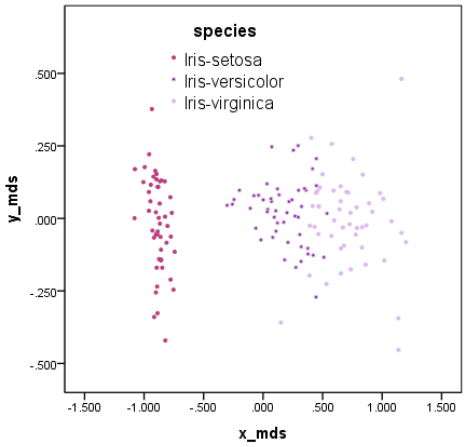
$$\Gamma\left(D^{P^Q}, D^{P^K}\right) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d^{P^Q}\left(\underline{y_i}, \underline{y_j}\right) d^{P^K}\left(\underline{y_i}, \underline{y_j}\right)$$

Se as distâncias forem unitárias no caso em que os objetos $\underline{y_i}$ e $\underline{y_j}$ pertencem ao mesmo grupo e nulas caso contrário, a estatística Γ coincide com o índice de (Russel e Rao 1940), uma medida de concordância pareada que se pode adicionar às já mencionadas

Para a avaliação interna de uma partição pode considerar-se $\Gamma\left(D; D^{P^K}\right)$ em que D se refere às distâncias calculadas sobre os objetos base de agrupamento – Estatística Gama (Hubert 1977)

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

- As propriedades de coesão e separação são inerentes à própria ideia de agrupamento
- Em geral, os índices de coesão-separação (ICS) baseiam-se na relação entre a variação intra-grupos ("Within") e entre-grupos ("Between")
- Alguns ICS foram generalizados de modo a avaliarem estruturas difusas – por exemplo, o índice PBM (Pakhira, Bandyopadhyay and Maulik 2004)



AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

Alguns exemplos são:

- Calinski e Harabasz
- Davies and Bouldin
- Silhueta

….

(Milligan and Cooper 1986) destacam o índice Calinski e Harabasz entre 30…
(Vendramin, Campello and Hruschka 2010) destacam o índice Silhueta entre os 30 anteriores+10…

$$Silh(P^K) = \frac{1}{K}\sum_{k=1}^{K} Silh(G_k)$$

$$Silh(G_k) = \frac{1}{\#G_k}\sum_{\underline{y}_i \in G_k} Silh\left(\underline{y}_i, G_k\right)$$

$$Silh\left(\underline{y}_i, G_k\right) = \frac{b\left(\underline{y}_i, G_k\right) - a(\underline{y}_i, G_k)}{max\left\{b\left(\underline{y}_i, G_k\right), a(\underline{y}_i, G_k)\right\}}$$

$$b\left(\underline{y}_i, G_k\right) = min_{k' \neq k} \sum_{\underline{y}_j \in G_{k'}} \frac{d\left(\underline{y}_i, \underline{y}_j\right)}{\#G_{k'}}$$

$$a\left(\underline{y}_i, G_k\right) = \sum_{\underline{y}_j \in G_k - \{\underline{y}_i\}} \frac{d\left(\underline{y}_i, \underline{y}_j\right)}{\#G_k - 1}$$

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

ICS ➡ f.o.

- Na prática de agrupamento os ICS são, por vezes, considerados como f.o. de um processo de agrupamento – por exemplo, o Índice Silhueta (Hruschka et al. 2006)

f.o. ➡ ICS

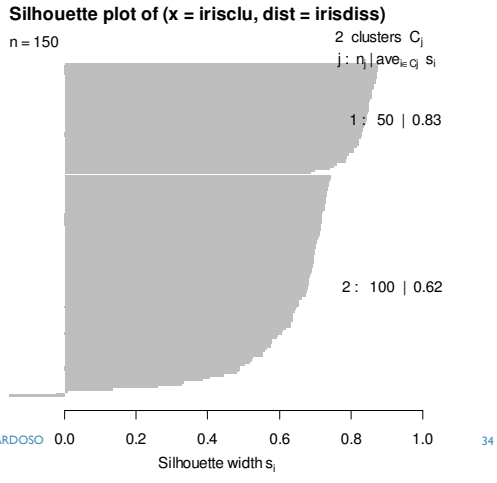
- Por outro lado, funções tradicionalmente consideradas como objetivo podem sugerir um novo ICS – por exemplo, o índice baseado no critério "Minimum Message Length" (Fred e Jain 2008)

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

Silhueta	Interpretação proposta
(-1) a 0,25	Nenhuma estrutura substancial foi descoberta
0,26 a 0,50	A estrutura é fraca e pode ser artificial; tente métodos adicionais sobre este conjunto de dados
0,51 a 0,70	A estrutura descoberta é razoável
0,71 a 1	Foi descoberta uma estrutura robusta

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

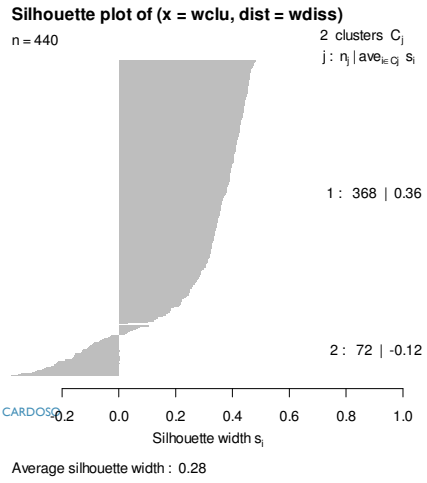
	Silhueta
K=2	**0.6864**
K=3	0.3951
Species	0.5032
Iris	



AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.2 ÍNDICES DE COESÃO-SEPARAÇÃO

Índices sobre dados base logaritmizados

	Silhueta
K=2	**0.2791**
K=4	0.1810
Canal de distribuição	0.2423
Wholesale	



AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.3 LIMIARES ÍNDICES DE COESÃO-SEPARAÇÃO

Kaufman et al. 1990) :" …experience with the Silhouette index which has led us to a rather subjective interpretation…" (p. 88).

Para determinar os limiares de um ICS (Cardoso e Carvalho 2009):

1. Gerar múltiplas amostras sob hipótese de homogeneidade (H_{0h})
2. A partir de distribuição empírica do ICS sob H_{0h} determinar um limiar (média, por exemplo)
3. Comparar valor de limiar com o valor de ICS para o agrupamento sobre a amostra original

Finalmente, selecionar o agrupamento correspondente a comparação mais favorável com o limiar

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.4 ESTABILIDADE E VALIDAÇÃO CRUZADA

- A estabilidade tem sido reconhecida como uma propriedade desejável de um agrupamento - e.g. (Jain et al. 1988). Uma solução de agrupamento diz-se estável se se mantiver razoavelmente inalterada quando o processo de agrupamento for sujeito a pequenas modificações…
- Na avaliação da estabilidade que decorre da comparação de agrupamentos resultantes de diferentes amostras, pode usar-se um procedimento de validação cruzada (McIntyre e Blashfield 1980), (Breckenridge 1989),uma metodologia muito comum na análise supervisionada

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.4 ESTABILIDADE E VALIDAÇÃO CRUZADA

Etapa	Ação	Resultado
1	Particionar amostra original	Amostras de treino e de teste
2	Agrupar amostra de treino	Grupos na amostra de treino
3	Construir um Classificador supervisionado por grupos na amostra de treino. Usar o Classificador na amostra de teste	Classes na amostra de teste
4	Agrupar amostra de teste	Grupos na amostra de teste
5	Calcular Índices de Concordância entre grupos e classes obtidos sobre amostra de teste.	Valor de referência para avaliação de estabilidade

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.4 ESTABILIDADE E VALIDAÇÃO CRUZADA

- Questão 1 –Selecionar um classificador adequado.

  Proposta: A utilização de um procedimento de estimação de um modelo de mistura finita para agrupamento permite, não só, constituir os grupos mas, também, obter um classificador que resulta da própria estimação do modelo e que pode ser utilizado sobre uma amostra de teste (Cardoso 2007)

- Questão 2 - Dispor de uma amostra original com dimensão suficiente.

  Proposta: A validação cruzada que radica no uso de uma amostra ponderada (Cardoso et al. 2010)

  - "imita" a constituição de subamostras aleatórias de treino e teste
  - elimina a necessidade de construir um classificador (Questão 1 não se coloca)

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.4 ESTABILIDADE E VALIDAÇÃO CRUZADA

- Nos exemplos Iris e Wholesale, o estudo da associação entre as propriedades de coesão-separação e a de estabilidade – (Cardoso et al. 2010) – leva a crer que maior separação-coesão estará associada a maior estabilidade. Por outro lado, (Amorim e Cardoso, 2015), a estabilidade não se associa necessariamente à validade externa.

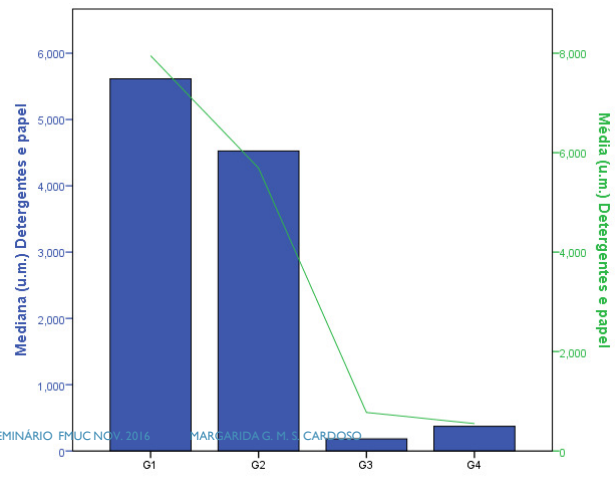
AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.5 OUTROS MODOS DE AVALIAÇÃO

- A avaliação de uma solução de agrupamento pode fazer-se na perspetiva de cada grupo - e.g. (Mirkin 1996). Neste caso avalia-se:
  - Separação ou diferença face à totalidade da amostra
  - A consistência com que um grupo é "descoberto" como resultado de vários agrupamentos efetuados sobre subamostras diferentes - estabilidade "cluster-wise" (Hennig 2007)

- Finalmente é imprescindível interpretar os grupos constituídos, já que a facilidade de interpretação de uma solução de agrupamento é inseparável da sua avaliação e, em última análise, condição da sua utilidade. Para este efeito podem ser muito úteis os métodos de classificação supervisionada.

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO
3.5 OUTROS MODOS DE AVALIAÇÃO

Gastos	**CART:** Medida de Importância
Frescos	0,140
Lácteos	**0,142**
Mercearia	**0,199**
Congelados	0,128
Detergentes e Papel	**0,277**
Charcutaria	0,033



AVALIAÇÃO INTERNA DE UM AGRUPAMENTO 3.5 OUTROS MODOS DE AVALIAÇÃO

Árvore CART para discriminar entre grupos:

Detergentes & papel ≤1629,5 e Frescos ≤ 1073 e Charcutaria ≤ 163,5→ **G3**

Detergentes & papel ≤ 36,5 e Frescos > 1073 e Detergentes & papel ≤ 36,5 → **G3**

Detergentes & papel ≤ 36,5 e Frescos > 1073 e Detergentes & papel >36,5 → **G4**

Detergentes & papel > 1629,5 e Congelados ≤ 366,5 → **G1**

Detergentes & papel > 1629,5 e Congelados > 366,5 e Frescos ≤ 2145→ **G1**

Detergentes & papel > 1629,5 e Congelados > 366,5 e Frescos > 2145→ **G2**

Erro	14 %
Erro (10-fold)	20%

AVALIAÇÃO INTERNA DE UM AGRUPAMENTO SUMÁRIO

1. A Estatística Gama pode ser usada para relacionar distâncias originais entre dados e um agrupamento
2. Os índices de coesão-separação (ICS) medem propriedades inerentes ao próprio conceito de agrupamento e podem ser usados como f.o. de agrupamento
3. Considerando os valores de um ICS sob hipótese de homogeneidade pode estabelecer-se limiar para o ICS que se associa a um agrupamento
4. A validação cruzada de um agrupamento para medir a sua estabilidade coloca algumas questões que podem ser ultrapassadas mediante uma análise sobre amostra ponderada
5. Outros modos de avaliação podem ser considerados: os seus resultados devem contribuir para interpretar e tornar útil um resultado de agrupamento

NOTAS FINAIS

- A avaliação externa de agrupamentos deverá ser vista, mais propriamente, como avaliação de algoritmos de agrupamento
- Na avaliação interna de partições resultantes de agrupamento, evidenciam-se as propriedades de coesão-separação e de estabilidade. Estas propriedades poderão ser úteis na eliminação de partições candidatas, mas não necessariamente na seleção da partição "real".
- A escolha de índices de concordância (para usar no estudo da estabilidade temporal de um agrupamento, por exemplo) é objecto de estudo atualmente, procurando-se uma tipificação destes índices sob hipótese de "concordância por acaso".

Albatineh, A. N. (2010), "Means and Variances for a Family of Similarity Indices Used in Cluster Analysis," *Journal of Statistical Planning and Inference*, 140, 2828-2838.
Albatineh, A. N., and Niewiadomska-Bugaj, M. (2011), "Correcting Jaccard and Other Similarity Indices for Chance Agreement in Cluster Analysis," *Advances in Data Analysis and Classification*, 5, 179-200.
Bache, K., and Lichman, M. (2013), "Uci Machine Learning Repository."
Baudry, J.-P., Cardoso, M. G. M. S., Celeux, G., Amorim, M. J., and Ferreira, A. S. (2015), "Enhancing the Selection of a Model-Based Clustering with External Categorical Variables," *Advances in Data Analysis and Classification*, 9, 177-196.
Breckenridge, J. N. (1989), "Replicating Cluster Analysis: Method, Consistency, and Validity," *Multivariate Behavioral Research*, 24, 147-161.
Fred, A. L., and Jain, A. K. (2008), "Cluster Validation Using a Probabilistic Attributed Graph," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, IEEE, pp. 1-4.
Hennig, C. (2007), "Cluster-Wise Assessment of Cluster Stability," *Computational Statistics & Data Analysis*, 52, 258-271.
Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of classification*, 2, 193-218.
Hubert, L. J., and Baker, F. B. (1977), "The Comparison and Fitting of Given Classification Schemes," *Journal of Mathematical Psychology*, 16, 233-253.

Jain, A. K. (2010), "Data Clustering: 50 Years Beyond K-Means," *Pattern recognition letters*, 31, 651-666.
Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), "Data Clustering: A Review," *ACM computing surveys (CSUR)*, 31, 264-323.
Kaufman, L., and Rousseeuw, P. J. (1990), "Finding Groups in Data. An Introduction to Cluster Analysis," *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990*, 1.
McIntyre, R. M., and Blashfield, R. K. (1980), "A Nearest-Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure," *Multivariate Behavioral Research*, 15, 225-238.
Melnykov, V., and Maitra, R. (2010), "Finite Mixture Models and Model-Based Clustering," *Statistics Surveys*, 4, 80-116.
Milligan, G. W., and Cooper, M. C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50, 159-179.
Milligan, G. W., and Cooper, M. C. (1986), "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research*, 21, 441-458.
Mirkin, B. (1996), *Mathematical Classification and Clustering* (Vol. 11), Springer Science & Business Media.
Strehl, A., and Ghosh, J. (2002), "Cluster Ensembles-a Knowledge Reuse Framework for Combining Partitionings," in *AAAI/IAAI*, pp. 93-99.
Vendramin, L., Campello, R. J., and Hruschka, E. R. (2010), "Relative Clustering Validity Criteria: A Comparative Overview," *Statistical Analysis and Data Mining*, 3, 209-235.
Vinh, N. X., Epps, J., and Bailey, J. (2009), "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 1073-1080.