A procura da melhor partição em Classificação Hierárquica: A abordagem SEP/COP

Lúcia Sousa
Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Viseu
Fernanda Sousa
Faculdade de Engenharia e CITTA, Universidade do Porto

XVIII Jornadas de Classificação e Análise de Dados - JOCLAD 2011
Vila Real – 6 a 9 de Abril

Plano

- Classificação hierárquica
- Validação de partições
- A abordagem SEP/COP
- Procedimento Metodológico
- Experimentação – Dados simulados
- Discussão dos Resultados
- Comentários finais

Classificação Hierárquica

Objectivo da classificação: Agrupar os elementos do conjunto a classificar em classes coesas e bem separadas.

Quadro de dados: n indivíduos × p variáveis.

Escolhas inerentes: - Medida de proximidade
- Método de agregação de classes

Classificação Hierárquica → Hierarquia de partições (dendrograma)

Escolha da melhor partição

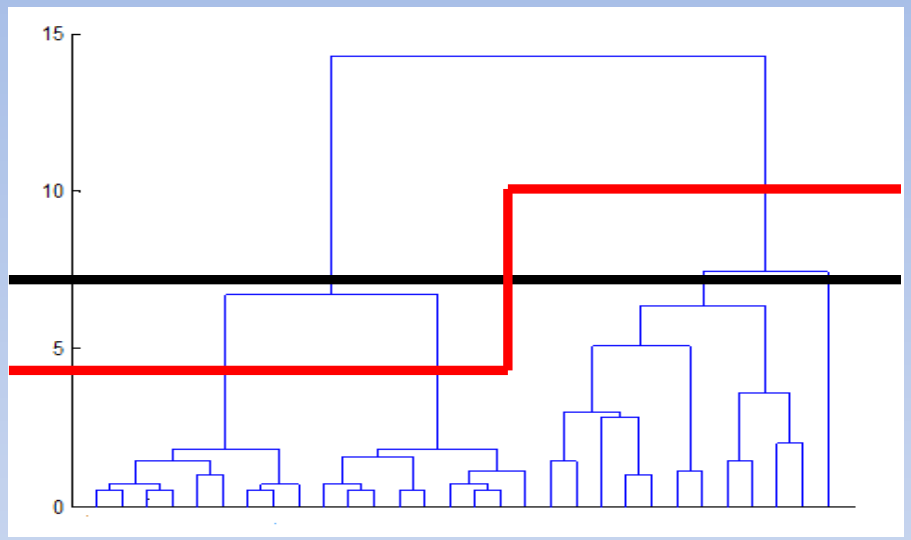
Validação de Partições

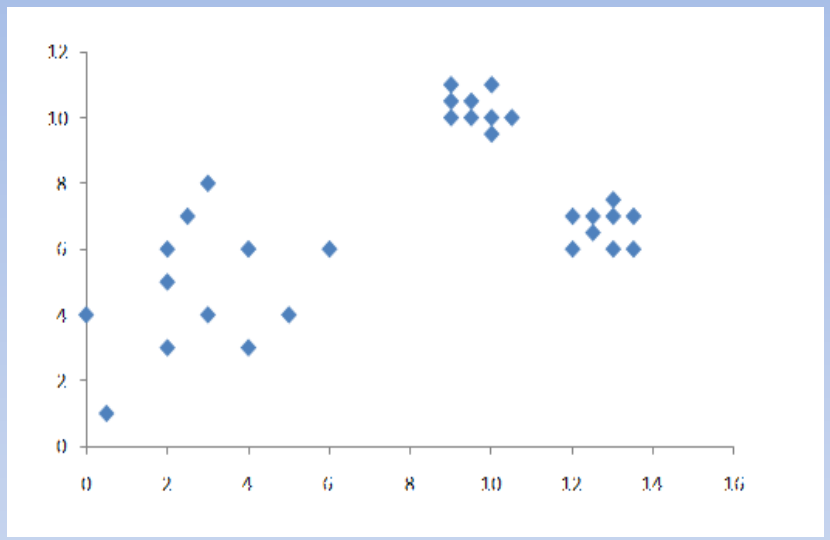
Índices de validação de partições: Comparam diferentes partições recorrendo à análise da coesão e à separação das classes que as constituem.

- Índices de validação externa: comparação da partição a validar com a partição de referência.
- Índices de validação relativa: comparação de duas partições obtidas, a partir do mesmo conjunto de dados, por processos diferentes.

Índice de Rand Corrigido, "*Adjusted Rand Index*" - ARI (Hubert & Arabie, 1985)

Abordagem SEP/COP





SEP/COP (Ibai Gurrutxaga et al., 2010) – Algoritmo SEP- *"Search over the Extended Partition set"*- procura a melhor partição no conjunto das partições estendidas (parciais) recorrendo ao índice de validação de partição COP- *"Context-independent Optmality and Partiality properties".*

Abordagem SEP/COP

Sejam:

- X o conjunto de elementos a classificar
- P^Y uma partição parcial de X

$$P^Y = \{C_1, \dots, C_k : \bigcup_{i=1}^{k} C_i = Y, C_i \cap C_j = \emptyset, \forall i \neq j, Y \subseteq X\}$$

- $H = \{P_1, \dots, P_R\}$ uma hierarquia de partições sobre X

$$\forall P_r, P_s \in H, r < s \Leftrightarrow \forall C_k \in P_r \; \exists C_l \in P_s : C_k \subseteq C_l$$

- E_H o conjunto de partições estendidas de uma hierarquia

$$E_H = \left\{ P : P \subseteq T, \bigcup_{C \in P} C = X, \forall C_k, C_l \in P : C_k \cap C_l = \emptyset \right\} \quad T = \bigcup_{C \in P, P \in H} C$$

- V(P) um índice de validação da partição P

Abordagem SEP/COP

Dados X e H, pretende-se encontrar a melhor partição, de acordo com V, no conjunto de partições estendidas de H.

Abordagem SEP/COP

O algoritmo SEP analisa cada sub-árvore independentemente, e decide de acordo com o índice de validação, qual a melhor partição parcial em cada nó da árvore.

O índice de validação COP:

- avalia as partições parciais,
- assegura as melhores partições parciais após as sucessivas agregações.

Abordagem SEP/COP

$$COP(P^Y, X) = \frac{1}{|Y|} \sum_{C \in P^Y} |C| \frac{intra(C)}{inter(C)}$$

$intra(C) = \frac{1}{|C|} \sum_{X \in C} d(X, \bar{C})$ - Variância intra-classes (coesão).

$inter(C) = \min_{x_i \notin C} \max_{x_j \in C} d(x_i, x_j)$ - Variância inter-classes (separação).

$$0 \leq COP \leq 1$$

$$COP(P^Y \cup P^Z, X) = \frac{1}{|Y| + |Z|} \left(\sum_{C \in P^Y} |C| \frac{intra(C)}{inter(C)} + \sum_{C \in P^Z} |C| \frac{intra(C)}{inter(C)} \right)$$

$$= \frac{1}{|Y| + |Z|} (|Y| COP(P^Y) + |Z| COP(P^Z))$$

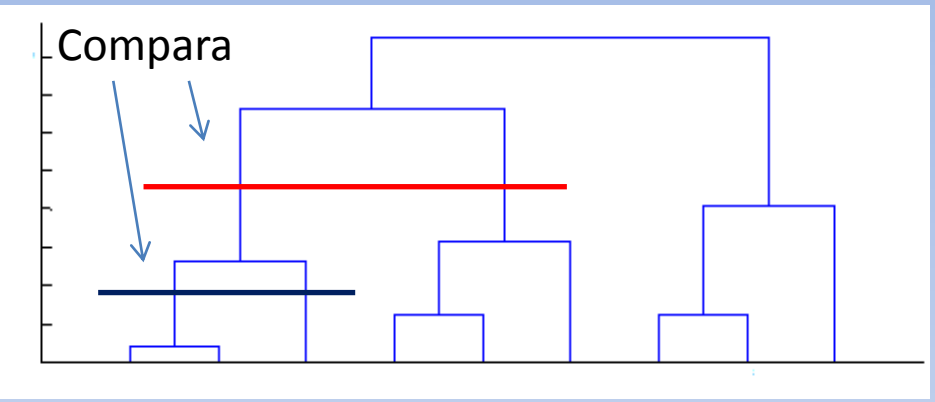
Abordagem SEP/COP

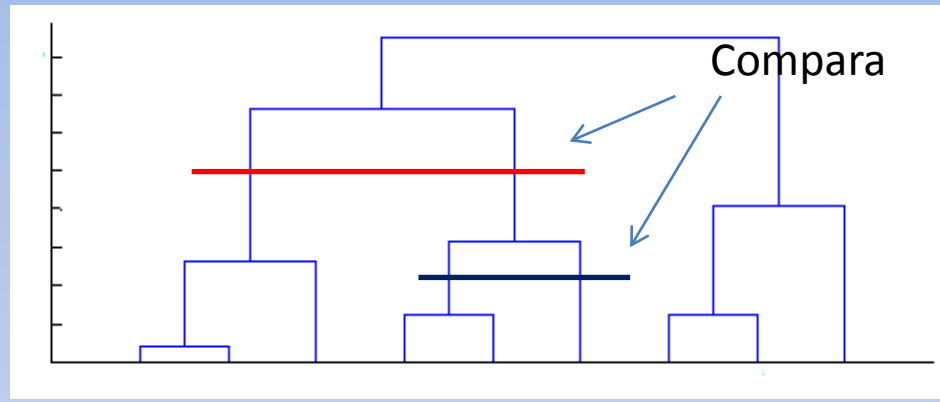
Descrição do algoritmo :

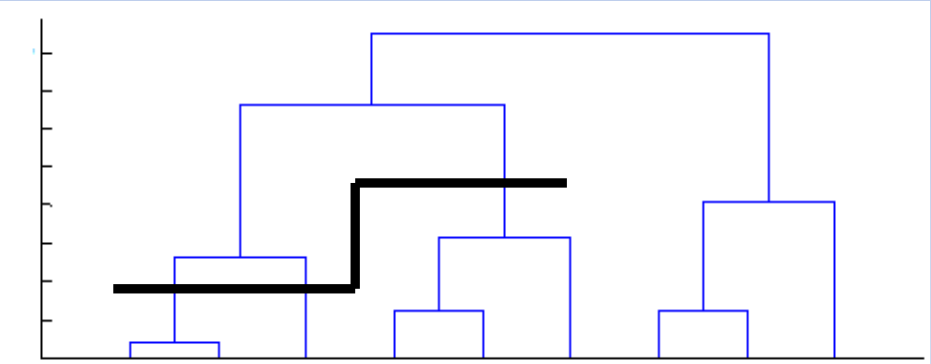
- A partir do nó raiz, no sentido de cima para baixo cria um vector com os nós interiores dos “descendentes esquerdo” e “descendentes direito” (árvores binárias).

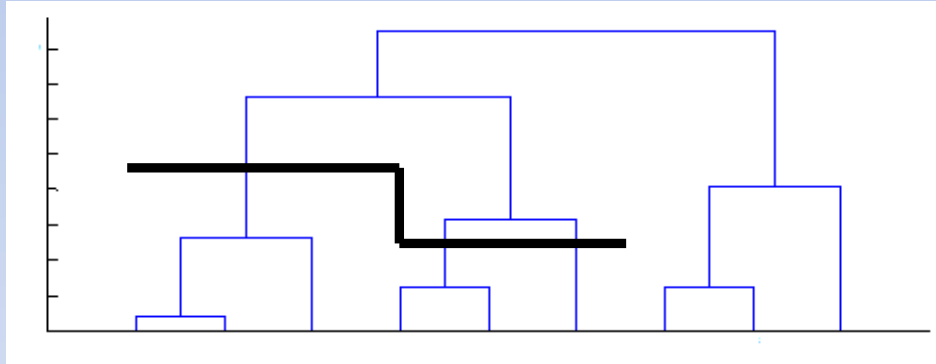
- De baixo para cima, para cada elemento do vector com os nós interiores dos “descendentes”:
  1) calcula o valor do índice COP para
     - a partição com a classe correspondente ao nó corrente
     - a união da melhor partição em cada nó “descendente” do nó corrente
  2) compara esses valores de COP
  3) decide pela melhor partição.

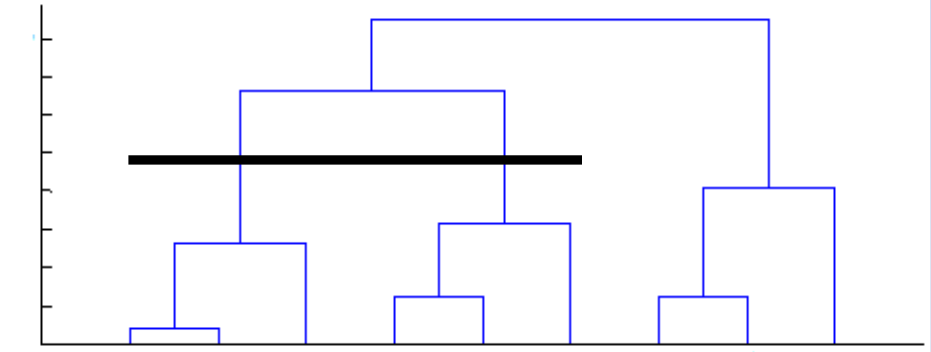
Abordagem SEP/COP

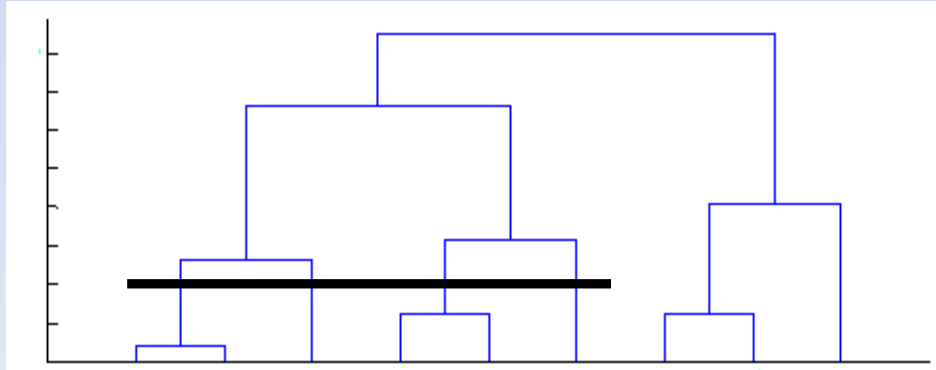




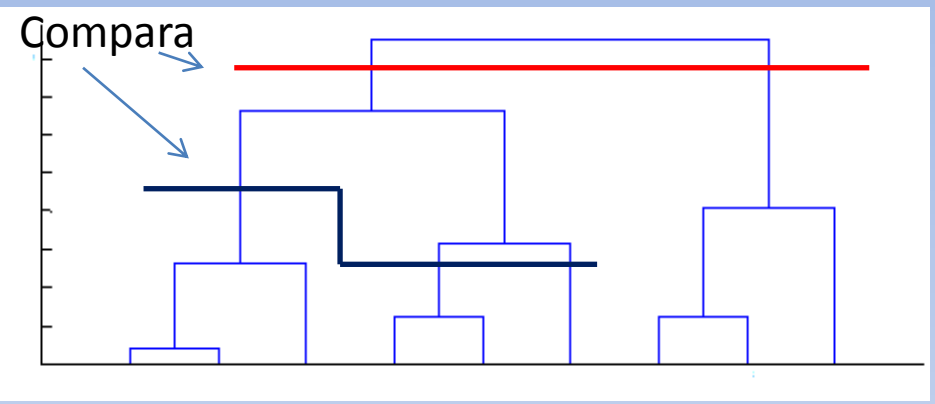


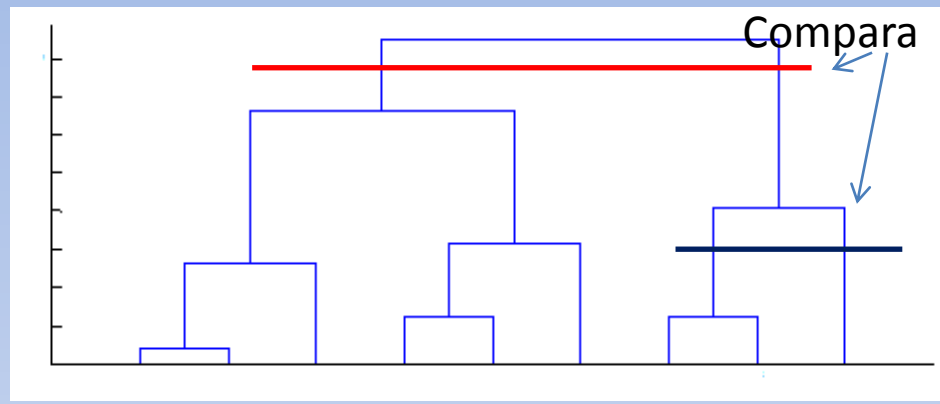


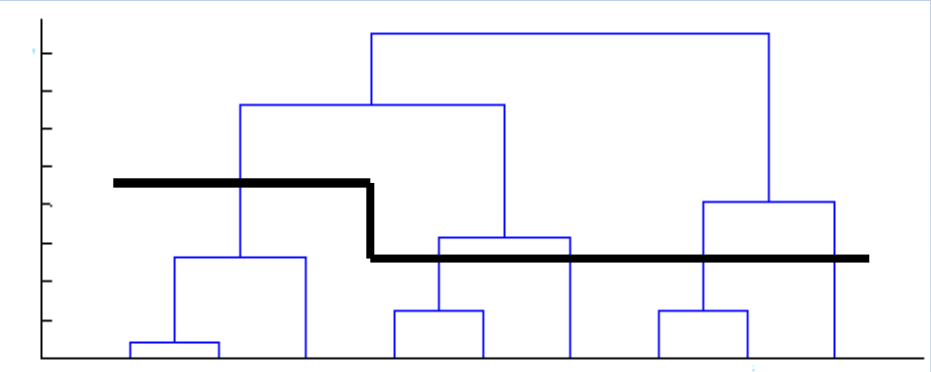


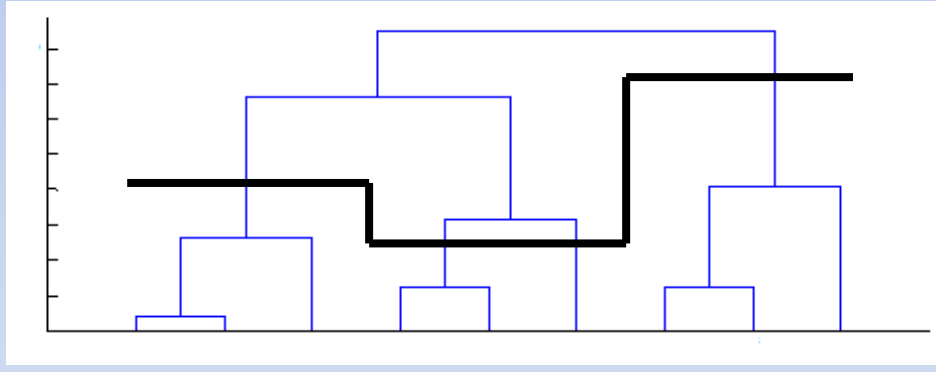


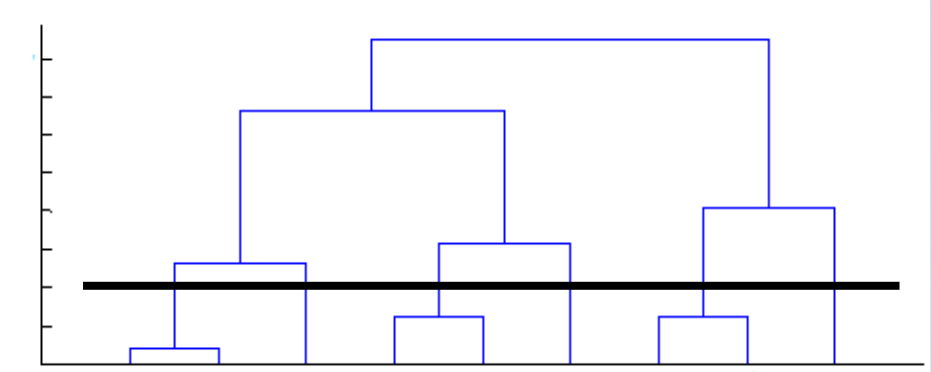
Abordagem SEP/COP

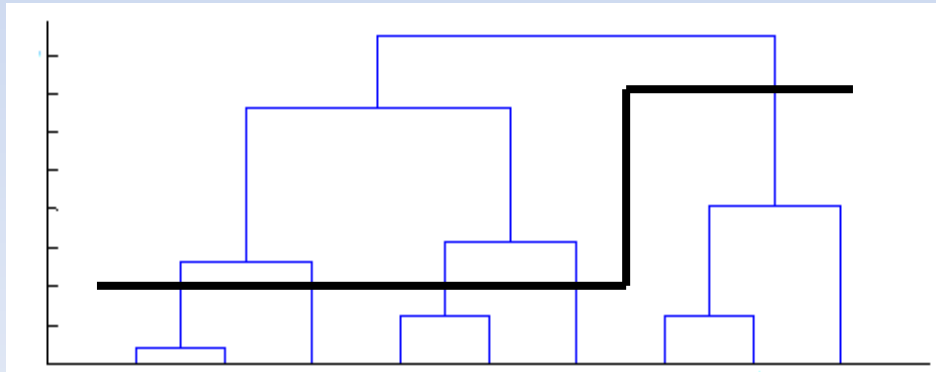




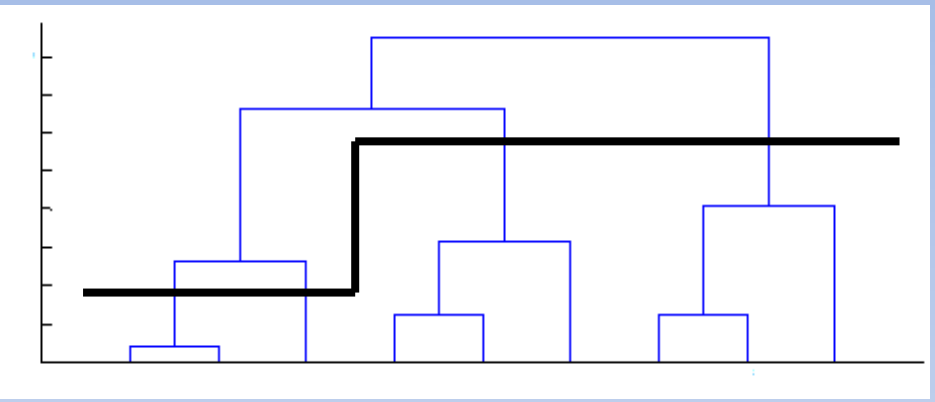


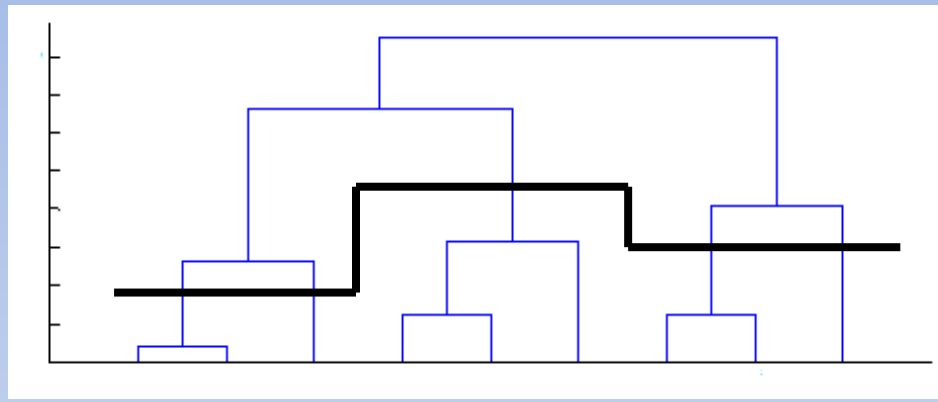


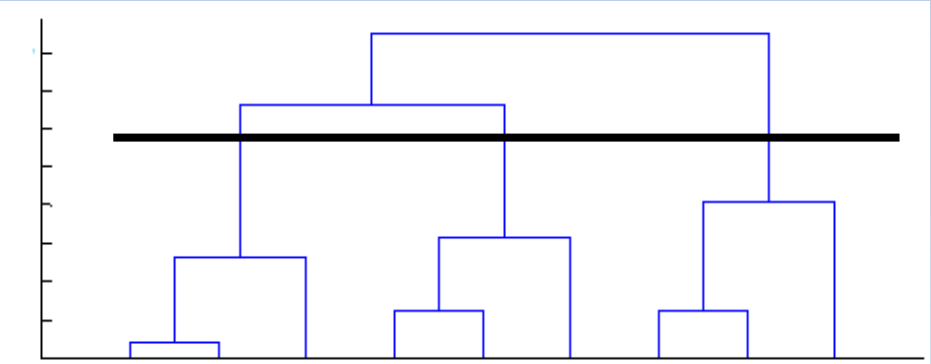


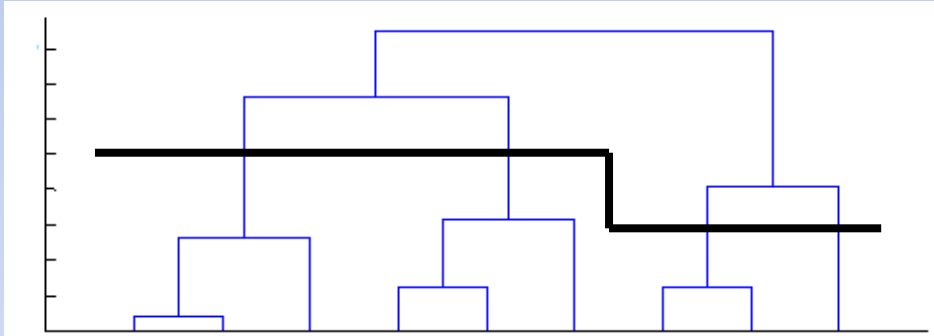


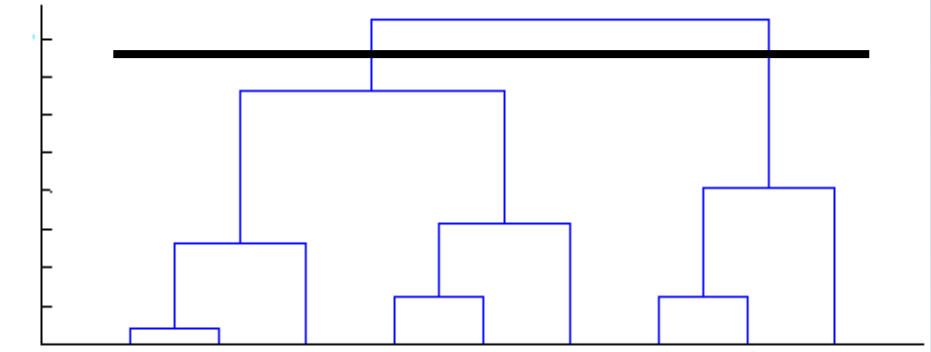
Abordagem SEP/COP

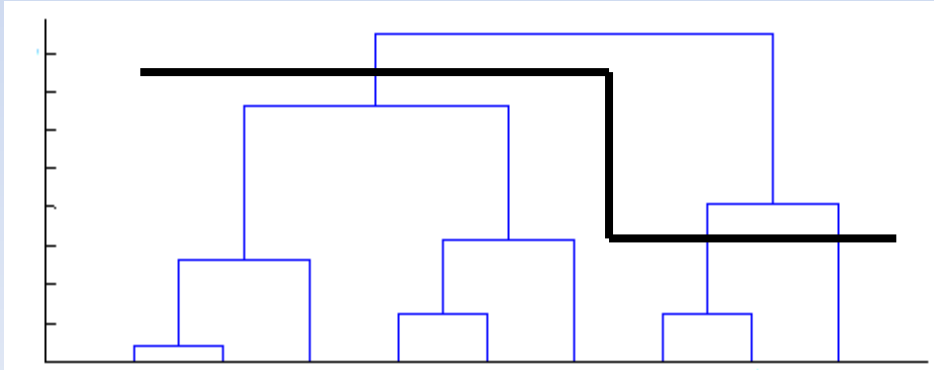












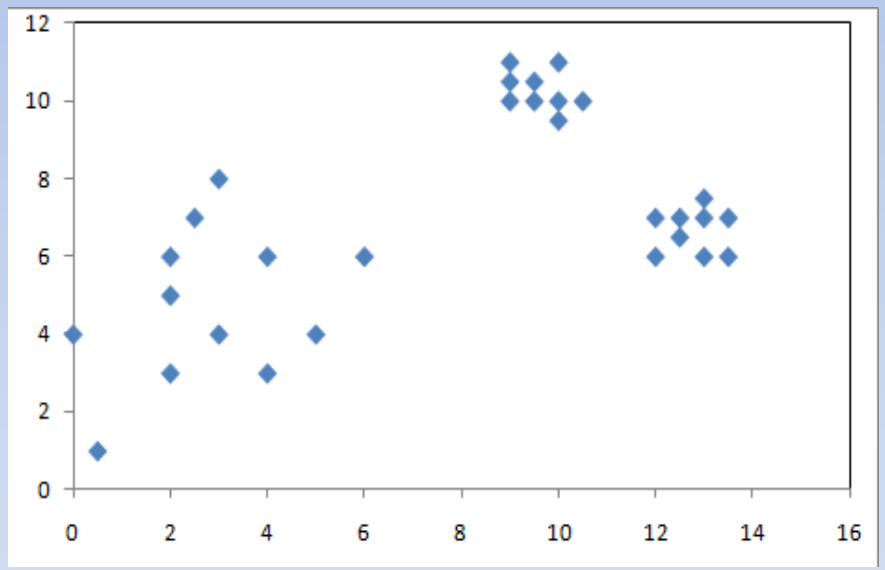
Procedimento Metodológico

Desenvolvimento/adaptação de códigos em *Matlab* e R que incluem:

- A obtenção de hierarquias, abordagem tradicional de classificação hierárquica, considerando:
  - Medida de proximidade - Distância Euclidiana
  - 3 métodos de agregação: *sl*, *cl*,*al*.
- A obtenção de partições, para os diferentes métodos de agregação usando, i) a abordagem SEP/COP,
  ii) a abordagem tradicional com o nº de classes da partição de referência.
- Comparação das partições – Validação externa e validação relativa, usando o índice ARI.

Dados simulados – Exemplo 1

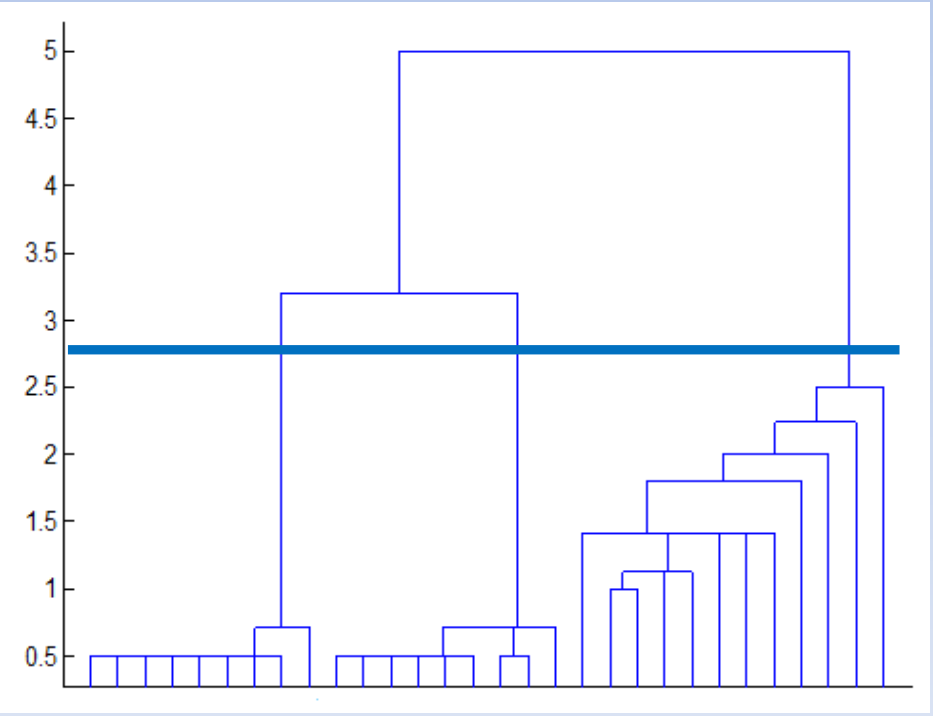
n=30, p=2, k=3

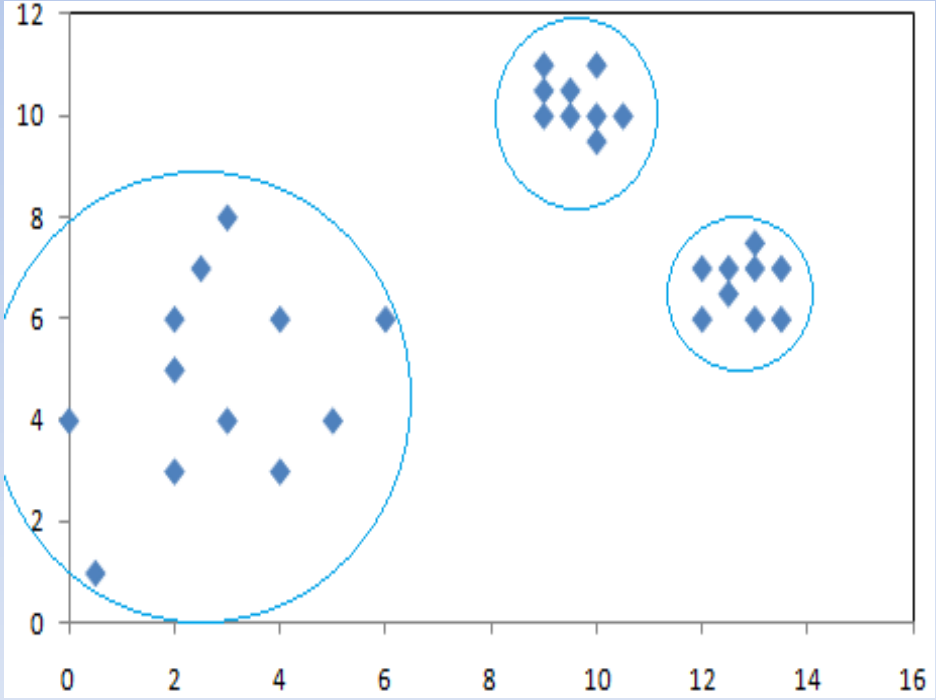


- **Na abordagem tradicional**: obtenção das hierarquias e respectivas partições (fixado o n.º de classes).
- **Na abordagem SEP/COP**: obtenção da melhor partição, com base no valor do índice COP.
- Comparação das partições obtidas (ARI).

Dados simulados - Exemplo 1

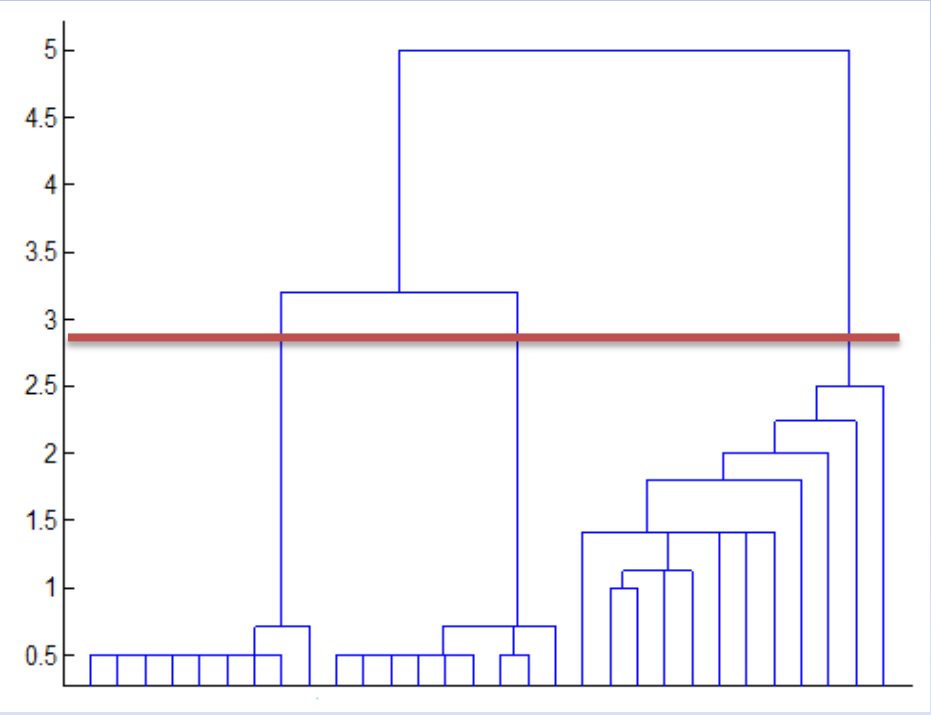
Abordagem tradicional "*single linkage*"

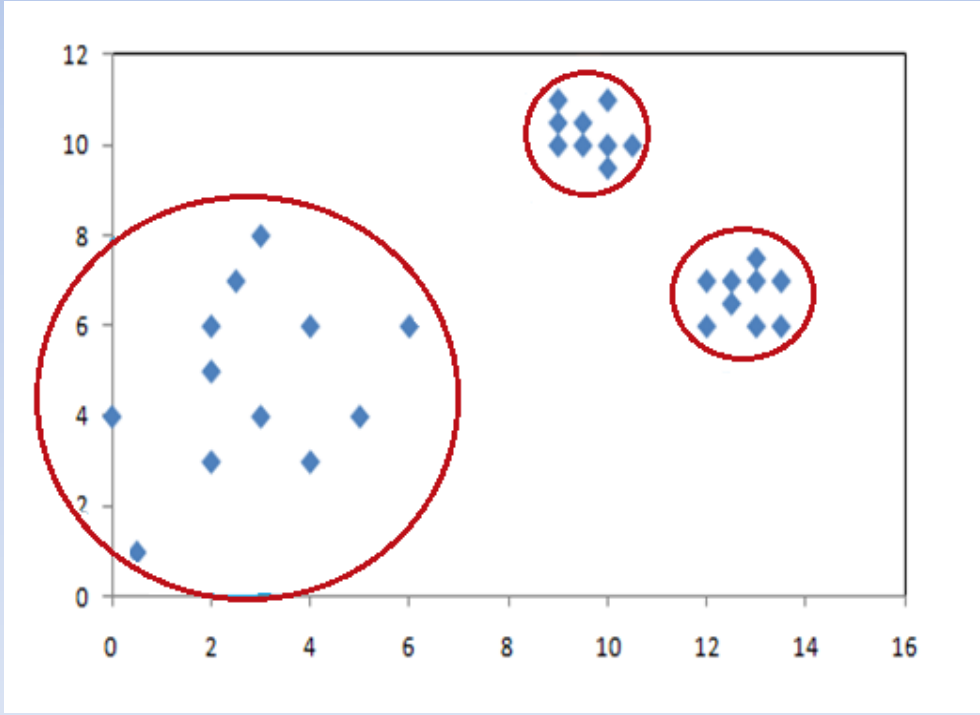




Dados simulados - Exemplo 1

Abordagem SEP/COP "*single linkage*"



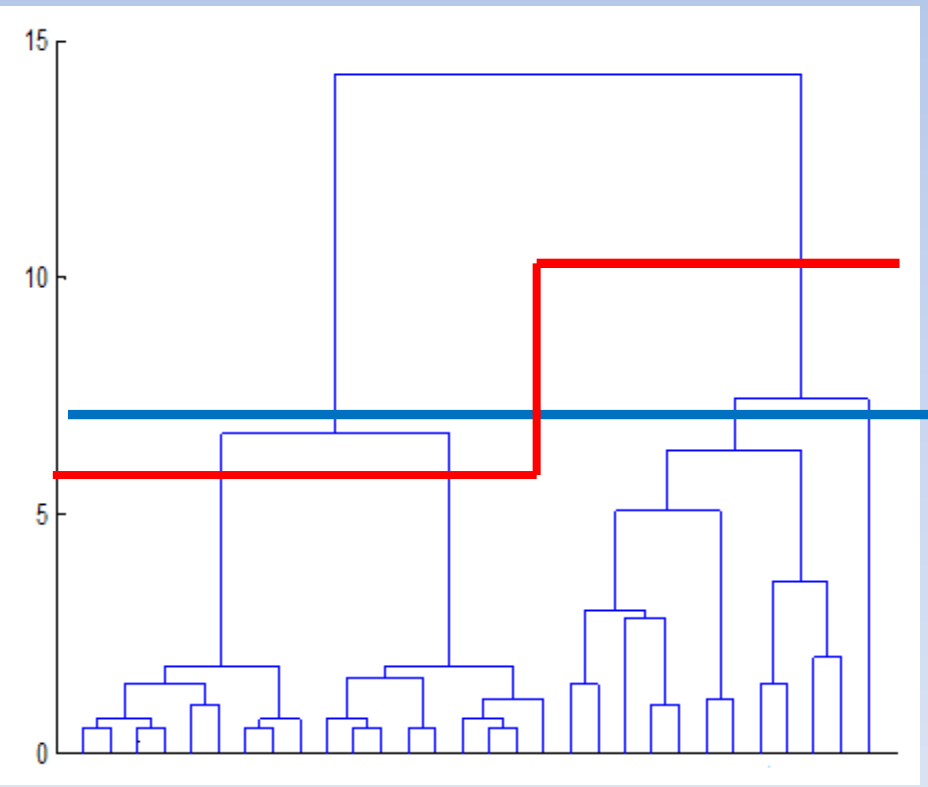


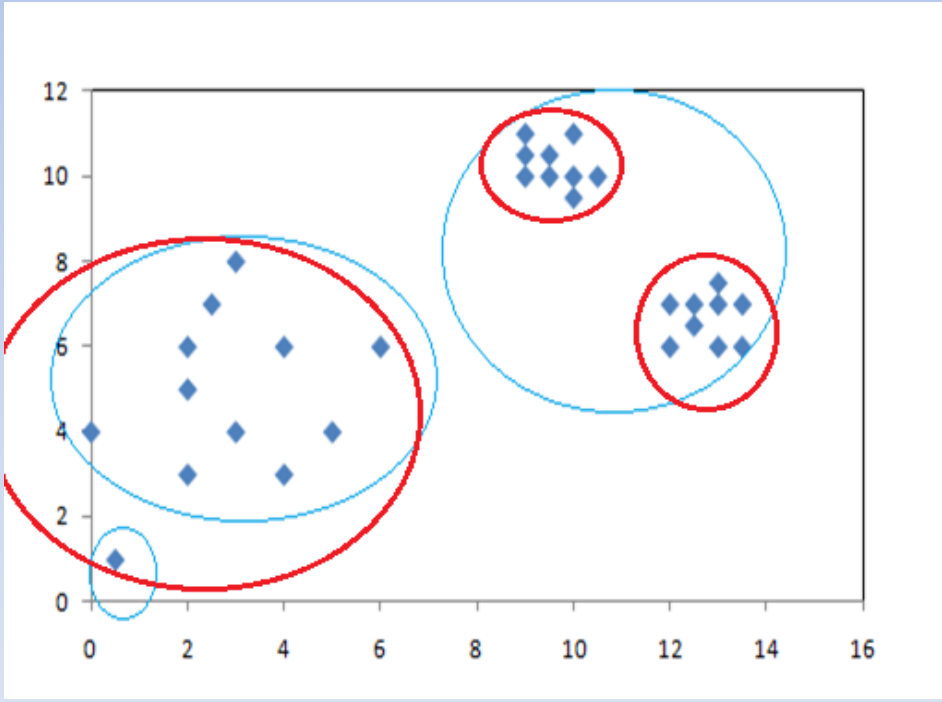
ARI=1

COP= 0.15266

Dados simulados - Exemplo 1

Abordagens tradicional e SEP/COP "*complete linkage*"





ARI=0.5701

COP= 0.15266

Experiências – Dados simulados

Experiência de dados simulados:

- Para k=1 a 1000:
  - ✓Gerar aleatoriamente 1 partição de referência de acordo com a distribuição pré-estabelecida.
  - ✓Aplicação das abordagens tradicional e SEP/COP (para os vários critérios de agregação).
  - ✓Comparação das partições obtidas com a partição de referência (ARI).
- Cálculo dos valores médio e desvio-padrão dos valores de ARI.
- Contagem do n.º de vezes que cada algoritmo, fornece a partição de referência(ARI=1).

Experiências – Dados simulados

Bases de dados simulados:

- A partir da distribuição Binormal
- 3 classes
- Variação na cardinalidade das classes
- Diferentes níveis de ruído introduzido: 4% e 10% de novos elementos a classificar.

Dados Simulados	n	Cardinais	% Ruído
Experiência 1	110	10,50,50	0%
Experiência 2	60	20,20,20	0%
Experiência 3	150	50,50,50	0%, 4%, 10%

Experiências – Dados simulados

Partição natural em 3 classes, 2 classes mais próximas e uma mais afastada.

Caracterização das classes:

Caso I

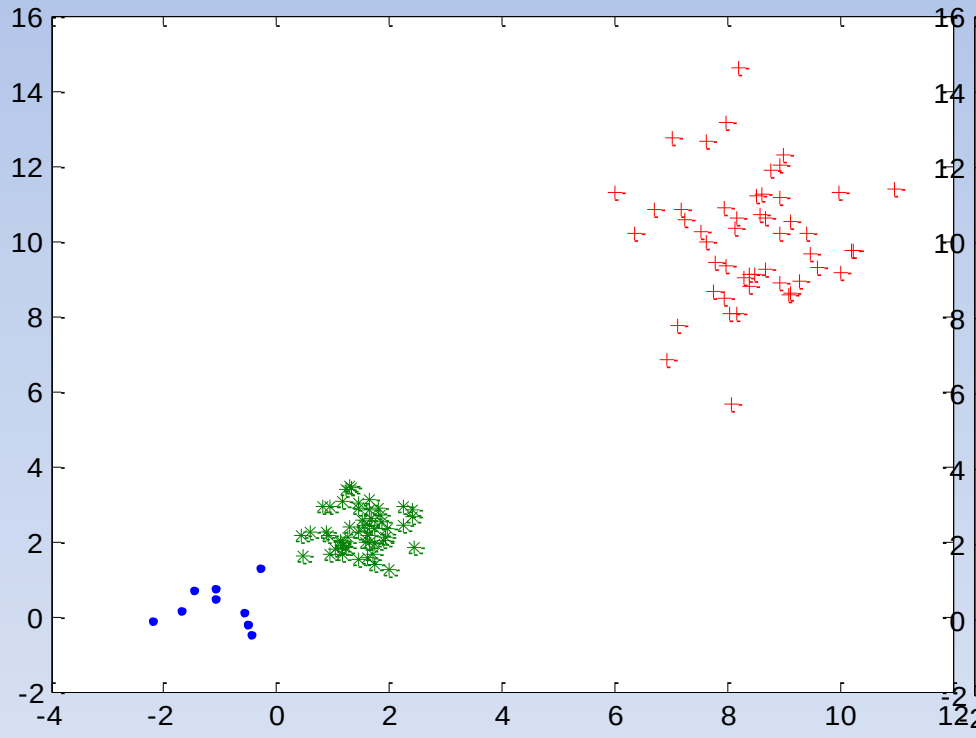
	μ_x	μ_y	σ_x^2	σ_y^2
C1	-1	0	0.3	0.3
C2	**1.5**	**2.5**	0.3	0.3
C3	8.5	10	1.5	2.25

Caso II

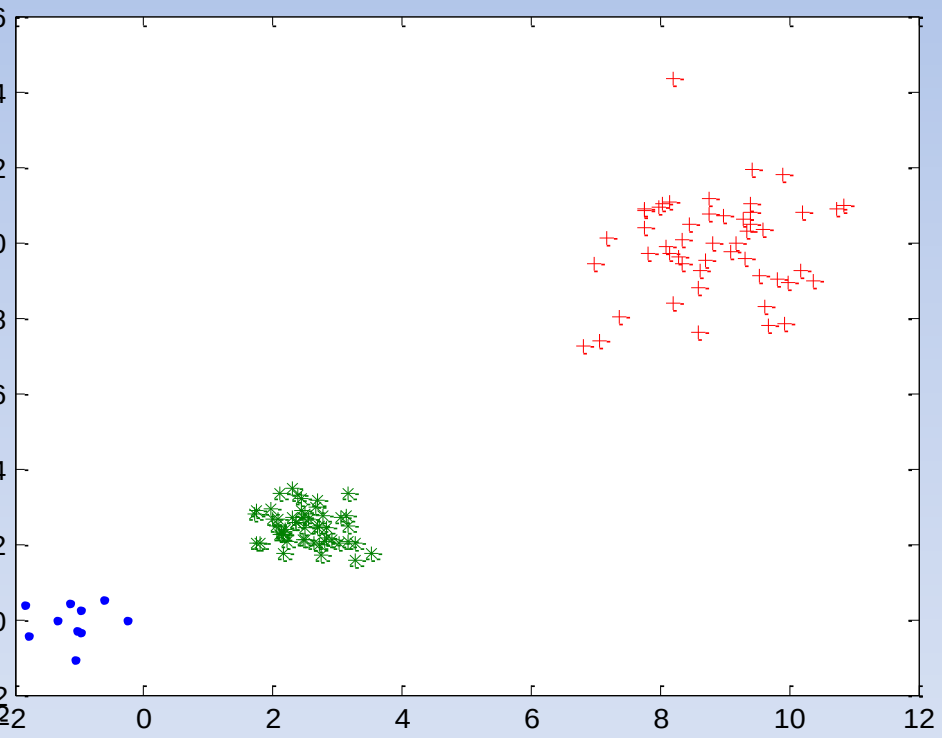
	μ_x	μ_y	σ_x^2	σ_y^2
C1	-1	0	0.3	0.3
C2	**2.5**	**2.5**	0.3	0.3
C3	8.5	10	1.5	2.25

Experiência1 – Dados simulados

Caso I

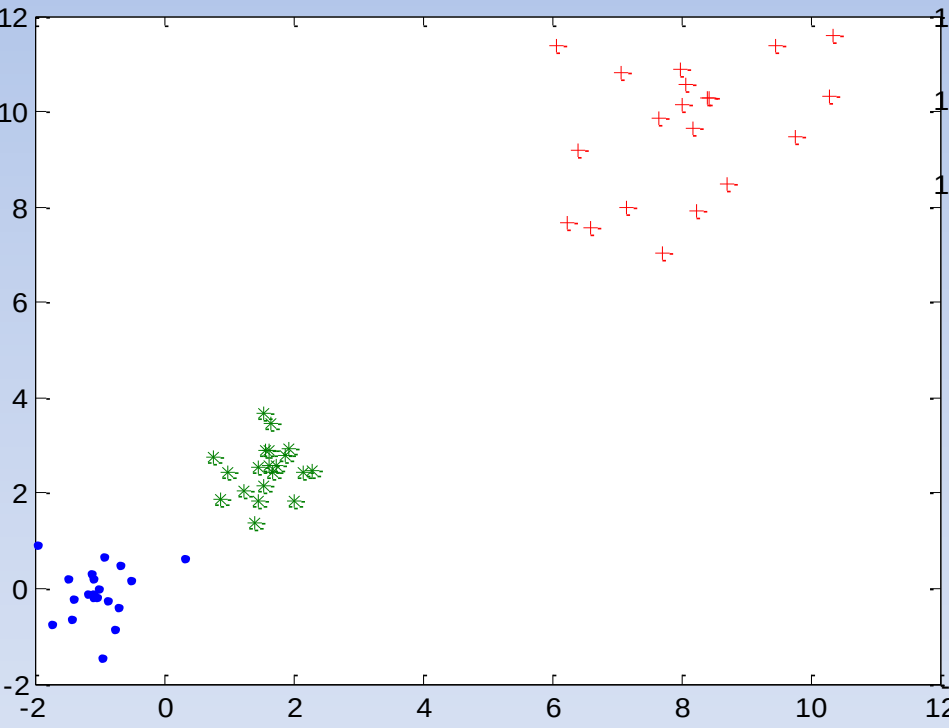


Caso II

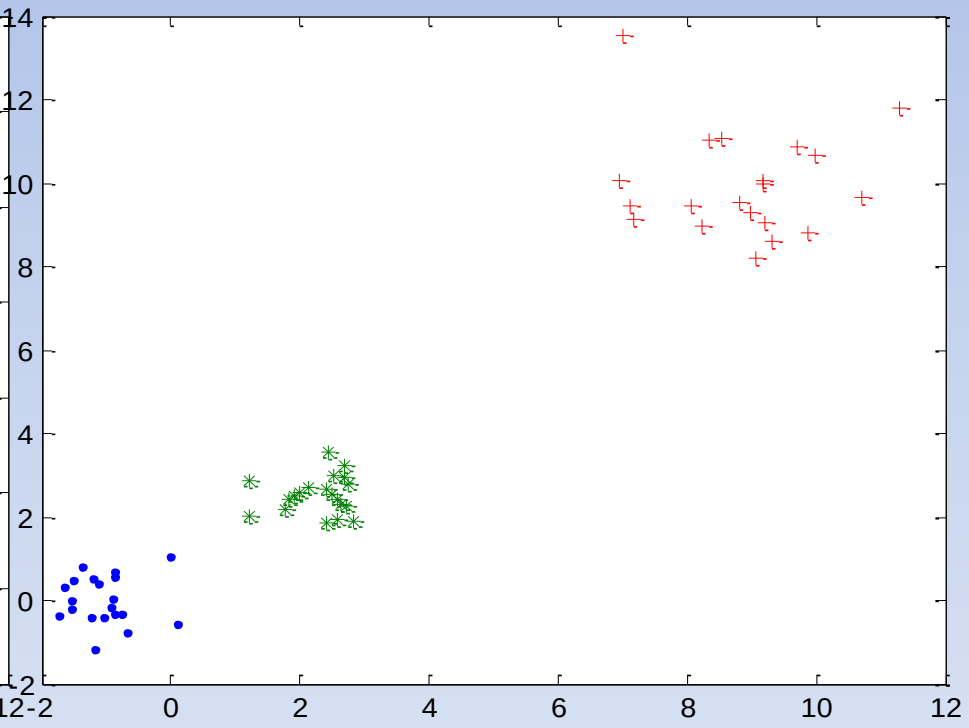


Experiência2 – Dados simulados

Caso I

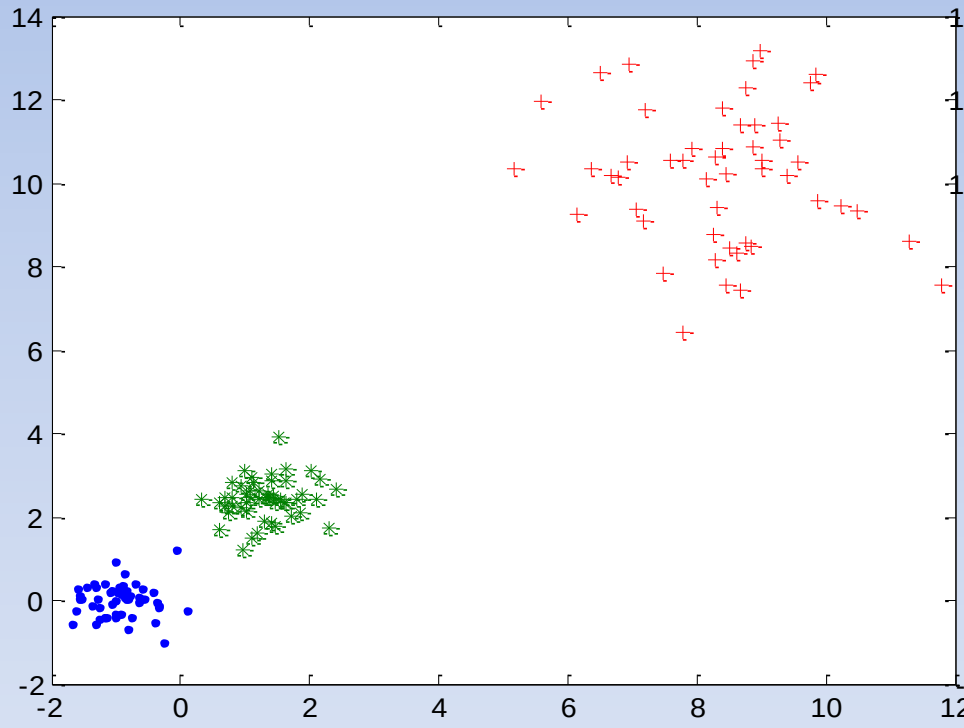


Caso II

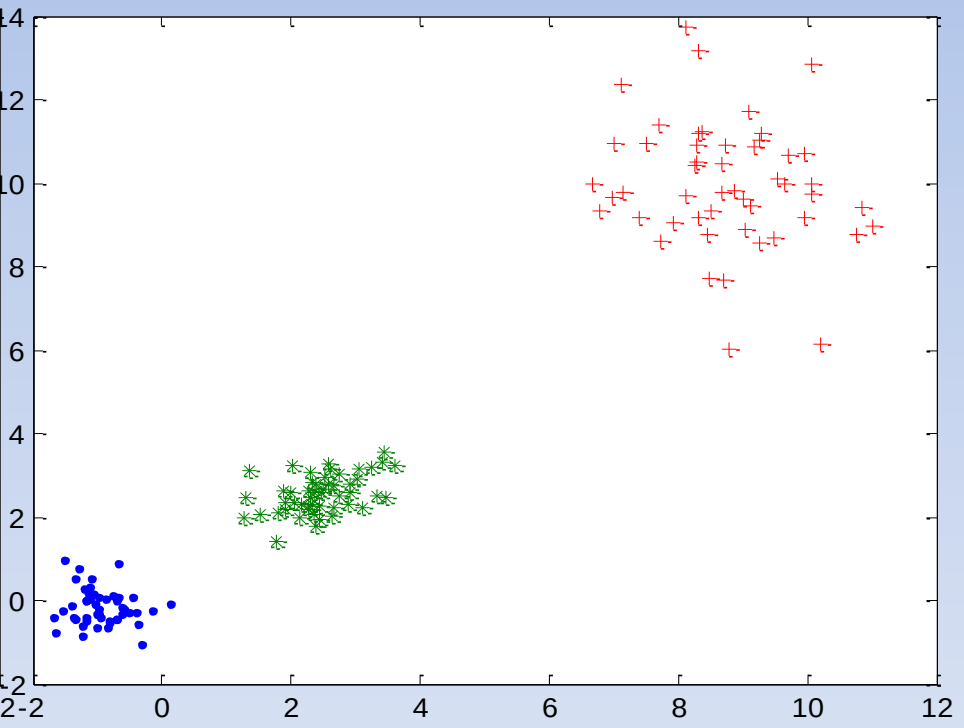


Experiência3 – Dados simulados

Caso I



Caso II



Resultados

Valores médios de ARI

		Cardinais 10, 50, 50		Cardinais 20, 20, 20	
		Tradicional	**SEP/COP**	**Tradicional**	**SEP/COP**
I	***sl***	0.9070 (0.0932)	0.8332 (0)	0.7266 (0.2306)	0.6578 (0.1802)
	cl	0.6688 (0.0717)	**0.8331 (0.0011)**	0.6114 (0.2391)	**0.6569 (0.1796)**
	al	0.8656 (0.0987)	0.8331 (0.0011)	0.7737 (0.2399)	0.6578 (0.1802)
II	***sl***	0.9755 (0.0626)	0.8543(0.0556)	0.9141 (0.1804)	**0.9929 (0.0549)**
	cl	0.7225 (0.1357)	**0.8544(0.0553)**	0.7655 (0.2645)	**0.9924 (0.0566)**
	al	0.9544 (0.0815)	0.8544 (0.0558)	0.9268 (0.1701)	**0.9925 (0.0565)**

Resultados

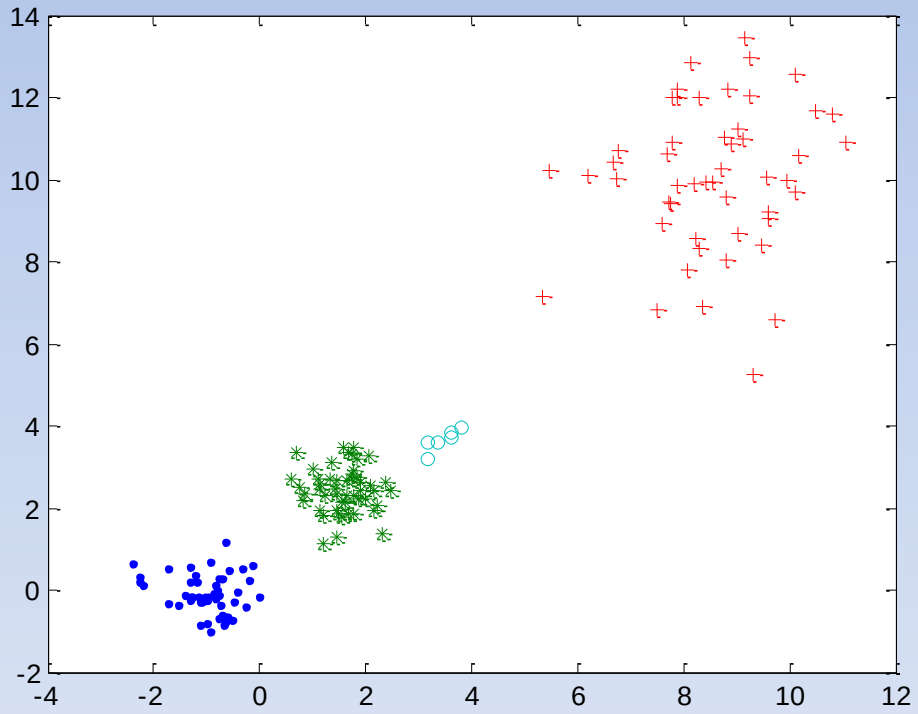
Valores médios de ARI

		Cardinais 50, 50, 50	
		Tradicional	**SEP/COP**
I	***sl***	0.6660 (0.1915)	0.6307 (0.1521)
	cl	0.4959 (0.1205)	**0.6307 (0.1521)**
	al	0.6982 (0.2148)	0.6299 (0.1513)
II	***sl***	0.8898 (0.1914)	**0.9981 (0.0273)**
	cl	0.6116 (0.2361)	**0.9976 (0.0305)**
	al	0.8843 (0.1952)	**0.9981 (0.0273)**

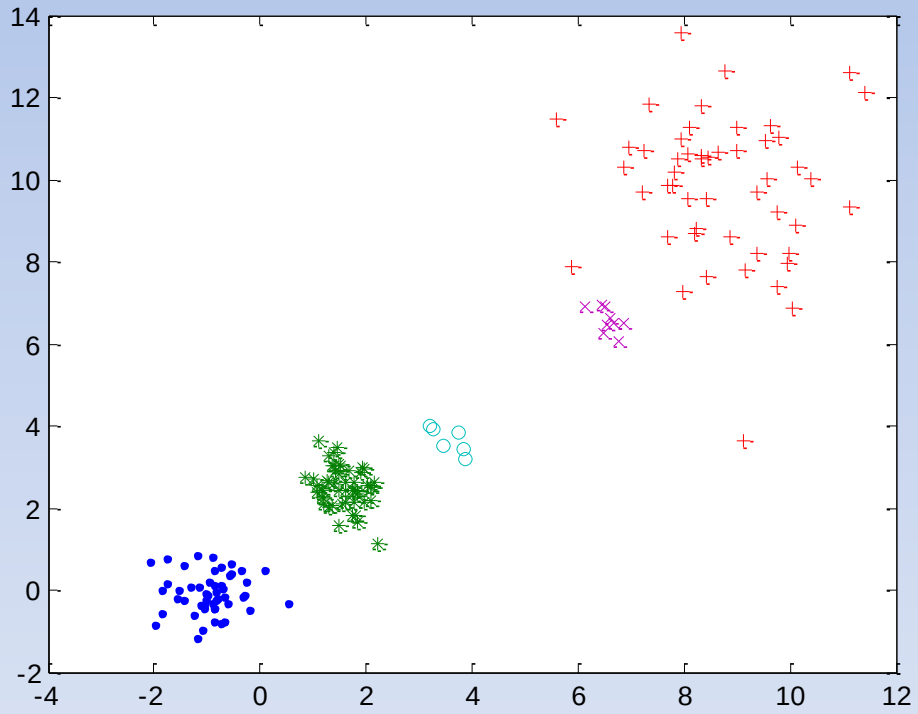
Experiência3 – Dados simulados com ruído

Caso I

4% ruído



10% ruído



Resultados

Valores médios de ARI

	4% ruído		10% ruído	
	Tradicional	**SEP/COP**	**Tradicional**	**SEP/COP**
sl	0.6601 (0.1978)	**0.7337 (0.2176)**	0.6804 (0.1870)	**0.9458 (0.1360)**
cl	0.7554 (0.2638)	0.7353 (0.2182)	0.5613 (0.1966)	**0.9567 (0.1242)**
al	0.7536 (0.2297)	0.7362 (0.2183)	0.5534 (0.1272)	**0.9551 (0.1262)**

Taxa de recuperação da partição exacta

		sl	***cl***	***al***
10% ruído	**Tradicional**	25.1%	15.5%	6.4%
	SEP/COP	**83.3%**	**86.8%**	**86.4%**

Experiência 4 – de acordo com [1]

Gerar 10 bases de dados, cada uma com 10 classes não sobrepostas.

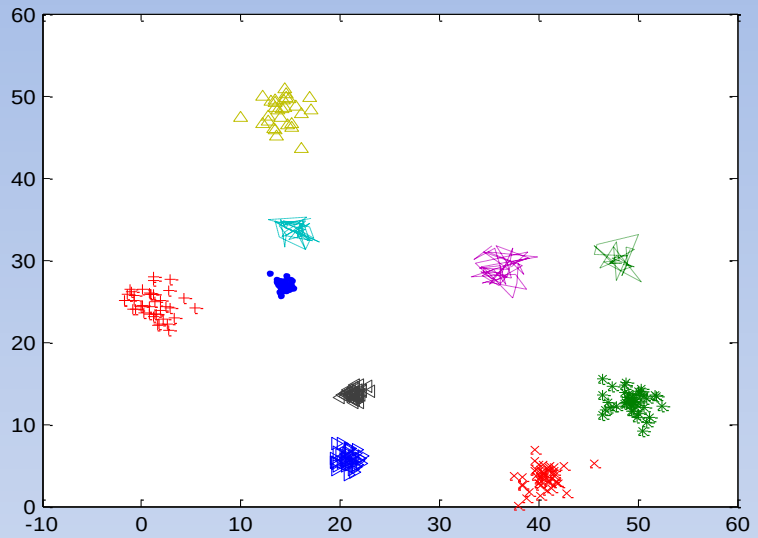
- Cada classe gerada a partir da distribuição binormal:
  - ✓Elementos do vector média são valores aleatórios entre 0 e 50
  - ✓As matrizes de covariâncias são da forma a*I, 0.1≤a≤0.3
  - ✓Os cardinais das classes são gerados aleatoriamente da uniforme entre 25 e 50
- Sendo μ_k e μ_l as médias de 2 classes, $\forall k, l$

$$d(\mu_k, \mu_l) > 3 * (a_k + a_l)$$

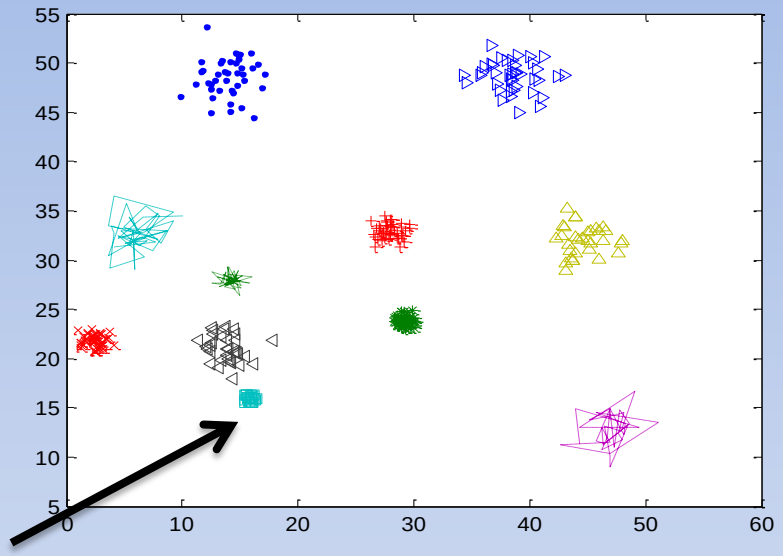
Dados Simulados	Classes	% Ruído
Experiência 4	10	0%, 5%, 10%, 20%

Experiência 4 – de acordo com [1]

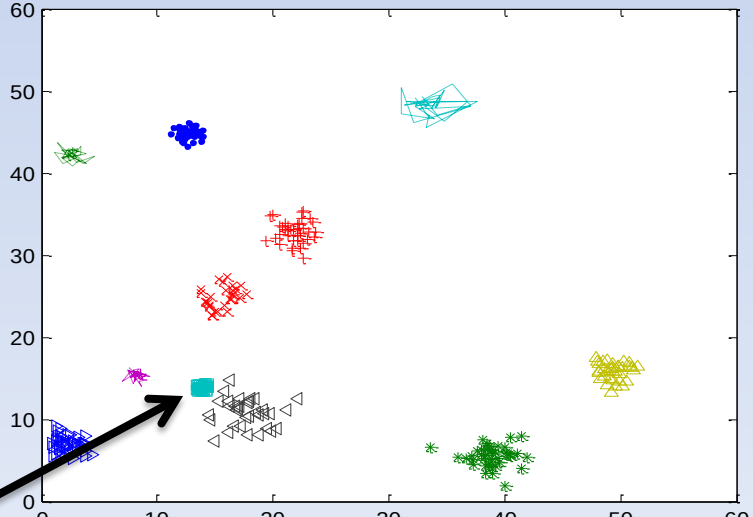
0% ruído



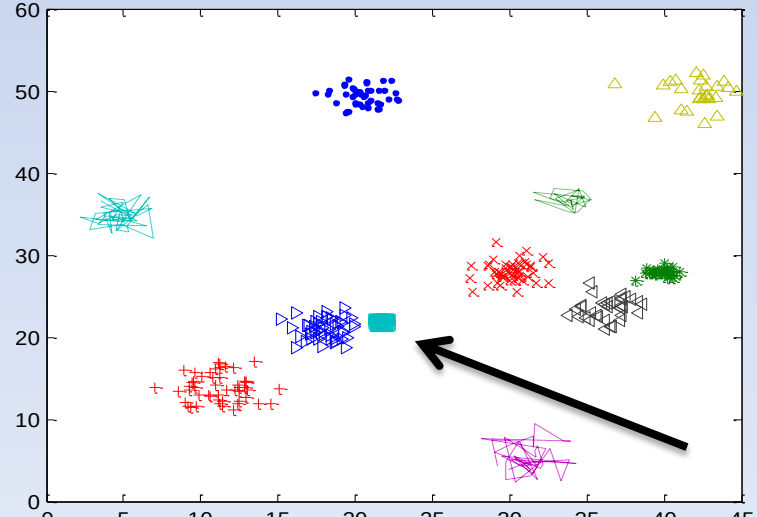
5% ruído



10% ruído



20% ruído



Resultados

Valores médios de ARI

	Tradicional	SEP/COP	Tradicional	SEP/COP
	0% ruído		5% ruído	
sl	0.9825 (0.0390)	0.9826 (0.0368)	0.8530 (0.0828)	**0.9306 (0.0467)**
cl	0.9873 (0.0401)	**0.9896 (0.0279)**	0.9102 (0.0549)	0.9024 (0.0719)
al	0.9886 (0.0361)	0.9885 (0.0275)	0.9066 (0.0357)	0.9024 (0.0719)
	10% ruído		20% ruído	
sl	0.8628 (0.0748)	**0.8916 (0.0579)**	0.7362 (0.0517)	**0.8560 (0.0650)**
Cl	0.8616 (0.0746)	**0.8914 (0.0522)**	0.7490 (0.0427)	**0.8504 (0.0693)**
Al	0.8608 (0.0750)	**0.8987 (0.0472)**	0.7468 (0.0460)	**0.8560 (0.0650)**

Análise dos resultados

- Na Experiência 1, partição de referência com classes de diferentes cardinalidades e diferentes separabilidades (Casos I e II), os algoritmos SEP/COP têm melhor desempenho que os algoritmos tradicionais no método *complete linkage* (coeficiente ARI com maiores valores médios e menores desvios padrões).

- Os resultados obtidos na abordagem SEP/COP dependem pouco do critério de agregação usado.

Análise dos resultados

- Nas Experiências 2 e 3, partições de referência com classes de iguais cardinalidades os algoritmos SEP/COP têm sempre melhor desempenho que os algoritmos tradicionais quando usado o critério *complete linkage*.
- Quando a separabilidade das classes aumenta (Caso II), os algoritmos SEP/COP têm melhor desempenho que os algoritmos tradicionais em todos os critérios de agregação.
- Na presença de ruído, o bom desempenho da abordagem SEP/COP é ainda mais notória (elevados valores de taxa de recuperação da partição de referência).

Análise dos resultados

- Na Experiência 4, partição de referência com classes de diferentes cardinalidades, homogeneidades e separabilidades, os algoritmos SEP/COP têm, em geral, melhor desempenho que os algoritmos tradicionais com todos os critérios de agregação, em particular quando se aumentam os níveis de ruído.

Comentários finais

- O objectivo deste trabalho foi o da comparação da abordagem tradicional com a abordagem SEP/COP, recentemente proposta, para a escolha da melhor partição aquando da interpretação de uma hierarquia.
- As duas abordagens foram implementadas computacionalmente recorrendo às linguagens Matlab e R.
- Realizaram-se experiências com dados simulados para a comparação do desempenho das duas abordagens.

Comentários finais

- Estas experiências não permitiram escolher uma das abordagens, já que nenhuma das abordagens se revelou sistematicamente melhor.
- Os algoritmos SEP/COP mostraram:
  - ✓ ser uma boa solução aquando de situações de equicardinalidade
  - ✓ depender pouco do critério de agregação usado
  - ✓ ser mais robustos à presença de ruído.
- Os resultados aqui encontrados vêm de encontro ao que é conhecido de validação em classificação, sendo uma área particularmente difícil de tirar conclusões genericamente válidas.

Agradecimento

- Ao professor Ibai Gurrutxaga, agradecemos o interesse, disponibilidade para trocas de impressões e o envio do código SEP/COP.

Referências

[1] Gurrutxaga I., Albisua I., Arbelaitz O., Martin J., Muguerza J., Perez J., Perona I. "SEP/COP: An efficient method to find the best partition in the hierarchical clustering based on a new cluster validity index". Pattern Recognition 43 (2010), pp. 3364-3373.

[2] Sousa F. "Novas metodologias e validação em classificação hierárquica ascendente". Tese de doutoramento (2000).

[3] Halkidi M., Batistakis Y., Vazirgiannis M. "On Clustering Validation Techniques". (2001).

[4] Halkidi M. e Vazirgiannis M. "Clustering Validity Assessment: Finding the optimal partitioning of a data set". (2001). ICDM 2001, pp. 187-197.

[5] Hubert L., Arabie P. "Comparing Partitions". Journal of Classification 2 (1985), pp. 193-218.

[6] Fred A.,Leitão J. " A New Cluster Isolation Criterion Based on Dissimilarity Increments". IEEE Transactions on pattern analysis and machine intelligence vol.25 no.8 (2003).