

Aprendizado de Máquina Não Supervisionado

João Carlos Xavier Júnior

jcxavier@imd.ufn.br

Aprendizado não Supervisionado

- ❑ O que significa **Aprendizado não Supervisionado**?
 - ❖ Nesse tipo de aprendizado, os dados **não** possuem **atributo classe**.
- ❑ Você sabe o que significa **atributo classe**?

Aprendizado não Supervisionado

□ Exemplo: dados **sem** classe

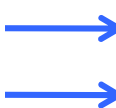


Nº	Idade_N	Gênero	Estado Civil	Filhos_N	Escolaridade	Renda_N	Cartão_Cr.	Imóvel_P	Classe
1	0,4333	Masc	Divorciado	0,5000	Superior	0,4565	Sim	Sim	
2	0,3000	Femi	Solteiro	0,0000	Médio	0,2935	Sim	Não	
3	1,0000	Masc	Viúvo	1,0000	Fundamental	1,0000	Sim	Não	
4	0,0333	Femi	Casado	0,5000	Superior	0,0761	Não	Sim	
5	0,7667	Femi	Casado	0,2500	Superior	0,2283	Sim	Sim	
6	0,5667	Masc	Casado	0,7500	Médio	0,2500	Sim	Não	
7	0,1333	Femi	Solteiro	0,2500	Superior	0,3696	Sim	Não	
8	0,2333	Femi	Casado	0,7500	Pós-graduação	0,7283	Sim	Sim	
9	0,3667	Masc	Divorciado	0,0000	Superior	0,5217	Não	Não	
10	0,0000	Masc	Solteiro	0,0000	Médio	0,0000	Não	Não	

CreditoPessoal.csv

Aprendizado não Supervisionado

❑ Exemplo: dados **com** classe



#	preg	plas	pres	skin	insu	mass	pedi	age	class
1	6	148	72	35	0	33,6	0,627	50	positive
2	1	85	66	29	0	26,6	0,351	31	negative
3	8	183	64	0	0	23,3	0,672	32	positive
4	1	89	66	23	94	28,1	0,167	21	negative
5	0	137	40	35	168	43,1	2,288	33	positive
6	5	116	74	0	0	25,6	0,201	30	negative
7	3	78	50	32	88	31	0,248	26	positive
8	10	115	0	0	0	35,3	0,134	29	negative
9	2	197	70	45	543	30,5	0,158	53	positive
10	8	125	96	0	0	0	0,232	54	positive
11	4	110	92	0	0	37,6	0,191	30	negative
12	10	168	74	0	0	38	0,537	34	positive
13	10	139	80	0	0	27,1	1,441	57	negative
14	1	189	60	23	846	30,1	0,398	59	positive
15	5	166	72	19	175	25,8	0,587	51	positive

Diabetes.csv

O que é um Agrupamento?

- ❑ **Clustering** ou **agrupamento** é uma técnica de aprendizado **não-supervisionado**, ou seja, quando não há uma **classe associada** a cada exemplo.
- ❑ As instâncias de uma base de dados são colocadas em **clusters** (**grupos**), que normalmente descrevem algum mecanismo existente no processo que as gerou.
- ❑ Dessa forma, algumas instâncias são mais **similares** entre si do que as restantes.

O que é um Agrupamento?

- ❑ Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles.
- ❑ Utilizado para encontrar padrões inesperados nos dados.
- ❑ Inerentemente é um problema não **definido claramente**.

- ❑ Como separar os animais abaixo? Qual o critério?



Descrição do Problema



Conhecimento do Usuário/Especialista

Interpretação

Validação

Algoritmo
de
Clustering

Escolha dos
objetos e seus
atributos

Calculando a Distância

- ❑ A **distância** é o método mais natural para **dados numéricos**.
- ❑ Valores pequenos indicam **maior similaridade**.
- ❑ Métricas de Distância:
 - ❖ Euclideana
 - ❖ Manhattan
 - ❖ Etc.
- ❑ Não generaliza muito bem para dados não numéricos.
 - ❖ **Qual a distância entre “masculino” e “feminino”?**

Calculando a Distância

□ Base normalizada:

$$Dist_E = \sqrt{\sum_{r=1}^m (x_{i,r} - x_{j,r})^2}$$

Nº	Idade_N	Gênero	Estado Civil	Filhos_N	Escolaridade	Renda_N	Cartão_Cr.	Imóvel_P
1	0,4333	Masc	Divorciado	0,5000	Superior	0,4565	Sim	Sim
2	0,3000	Femi	Solteiro	0,0000	Médio	0,2935	Sim	Não
3	1,0000	Masc	Viúvo	1,0000	Fundamental	1,0000	Sim	Não
4	0,0333	Femi	Casado	0,5000	Superior	0,0761	Não	Sim
5	0,7667	Femi	Casado	0,2500	Superior	0,2283	Sim	Sim
6	0,5667	Masc	Casado	0,7500	Médio	0,2500	Sim	Não
7	0,1333	Femi	Solteiro	0,2500	Superior	0,3696	Sim	Não
8	0,2333	Femi	Casado	0,7500	Pós-graduação	0,7283	Sim	Sim
9	0,3667	Masc	Divorciado	0,0000	Superior	0,5217	Não	Não
10	0,0000	Masc	Solteiro	0,0000	Médio	0,0000	Não	Não



Calculando a Distância

□ Resultado:

[1-2]	0,0178	1,0000	1,0000	0,2500	1,0000	1,0000	0,0266	0,0000	1,0000	5,2944	2,3009
[1-3]	0,3211	0,0000	1,0000	0,2500	1,0000	0,0000	0,2954	0,0000	1,0000	3,8665	1,9663
[1-4]	0,1600	1,0000	1,0000	0,0000	0,0000	1,0000	0,1447	1,0000	0,0000	4,3047	2,0748
[1-5]	0,1111	1,0000	1,0000	0,0625	0,0000	0,0000	0,0521	0,0000	0,0000	2,2257	1,4919
[1-6]	0,0178	0,0000	1,0000	0,0625	1,0000	1,0000	0,0427	0,0000	1,0000	4,1229	2,0305
[1-7]	0,0900	1,0000	1,0000	0,0625	0,0000	0,0000	0,0076	0,0000	1,0000	3,1601	1,7777
[1-8]	0,0400	1,0000	1,0000	0,0625	1,0000	1,0000	0,0738	0,0000	0,0000	4,1763	2,0436
[1-9]	0,0044	0,0000	0,0000	0,2500	0,0000	0,0000	0,0043	1,0000	1,0000	2,2587	1,5029
[1-10]	0,1878	0,0000	1,0000	0,2500	1,0000	1,0000	0,2084	1,0000	1,0000	5,6462	2,3762
[2-3]	0,4900	1,0000	1,0000	1,0000	1,0000	1,0000	0,4992	0,0000	0,0000	5,9892	2,4473
[2-4]	0,0711	0,0000	1,0000	0,2500	1,0000	0,0000	0,0473	1,0000	1,0000	4,3684	2,0901
[2-5]	0,2178	0,0000	1,0000	0,0625	1,0000	1,0000	0,0043	0,0000	1,0000	4,2845	2,0699
[2-6]	0,0711	1,0000	1,0000	0,5625	0,0000	0,0000	0,0019	0,0000	0,0000	2,6355	1,6234
[2-7]	0,0278	0,0000	0,0000	0,0625	1,0000	1,0000	0,0058	0,0000	0,0000	2,0961	1,4478
[2-8]	0,0044	0,0000	1,0000	0,5625	1,0000	0,0000	0,1890	0,0000	1,0000	3,7560	1,9380
[2-9]	0,0044	1,0000	1,0000	0,0000	1,0000	1,0000	0,0521	1,0000	0,0000	5,0565	2,2487
[2-10]	0,0900	1,0000	0,0000	0,0000	0,0000	0,0000	0,0861	1,0000	0,0000	2,1761	1,4752
[3-4]	0,9344	1,0000	1,0000	0,2500	1,0000	1,0000	0,8536	1,0000	1,0000	8,0381	2,8351
[3-5]	0,0544	1,0000	1,0000	0,5625	1,0000	0,0000	0,5956	0,0000	1,0000	5,2125	2,2831
[3-6]	0,1878	0,0000	1,0000	0,0625	1,0000	1,0000	0,5625	0,0000	0,0000	3,8128	1,9526
[3-7]	0,7511	1,0000	1,0000	0,5625	1,0000	0,0000	0,3974	0,0000	0,0000	4,7111	2,1705
[3-8]	0,5878	1,0000	1,0000	0,0625	1,0000	1,0000	0,0738	0,0000	1,0000	5,7241	2,3925
[3-9]	0,4011	0,0000	1,0000	1,0000	1,0000	0,0000	0,2287	1,0000	0,0000	4,6298	2,1517
[3-10]	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	7,0000	2,6458
[4-5]	0,5378	0,0000	0,0000	0,0625	0,0000	1,0000	0,0232	1,0000	0,0000	2,6234	1,6197
[4-6]	0,2844	1,0000	0,0000	0,0625	1,0000	0,0000	0,0302	1,0000	1,0000	4,3772	2,0922
[4-7]	0,0100	0,0000	1,0000	0,0625	0,0000	1,0000	0,0861	1,0000	1,0000	4,1586	2,0393
[4-8]	0,0400	0,0000	0,0000	0,0625	1,0000	0,0000	0,4253	1,0000	0,0000	2,5278	1,5899
[4-9]	0,1111	1,0000	1,0000	0,2500	0,0000	1,0000	0,1986	0,0000	1,0000	4,5597	2,1353
[4-10]	0,0011	1,0000	1,0000	0,2500	1,0000	0,0000	0,0058	0,0000	1,0000	4,2569	2,0632

Calculando a Distância

□ Juntando (merging) objetos:

Nº	Idade_N	Gênero	Estado Civil	Filhos_N	Escolaridade	Renda_N	Cartão_Cr.	Imóvel_P
1	0,4333	Masc	Divorciado	0,5000	Superior	0,4565	Sim	Sim
2	0,3000	Femi	Solteiro	0,0000	Médio	0,2935	Sim	Não
3	1,0000	Masc	Viúvo	1,0000	Fundamental	1,0000	Sim	Não
4	0,0333	Femi	Casado	0,5000	Superior	0,0761	Não	Sim
5	0,7667	Femi	Casado	0,2500	Superior	0,2283	Sim	Sim
6	0,5667	Masc	Casado	0,7500	Médio	0,2500	Sim	Não
7	0,1333	Femi	Solteiro	0,2500	Superior	0,3696	Sim	Não
8	0,2333	Femi	Casado	0,7500	Pós-graduação	0,7283	Sim	Sim
9	0,3667	Masc	Divorciado	0,0000	Superior	0,5217	Não	Não
10	0,0000	Masc	Solteiro	0,0000	Médio	0,0000	Não	Não

Algoritmo Hierárquico ou Hierarchical clustering

Agrupamento Hierárquico

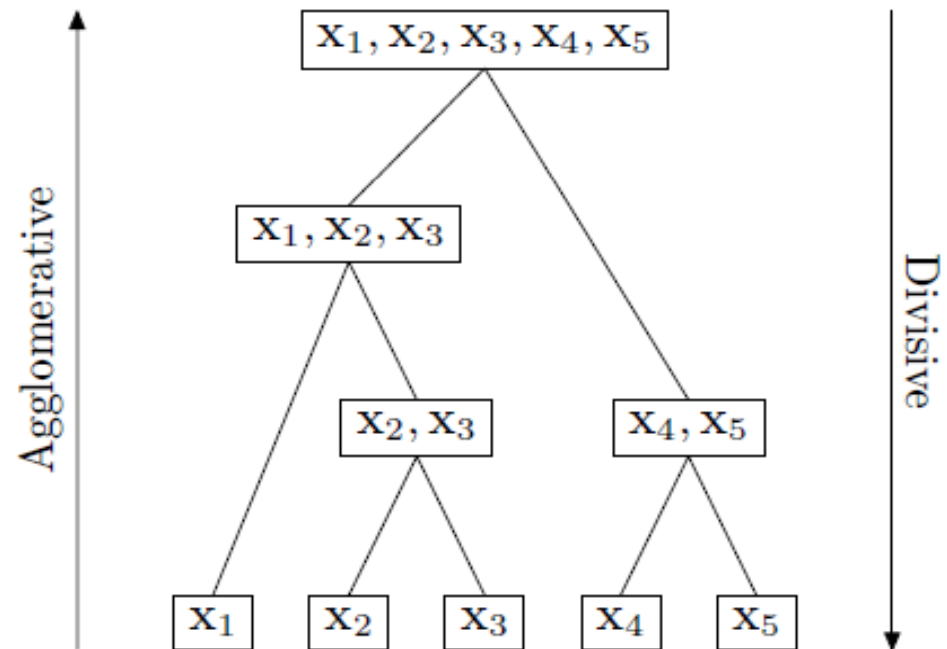
□ Envolvem a construção de uma hierarquia de uma estrutura do tipo árvore.

□ Tipos:

❖ Divisivo

❖ Aglomerativo

- Ligação Simples
- Ligação Completa
- Ligação Média.



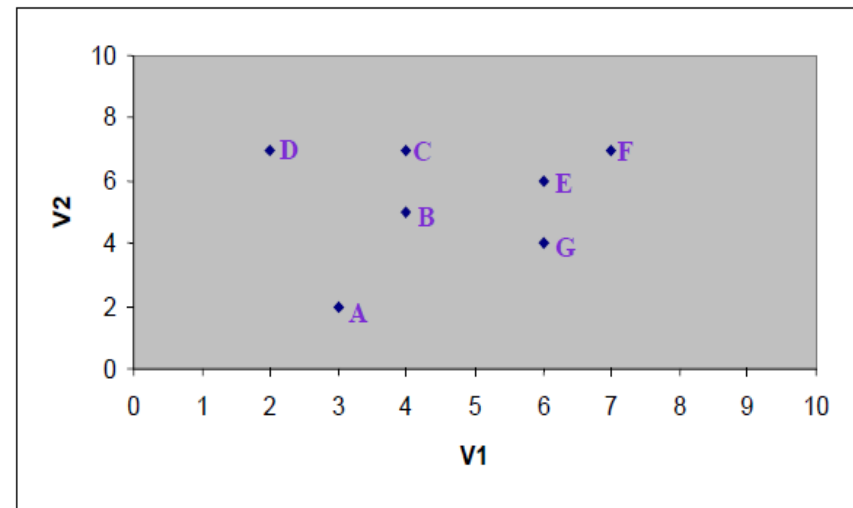
Métodos Aglomerativos

- ❑ Cada objeto começa como seu próprio grupo (cluster).
- ❑ Em passos seguintes, os dois grupos (ou objetos) mais próximos (similares) **são combinados** em um novo agregado.
 - ❖ O número de grupos é reduzido em uma unidade em cada passo.
- ❑ Ao final, todos os elementos são reunidos em um grande agregado.

Análise de agrupamentos

- ❑ **Tarefa:** identificar tipos de um determinado câncer com base na expressão gênica do tecido extraído do **tumor**.
- ❑ Uma pequena amostra de sete pacientes é selecionada.
 - ❖ A expressão gênica de dois genes (V_1 e V_2) foi medida para o tumor de cada paciente.

Paciente	V1	V2
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4



Análise de agrupamentos

- ❑ O objetivo principal da análise de agrupamentos é colocar as instâncias (objetos) mais **parecidas** ou **similares** em **grupos**.
- ❑ Como fazer???



Medida de Similaridade

❑ Distância Euclidiana:

- ❖ $d(A, B) = \text{SQRT}[(3-4)^2 + (2-5)^2] \rightarrow 3,162$
- ❖ $d(A, C) = \text{SQRT}[(3-4)^2 + (2-7)^2] \rightarrow 5,099$
- ❖ $d(C, F) = \text{SQRT}[(4-7)^2 + (7-7)^2] \rightarrow 3,000$
- ❖

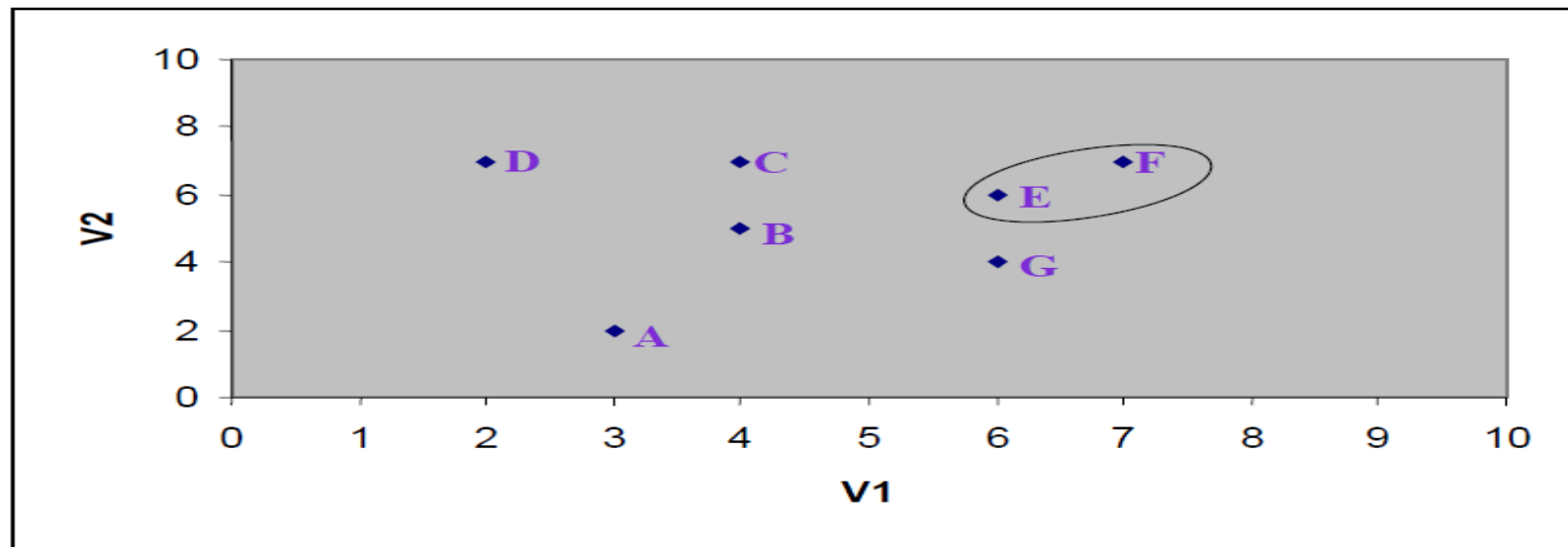
Paciente	V1	V2
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4

	A	B	C	D	E	F	G
A	0,0000						
B	3,1623	0,0000					
C	5,0990	2,0000	0,0000				
D	5,0990	2,8284	2,0000	0,0000			
E	5,0000	2,2361	2,2361	4,1231	0,0000		
F	6,4031	3,6056	3,0000	5,0000	1,4142	0,0000	
G	3,6056	2,2361	3,6056	5,0000	2,0000	3,1623	0,0000

Formação dos Grupos

□ Passo 1:

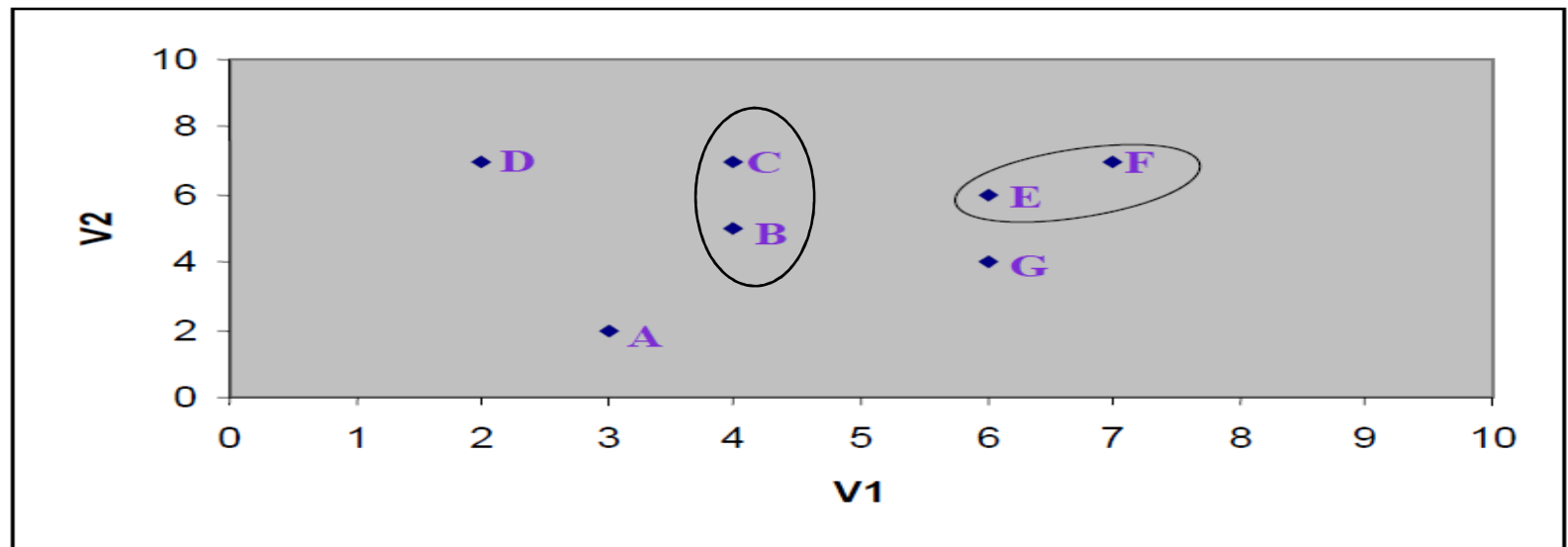
	A	B	C	D	E	F	G
A	0,0000						
B	3,1623	0,0000					
C	5,0990	2,0000	0,0000				
D	5,0990	2,8284	2,0000	0,0000			
E	5,0000	2,2361	2,2361	4,1231	0,0000		
F	6,4031	3,6056	3,0000	5,0000	1,4142	0,0000	
G	3,6056	2,2361	3,6056	5,0000	2,0000	3,1623	0,0000



Formação dos Grupos

Passo 2:

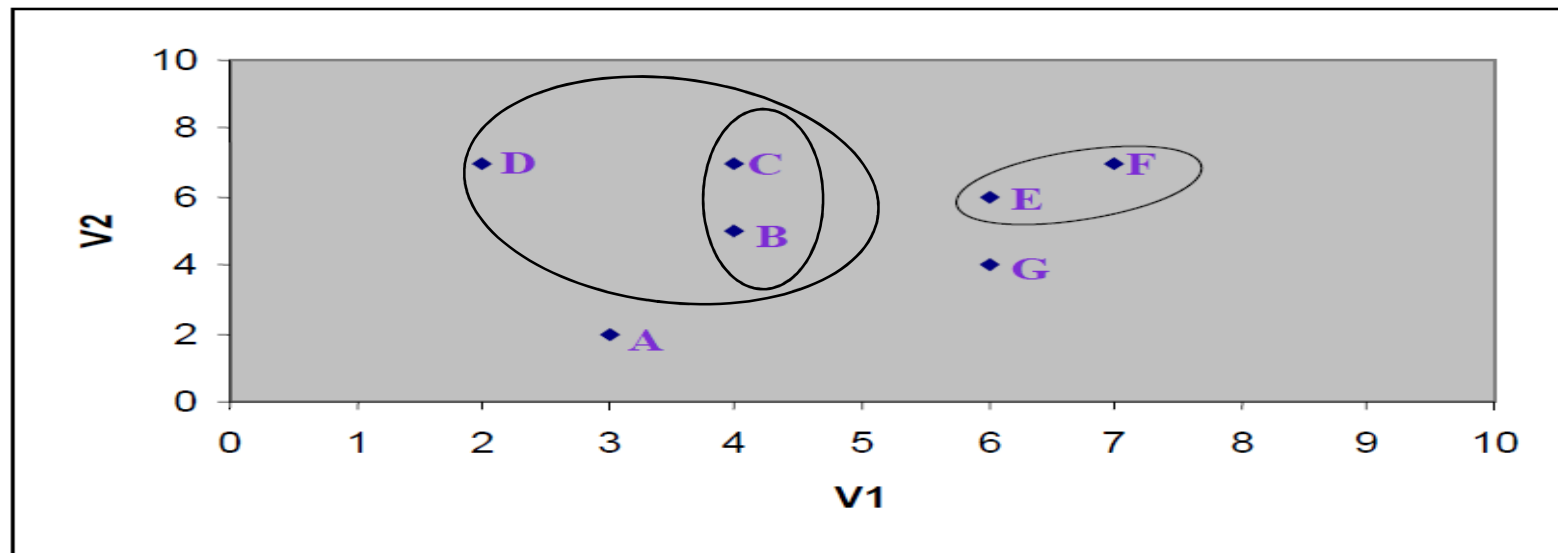
	A	B	C	D	EF
A	0,0000				
B	3,1623	0,0000			
C	5,0990	2,0000	0,0000		
D	5,0990	2,8284	2,0000	0,0000	
EF	5,7009	2,9155	2,5495	4,5277	0,0000
G	3,6056	2,2361	3,6056	5,0000	2,5495



Formação dos Grupos

□ Passo 3:

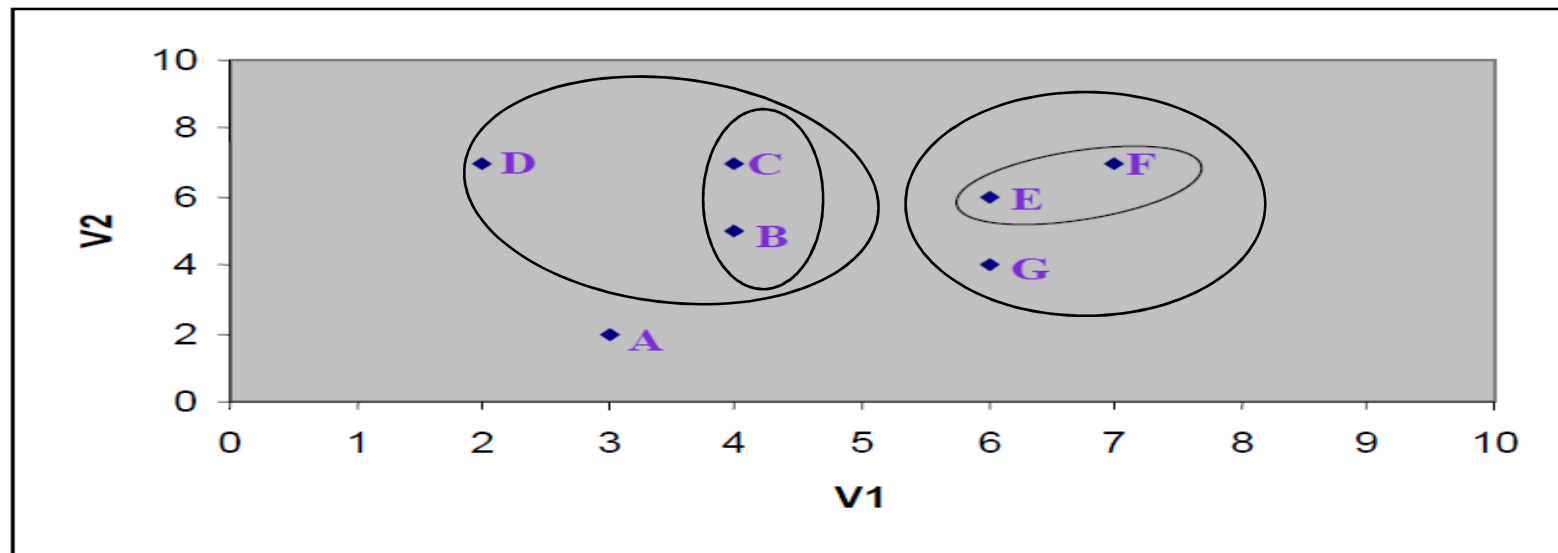
	A	BC	D	EF
A	0,0000			
BC	4,1231	0,0000		
D	5,0990	2,2361	0,0000	
EF	5,7009	2,5495	4,5277	0,0000
G	3,6056	2,8284	5,0000	2,5495



Formação dos Grupos

□ Passo 4:

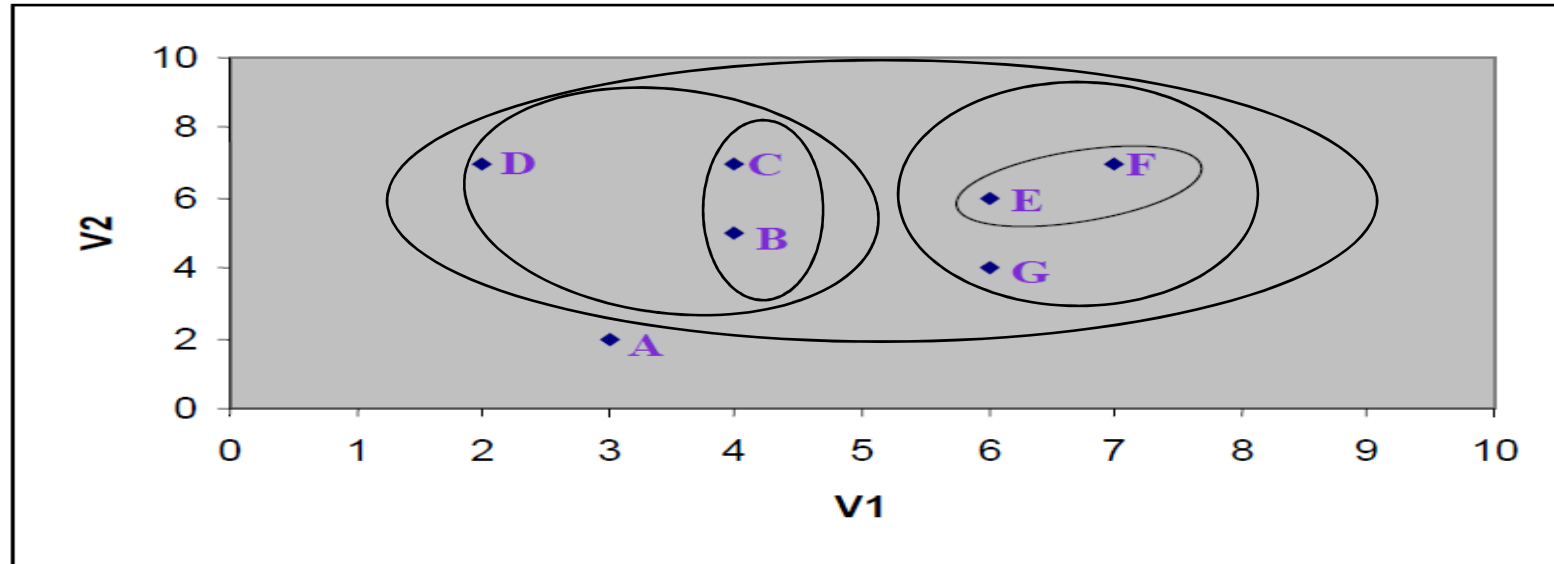
	A	BCD	EF
A	0,0000		
BCD	4,5000	0,0000	
EF	5,7009	3,5000	0,0000
G	3,6056	3,9051	2,5495



Formação dos Grupos

□ Passo 5:

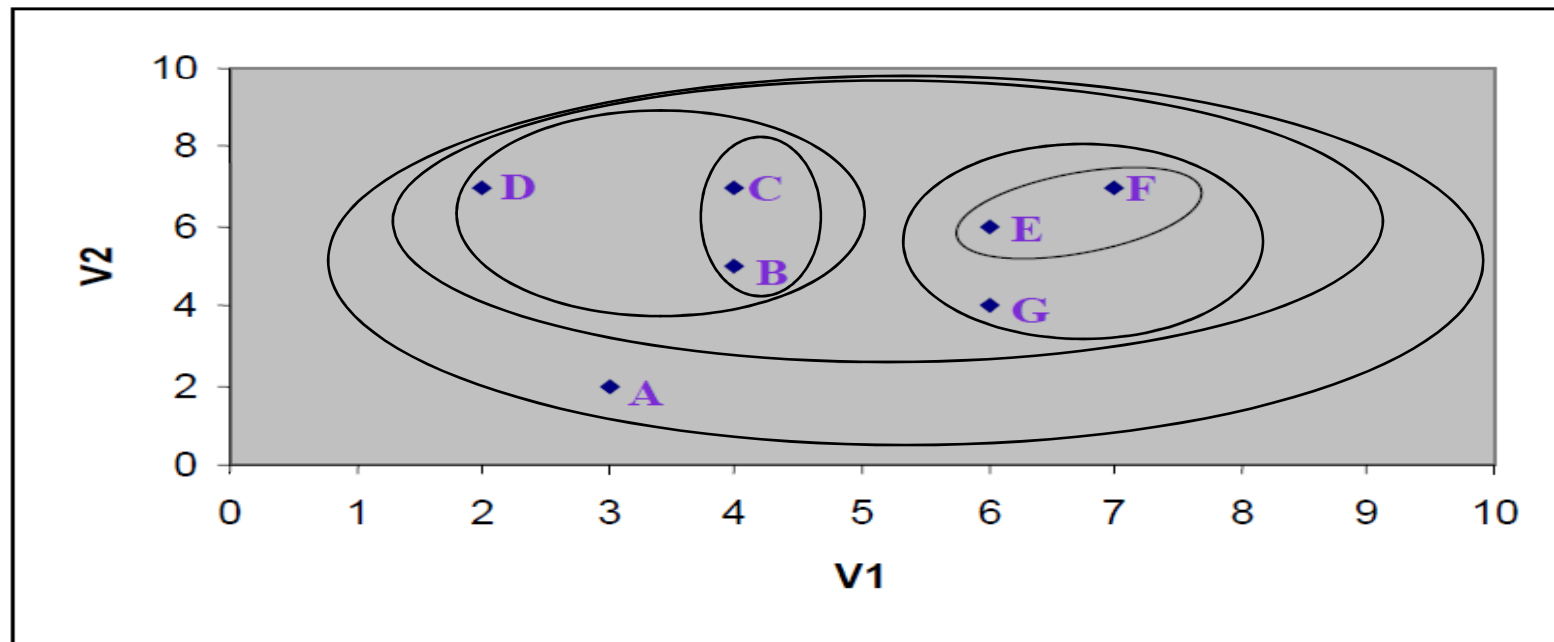
	A	BCD
A	0,0000	
BCD	4,5000	0,0000
EFG	4,5962	3,4821



Formação dos Grupos

Passo 6:

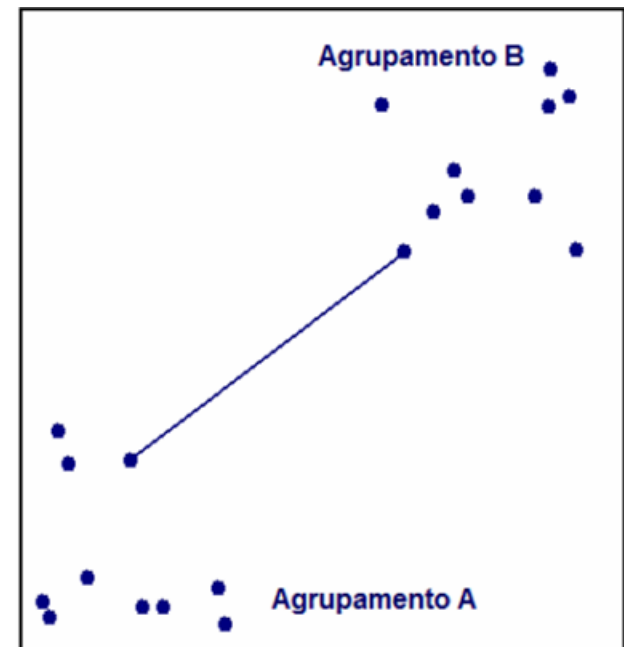
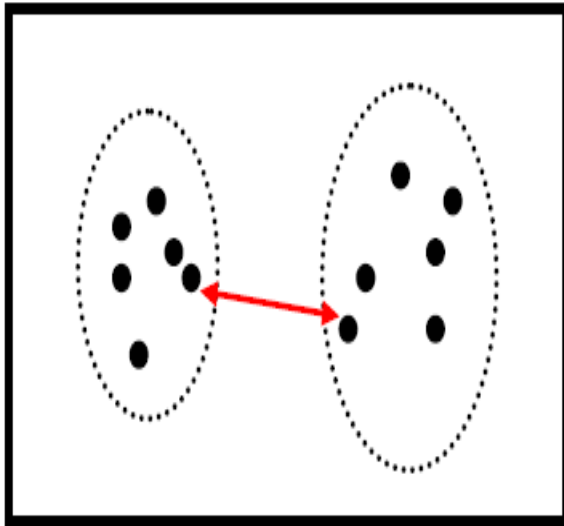
	A
A	0,0000
BCDEFG	4,2019



Ligação Simples

❑ Definição:

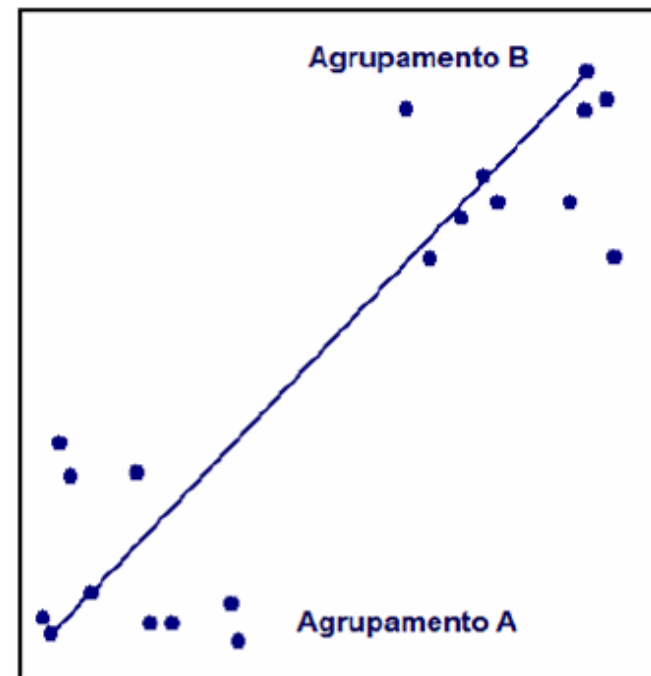
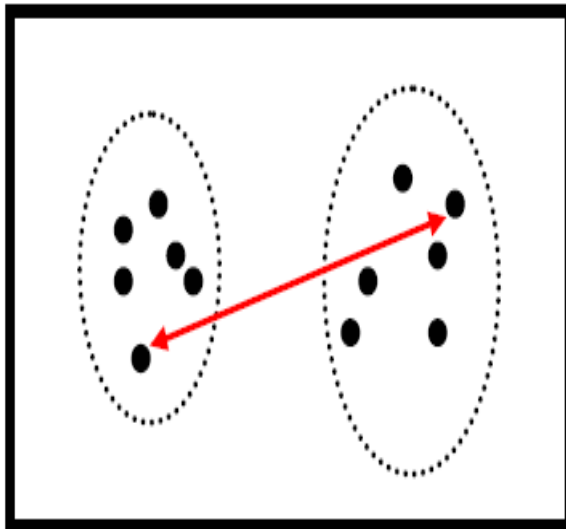
- ❖ Distância entre dois clusters é a distância entre os pontos mais próximos. Também chamado “agrupamento de vizinhos”.



Ligação Completa

❑ Definição:

- ❖ Distância entre dois clusters é a distância entre os pontos mais distantes.



Ligação Média

□ Ligação Média:

- ❖ Distância (similaridade) média de todos os objetos em um grupo para os demais em outro é usada como critério.
- ❖ São menos dependentes de valores extremos, como ocorre com a ligação simples ou completa.
- ❖ Tendem a combinar grupos com pequena variação interna.

Métodos Hierárquicos

❑ Características:

- ❖ Abordagem aglomerativa.
- ❖ Determinístico.

❑ Dificuldades:

- ❖ A estrutura é sempre uma árvore.
- ❖ Os objetos só podem ser agrupados baseando-se em decisões locais, que, uma vez tomadas, não podem ser re-avaliadas.

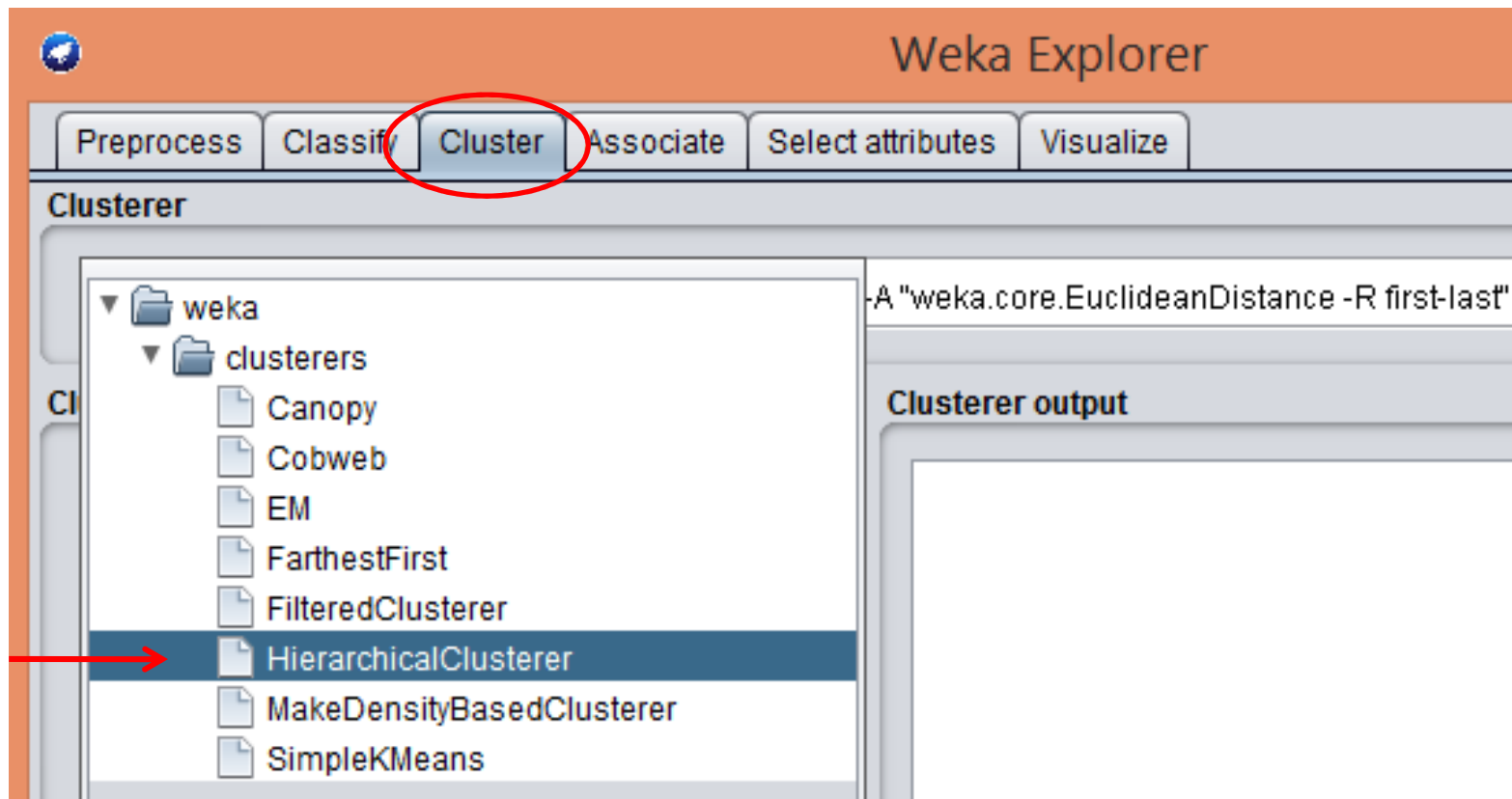
Hierárquico

- ❑ Abrir dataset PessoaNorm.csv;
- ❑ Utilizar o algoritmo Hierárquico Aglomerativo
 - ❖ numClusters = 3;
 - ❖ linkType = COMPLETE;
 - ❖ distanceFunction = EuclideanDistance.
- ❑ Salvar arquivo resultante:
 - ❖ PessoaNorm_Hiera3k_cLink.arff

❖ <https://www.dropbox.com/sh/b03djhgsziugl7h/AAARsUXaqAA7TYSsdk1-ZALKa?dl=0>

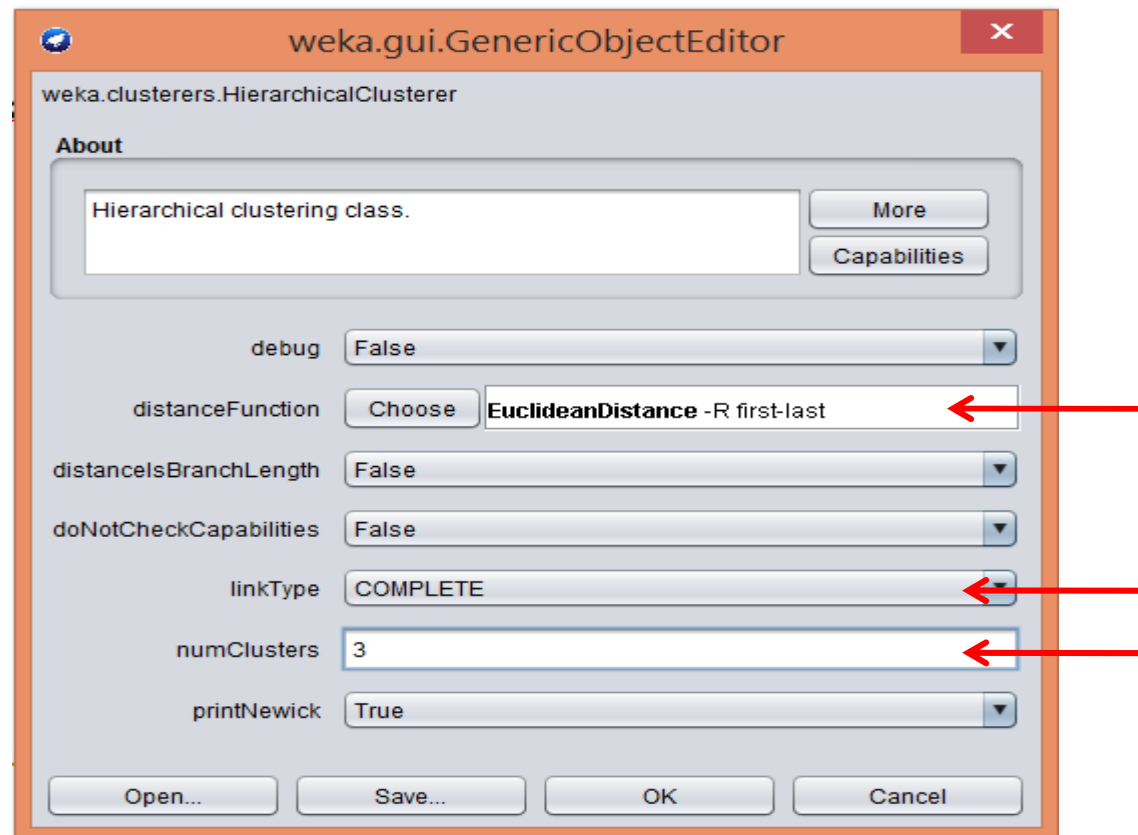
Hierárquico

- ❑ Utilizando **Hierarchical clustering** (WEKA):



Hierárquico

- ❑ Configurando o **Hierarchical clustering**:



Hierárquico

❑ Analisando os resultados....

Start Stop

Result list (right-click for options)

09:16:15 - HierarchicalClusterer

Time taken to build model (full training data) : 0 seconds

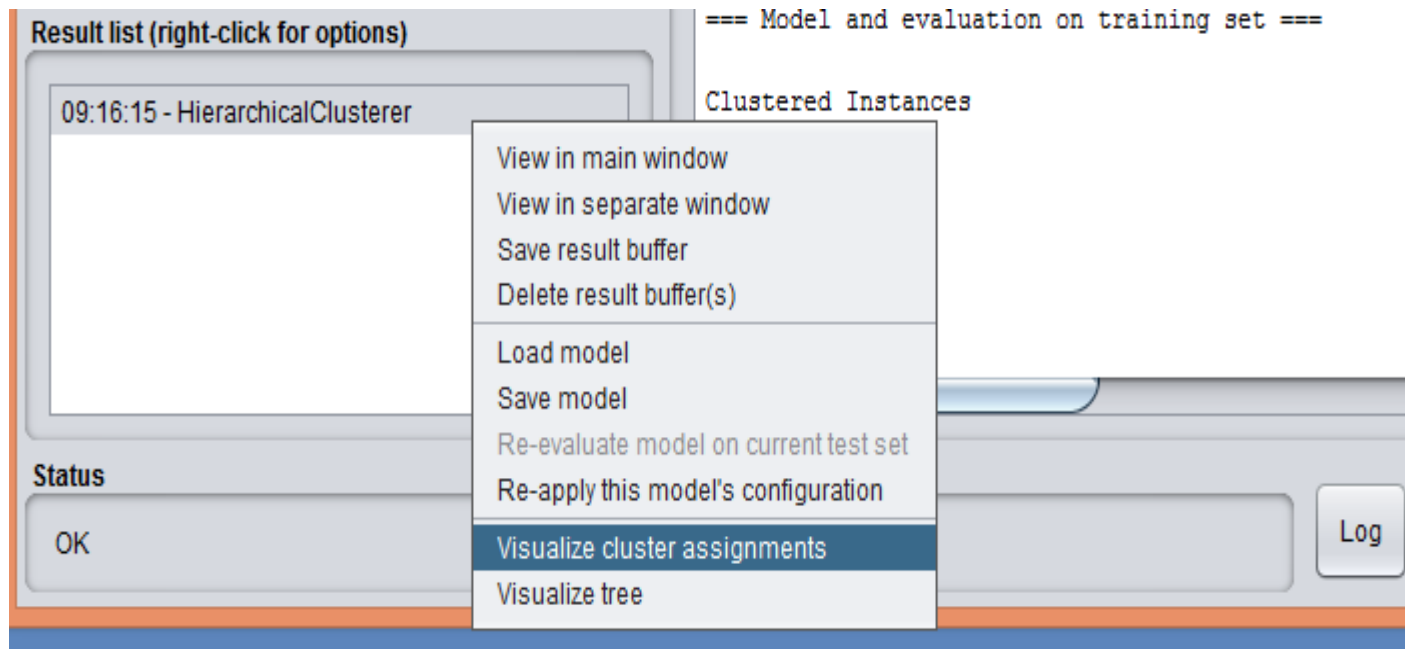
=== Model and evaluation on training set ===

Clustered Instances

0	4 (40%)
1	4 (40%)
2	2 (20%)

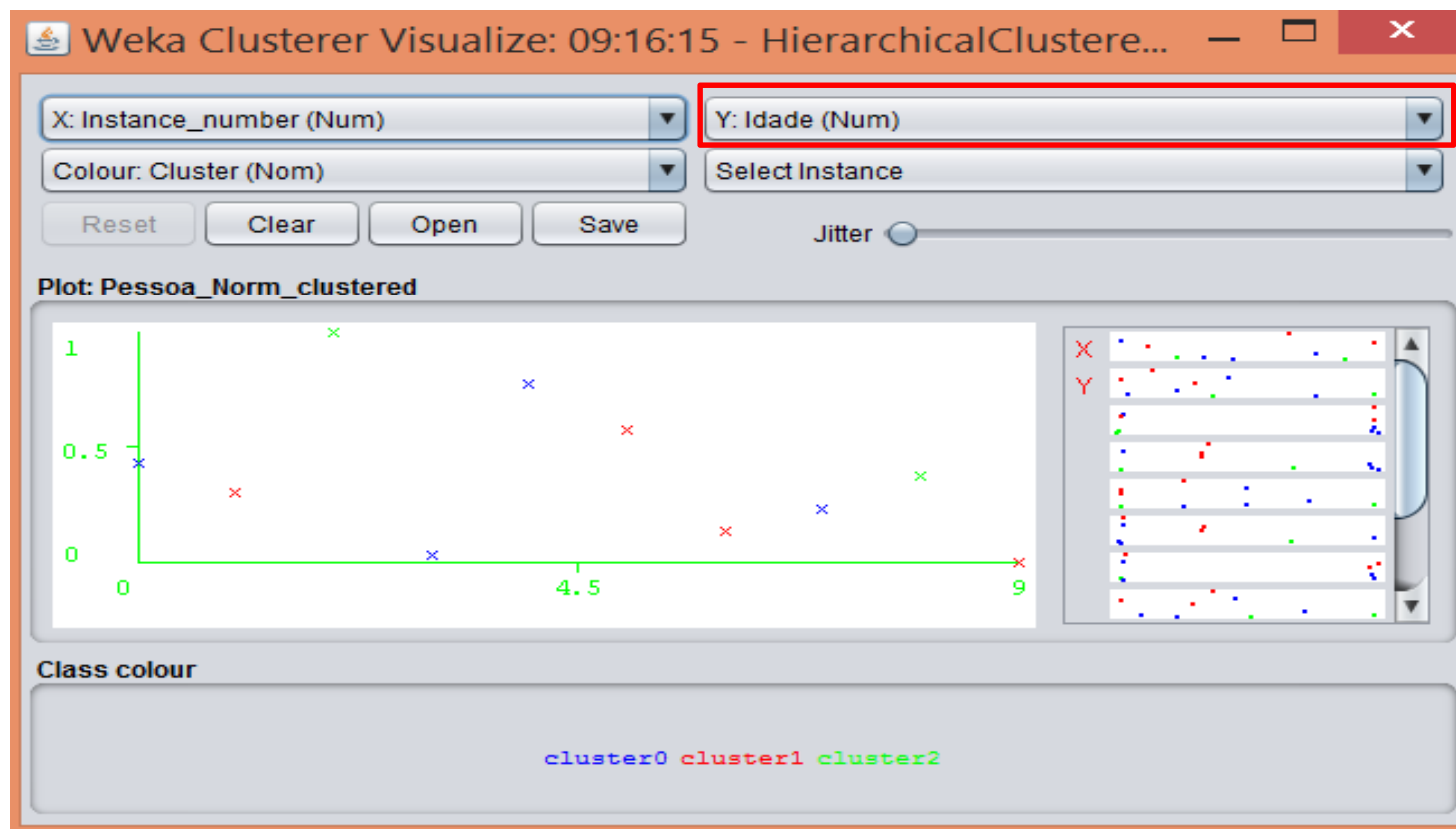
Hierárquico

- ❑ Visualizando os grupos gerados ...



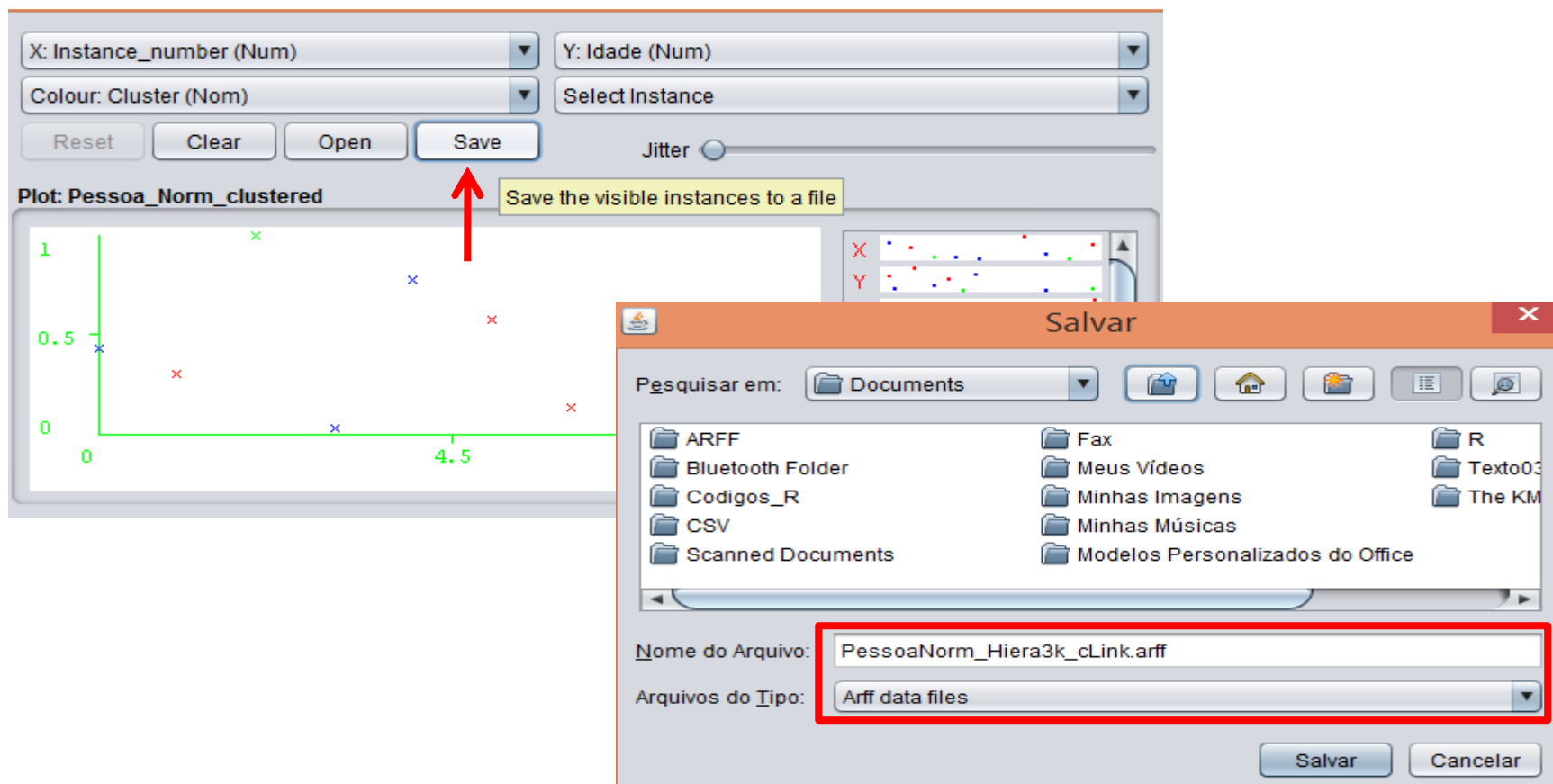
Hierárquico

- Visualizando os grupos gerados ...



Hierárquico

- ❑ Salvando partição resultante (arquivo com grupos) ...



Hierárquico

□ Partição resultante:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
1	0,43	Masc	Divorciado	0,50	Superior	0,46	Sim	Sim	cluster1
2	0,30	Femi	Solteiro	0,00	Medio	0,29	Sim	Nao	cluster2
3	1,00	Masc	Viuvo	1,00	Fundamental	1,00	Sim	Nao	cluster1
4	0,03	Femi	Casado	0,50	Superior	0,08	Nao	Sim	cluster3
5	0,77	Femi	Casado	0,25	Superior	0,23	Sim	Sim	cluster3
6	0,57	Masc	Casado	0,75	Medio	0,25	Sim	Nao	cluster1
7	0,13	Femi	Solteiro	0,25	Superior	0,37	Sim	Nao	cluster2
8	0,23	Femi	Casado	0,75	Pos_graduacao	0,73	Sim	Sim	cluster3
9	0,37	Masc	Divorciado	0,00	Superior	0,52	Nao	Nao	cluster1
10	0,00	Masc	Solteiro	0,00	Medio	0,00	Nao	Nao	cluster2

□ Partição resultante ordenada:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
1	0,43	Masc	Divorciado	0,50	Superior	0,46	Sim	Sim	cluster1
3	1,00	Masc	Viuvo	1,00	Fundamental	1,00	Sim	Nao	cluster1
6	0,57	Masc	Casado	0,75	Medio	0,25	Sim	Nao	cluster1
9	0,37	Masc	Divorciado	0,00	Superior	0,52	Nao	Nao	cluster1
2	0,30	Femi	Solteiro	0,00	Medio	0,29	Sim	Nao	cluster2
7	0,13	Femi	Solteiro	0,25	Superior	0,37	Sim	Nao	cluster2
10	0,00	Masc	Solteiro	0,00	Medio	0,00	Nao	Nao	cluster2
4	0,03	Femi	Casado	0,50	Superior	0,08	Nao	Sim	cluster3
5	0,77	Femi	Casado	0,25	Superior	0,23	Sim	Sim	cluster3
8	0,23	Femi	Casado	0,75	Pos_graduacao	0,73	Sim	Sim	cluster3

Hierárquico

- ❑ Abrir dataset **PessoaNorm.csv** novamente;
- ❑ Utilizar o algoritmo Hierárquico Aglomerativo
 - ❖ numClusters = 3;
 - ❖ linkType = SINGLE; e
 - ❖ linkType = AVERAGE;
 - ❖ distanceFunction = EuclideanDistance.
- ❑ Salvar os arquivos resultantes:
 - ❖ PessoaNorm_Hiera3k_sLink.arff; e
 - ❖ PessoaNorm_Hiera3k_avLink.arff.

Hierárquico Aglomerativo

❑ Analisando as partições

Clusterer

Choose **HierarchicalClusterer** -N 3 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

09:16:15 - HierarchicalClusterer	Clustered Instances
09:48:11 - HierarchicalClusterer	0 8 (80%)
	1 1 (10%)
	2 1 (10%)

Clusterer

Choose **HierarchicalClusterer** -N 3 -L AVERAGE -P -A "weka.core.EuclideanDistance -R first-last"

09:16:15 - HierarchicalClusterer	Clustered Instances
09:48:11 - HierarchicalClusterer	0 5 (50%)
09:50:55 - HierarchicalClusterer	1 4 (40%)
	2 1 (10%)

Dúvidas ...



K-Means ou K-Médias

História

- ❑ K-Means também chamado de **K-Médias**.
- ❑ É um algoritmo de Agrupamento (***Clustering***) que objetiva particionar n objetos em k grupos onde cada objeto pertence ao grupo mais próximo da média.
- ❑ Foi empregado primeiramente por James MacQueen em **1967**.

K-Means

- ❑ Difere do agrupamento hierárquico de várias maneiras.
Em particular:
 - ❖ Não há hierarquias, os dados são **particionados**.
 - ❖ Ou seja, a solução de seis grupos não é apenas a combinação de dois grupos a partir de uma solução com sete grupos, como no hierárquico.
- ❑ O resultado é apenas a **pertinência** final de cada **padrão** relacionado aos grupos.
- ❑ O número de grupos permitido (**k**) tem que ser **definido a priori**.

K-Means: Algoritmo

- ❑ Passo 1: os primeiros k centros dos grupos são escolhidos aleatoriamente.
- ❑ Passo2: cada objeto é atribuído ao grupo associado com o centro mais próximo.
- ❑ Passo3: compute um novo centro para cada grupo (média dos valores de todos os objetos - centróide).
- ❑ Passo4: repita passo2 (com os novos centros) e passo3 até que não haja mudança nos centros.

K-Means: exemplo (1/7)

- Passo 1: os primeiros k centros dos grupos são escolhidos aleatoriamente.

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
→ Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
→ Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Cliente_7	2,000	4,000	5,000	2,000	5,000

K-Means com Correlação de Pearson e $k=2$.

K-Means: exemplo (2/7)

- Passo2: cada objeto é atribuído ao grupo associado com o centro mais próximo.

	Ciente_1	Ciente_2	Ciente_3	Ciente_4	Ciente_5	Ciente_6	Ciente_7
Ciente_1	1,000						
Ciente_2	-0,147	1,000	0,000	0,516	-0,408	0,791	-0,516
Ciente_3	0,000	0,000	1,000				
Ciente_4	0,087	0,516	-0,824	1,000			
Ciente_5	0,963	-0,408	0,000	-0,060	1,000	-0,645	0,963
Ciente_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
Ciente_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

K-Means: Exemplo (3/7)

- ❑ Passo3: compute um novo centro para cada grupo (média dos valores de todos os objetos - centróide).

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
→ Centro_1	6,000	5,750	4,750	5,500	5,750

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
→ Centro_2	3,333	5,333	5,333	3,333	5,667

K-Means: Exemplo (4/7)

- Passo2: cada objeto é atribuído ao grupo associado com o centro mais próximo.

	Ciente_1	Ciente_2	Ciente_3	Ciente_4	Ciente_5	Ciente_6	Ciente_7	Centro_1	Centro_2
Ciente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1371	0,9723
Ciente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,93	-0,3498
Ciente_3	0	0	1	-0,8242	0	-0,3536	0,1648	-0,2599	0,068
Ciente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,737	-0,0698
Ciente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3797	0,9926
Ciente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,919	-0,6011
Ciente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4799	0,9723
Centro_1	-0,1371	0,93	-0,2599	0,737	-0,3797	0,919	-0,4799	1	-0,322
Centro_2	0,9723	-0,3498	0,068	-0,0698	0,9926	-0,6011	0,9723	-0,322	1

K-Means: Exemplo (5/7)

- Passo3: compute um novo centro para cada grupo (média dos valores de todos os objetos - centróide).

	X1	X2	X3	X4	X5
Cliente_2	9.000	9000	8000	9000	9000
Cliente_4	6.000	6000	3000	3000	4000
Cliente_6	4.000	3000	2000	3000	3000
→ Centro_1	6.333	6.000	4.333	5.000	5.333

	X1	X2	X3	X4	X5
Cliente_1	7.000	10000	9000	7000	10000
Cliente_3	5.000	5000	6000	7000	7000
Cliente_5	1.000	2000	2000	1000	2000
Cliente_7	2.000	4000	5000	2000	5000
→ Centro_2	3.750	5.250	5.500	4.250	6.000

K-Means: Exemplo (6/7)

- Passo2: cada objeto é atribuído ao grupo associado com o centro mais próximo.

	Ciente_1	Ciente_2	Ciente_3	Ciente_4	Ciente_5	Ciente_6	Ciente_7	Centro_1	Centro_2
Ciente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1106	0,9175
Ciente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,75	-0,3323
Ciente_3	0	0	1	-0,8242	0	-0,3536	-0,6281	0,3377	
Ciente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,9389	-0,2939
Ciente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3062	0,9372
Ciente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,8883	-0,6686
Ciente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4564	0,962
Centro_1	-0,1106	0,75	-0,6281	0,9389	-0,3062	0,8883	-0,4564	1	-0,4473
Centro_2	0,9175	-0,3323	0,3377	-0,2939	0,9372	-0,6686	0,962	-0,4473	1

K-Means: Exemplo (7/7)

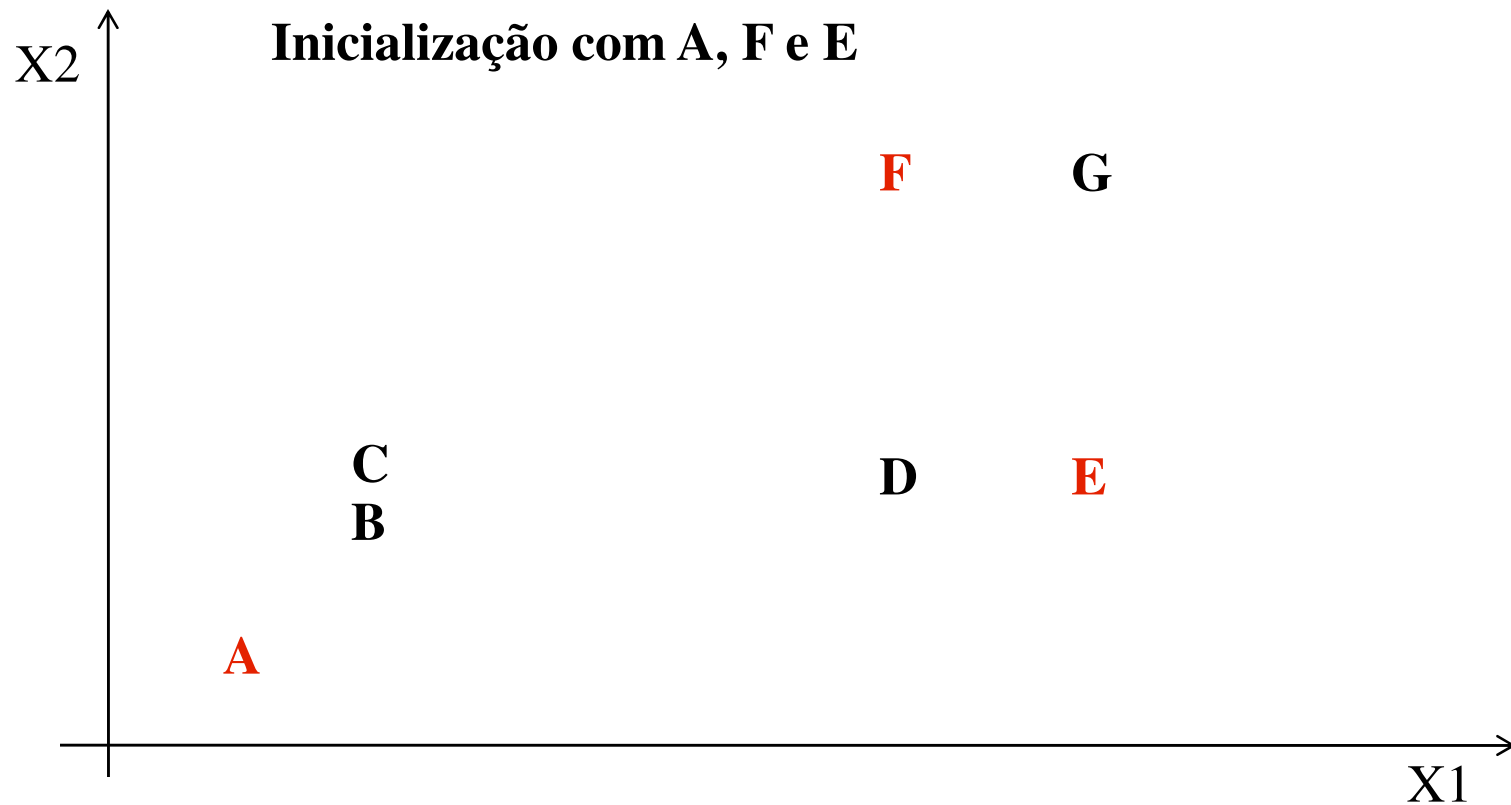
- Fim, pois não houve mudança nos centros.

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
→ Centro_1	6,333	6,000	4,333	5,000	5,333

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
→ Centro_2	3,750	5,250	5,500	4,250	6,000

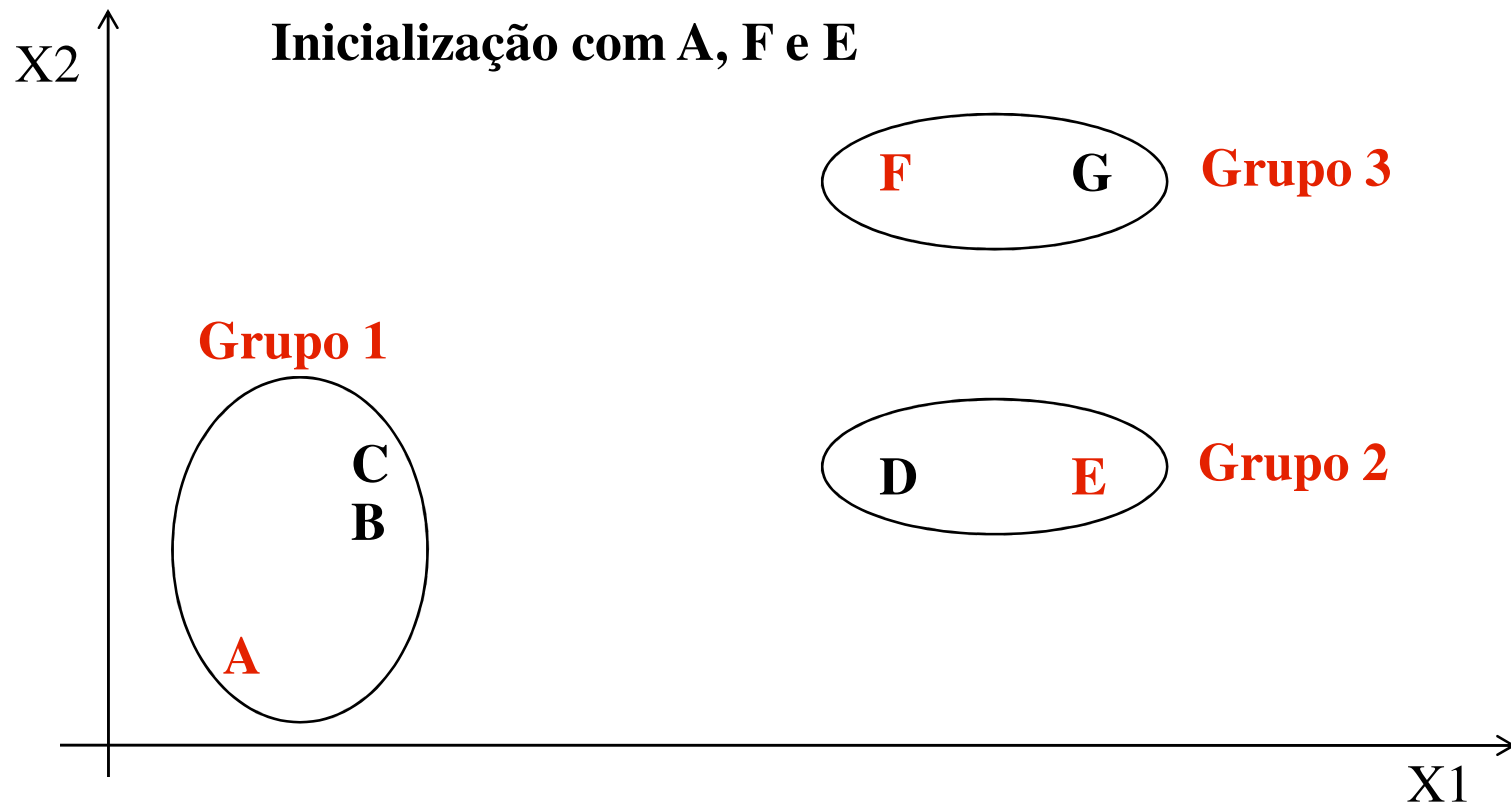
K-Means

❑ Sensibilidade à condição inicial:



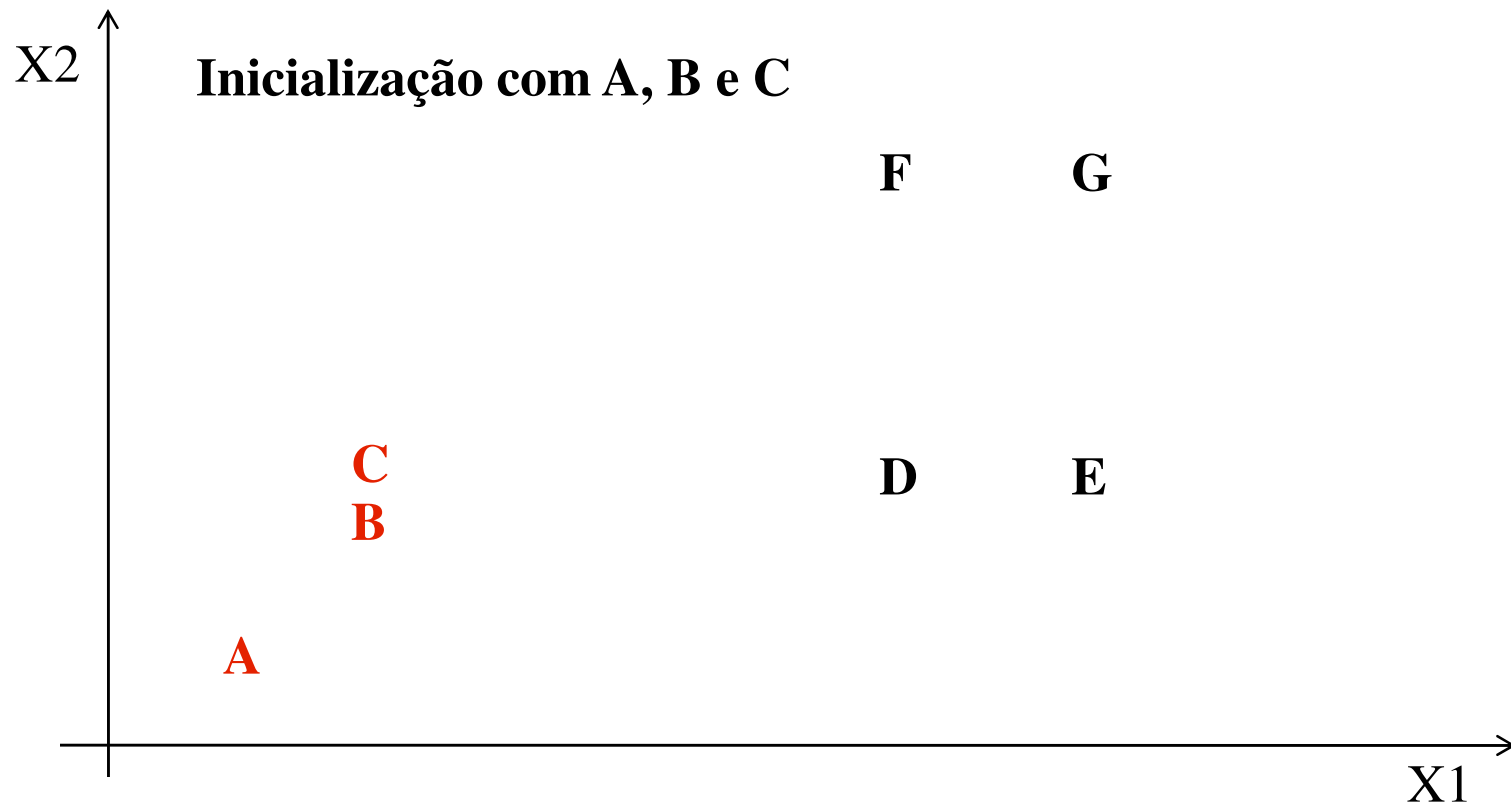
K-Means

❑ Sensibilidade à condição inicial:



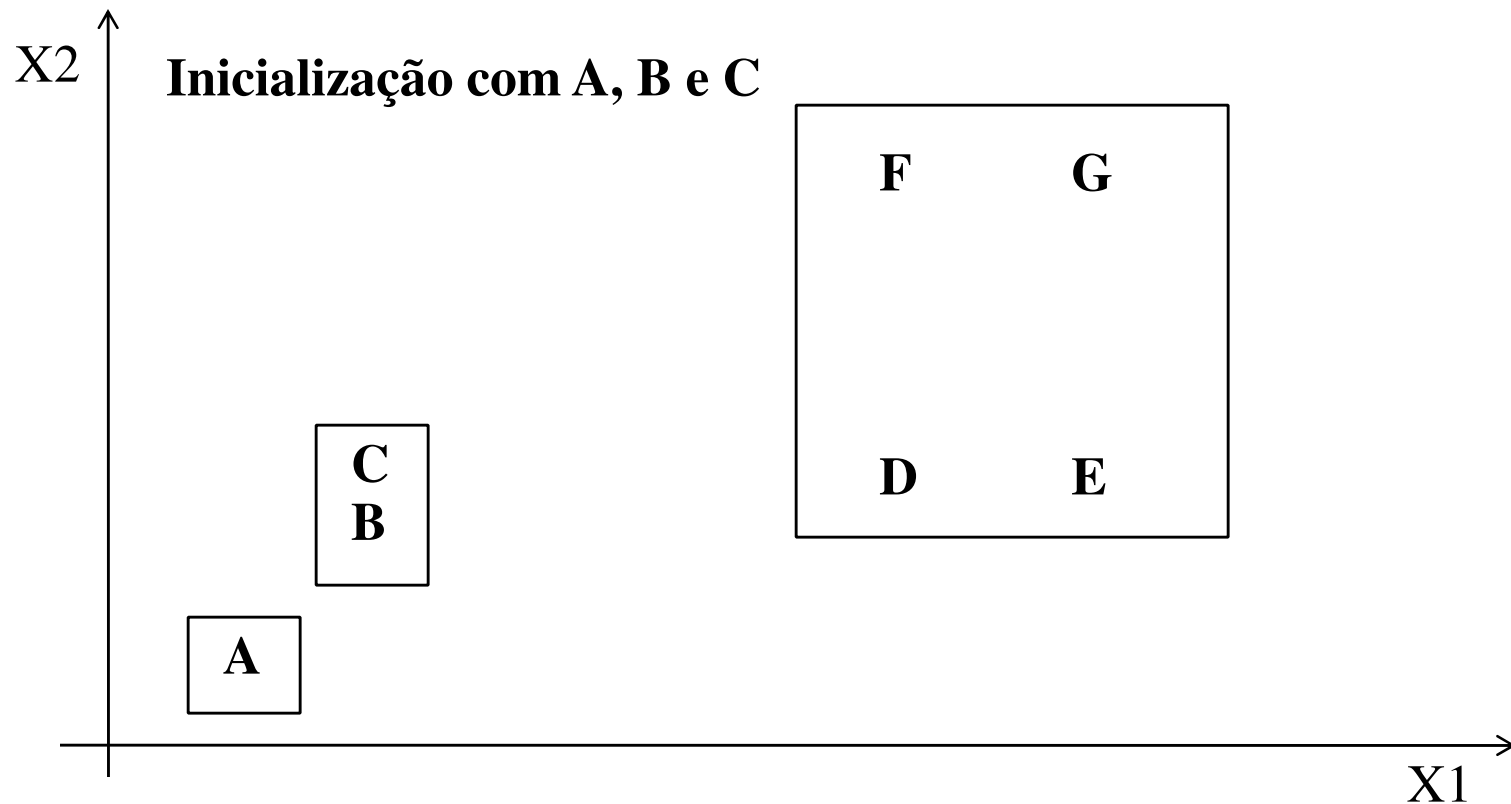
K-Means

❑ Sensibilidade à condição inicial:



K-Médias

- ❑ Sensibilidade à condição inicial:



K-Means

❑ Características:

- ❖ Partição.
- ❖ O número de grupos deve ser definido a priori.
- ❖ Não-determinístico: inicializações aleatórias dos centros.
- ❖ Grupos (clusters) esféricos.

❑ Dificuldades:

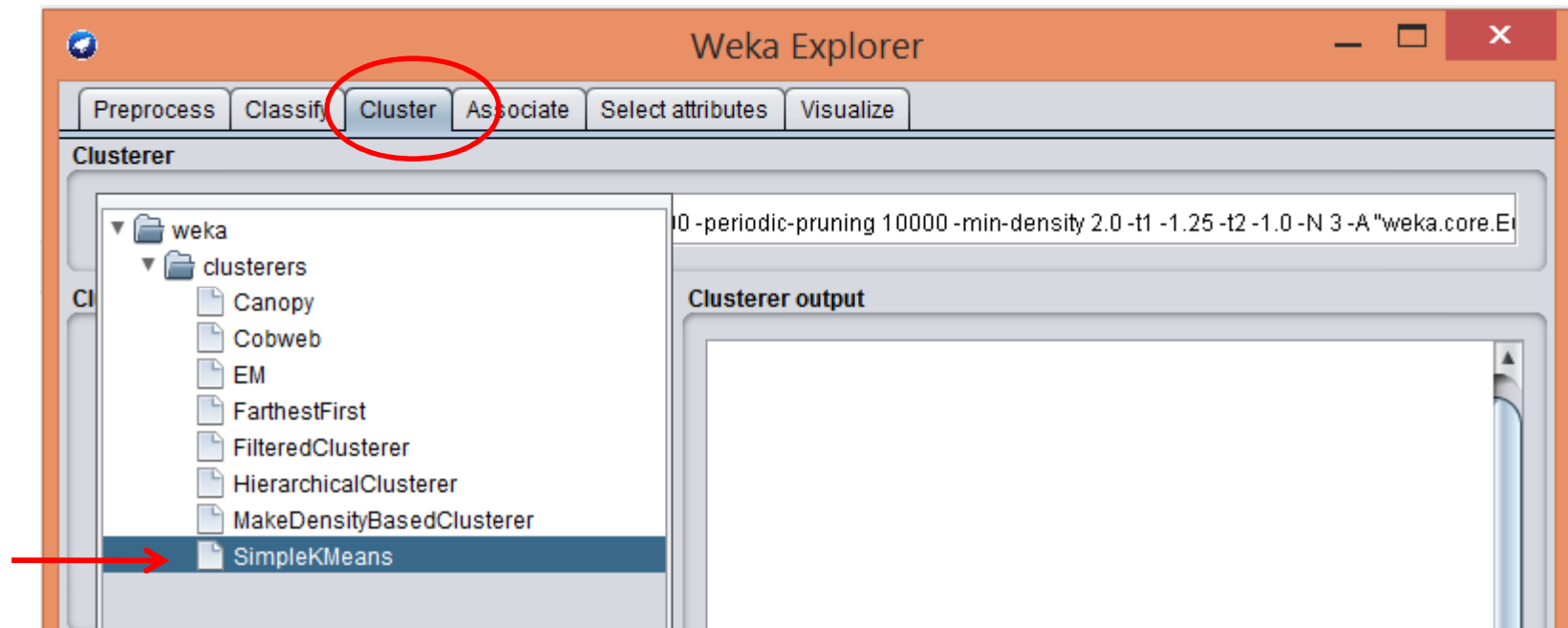
- ❖ Inicialização dos centros.

K-Means

- ❑ Abrir dataset PessoaNorm.csv;
- ❑ Utilizar o algoritmo Simple KMeans
 - ❖ numClusters = 3;
 - ❖ seed= 10;
 - ❖ distanceFunction = EuclideanDistance.
- ❑ Salvar arquivo resultante:
 - ❖ PessoaNorm_KMeans3k_s10.arff

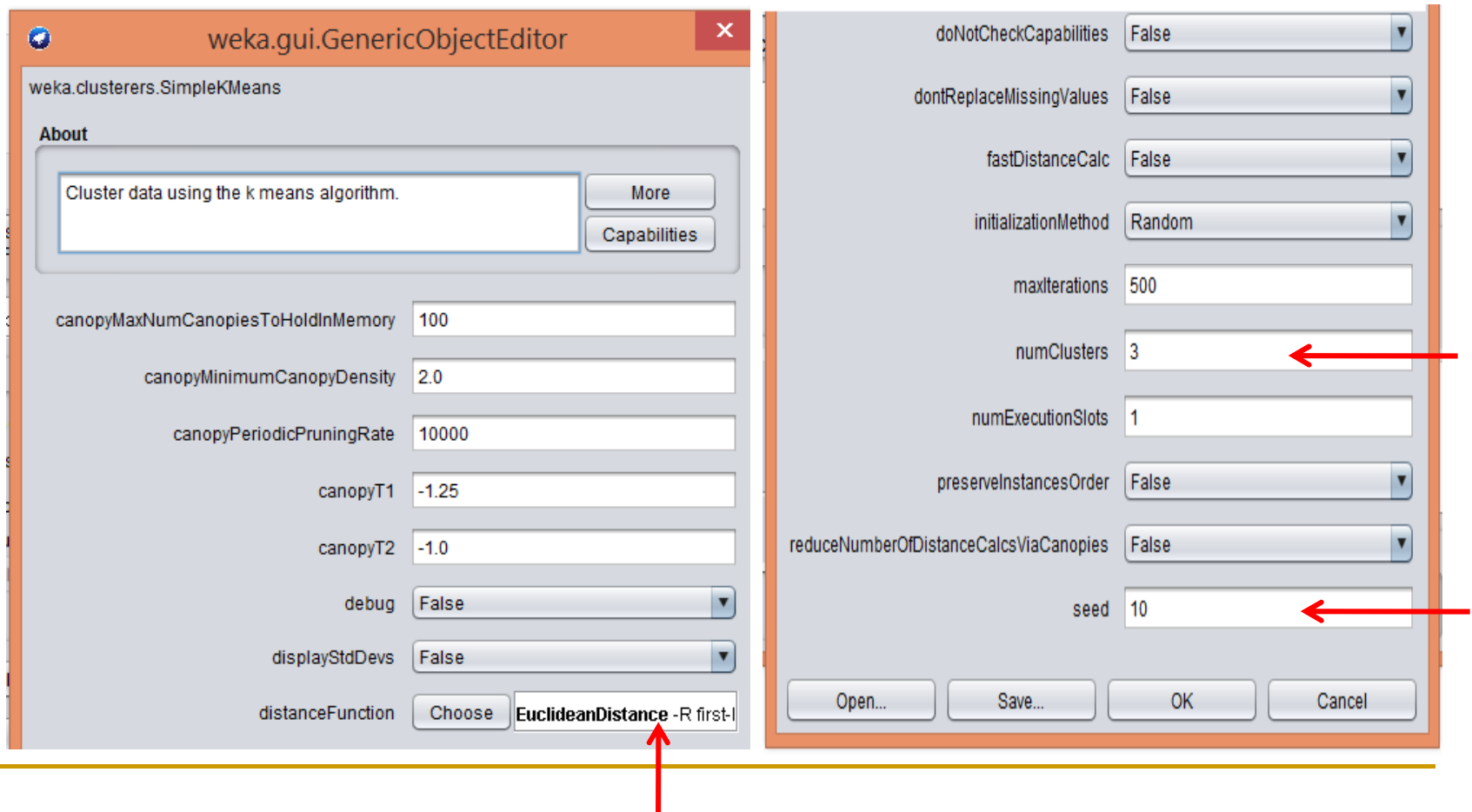
K-Means

- Utilizando **Simple KMeans**(WEKA):



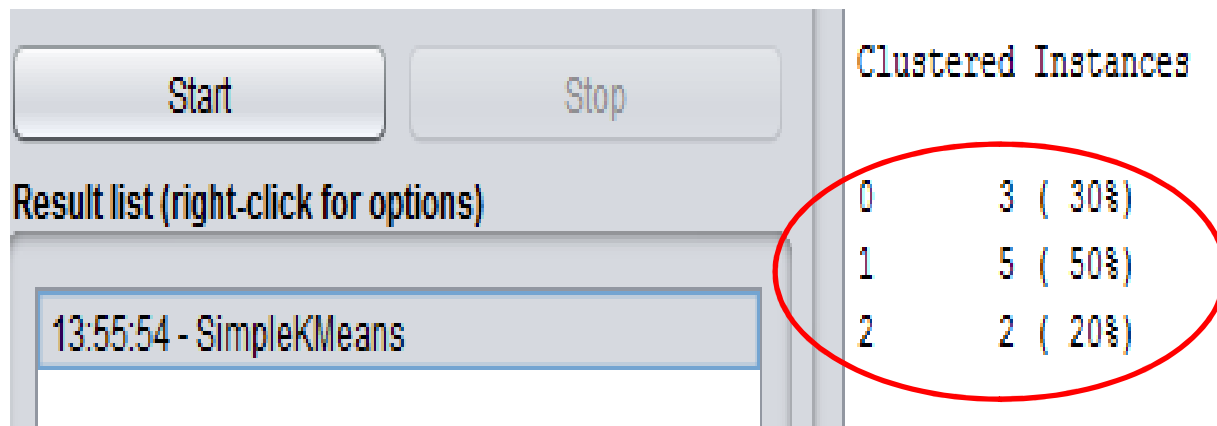
K-Means

❑ Configurando o Simple KMeans:



K-Means

- ❑ Analisando os resultados....



The screenshot shows a software interface for K-Means clustering. It features a 'Start' button and a 'Stop' button. Below them is a 'Result list (right-click for options)' section. A table titled 'Clustered Instances' displays the results of the clustering process. The table has three columns: Cluster ID, Number of Instances, and Percentage of Total Instances. The data is as follows:

Cluster ID	Number of Instances	Percentage of Total Instances
0	3	30%
1	5	50%
2	2	20%

A red oval highlights the 'Clustered Instances' table.

Validação de Agrupamentos

Validação de Agrupamentos

- ❑ Para a análise de agrupamentos, a questão é:
 - ❖ Como avaliar a “**qualidade**” dos grupos resultantes?
- ❑ Por que avaliá-los?
 - ❖ Comparar **diferentes algoritmos** de agrupamento.
 - ❖ Comparar **duas partições**.
 - ❖ Comparar **dois grupos** (clusters).

Medidas para Validação

□ Medidas numéricas aplicadas para avaliar aspectos da validação de agrupamentos.

❖ Índices Externos:

- Avaliam o agrupamento gerado baseado em uma estrutura pré-especificada (conjunto de dados).
 - Índice Rand ajustado (adjusted Rand) e índice de Jaccard.

❖ Índices Internos:

- Medem a qualidade de um agrupamento usando apenas os dados originais (instâncias ou matriz de similaridade).
 - Índice Davies-Bouldin, Silhuetas, Índice Dunn,

Índices Internos

- ❑ Não há um **rotulo** (classe) para os dados.
- ❑ Medem a qualidade de um agrupamento usando apenas os dados originais.
 - ❖ Utilizam alguma medida de similaridade (**compacticidade**).
- ❑ Principais medidas:
 - ❖ Índice Davies-Bouldin,
 - ❖ Silhuetas,
 - ❖ Índice Dunn, ...

Índice Davies-Bouldin (DB)

- Dada uma partição $\{C_1, C_2, \dots, C_k\}$, definimos a **similaridade relativa** entre dois grupos, C_i e C_j , como:

$$RS_{i,j} = \frac{E_i + E_j}{d(m_i, m_j)} \qquad E_i = \frac{1}{C_i} \sum_{x \in C_i} (x - z_i)^2$$

Onde:

- $d(m_i, m_j)$ é a distância entre as médias do grupo i e grupo j ;
- E_i é a distância quadrada média dos pontos no i -ésimo grupo para o centroide (média) desse grupo.

Índice Davies-Bouldin (DB)

- Com $RS_{i,j}$, podemos calcular a similaridade relativa máxima entre o grupo i e cada um dos outros (MRS_i):

$$MRS_i = \max_{i,j} \{RS_{i,j}\}$$

- O índice **Davies-Bouldin** (DB) para a partição $\{C_1, C_2 \dots C_k\}$ é a média de MRS_i ($i = 1, 2 \dots k$):

$$DB(k) = \frac{1}{k} \sum_{i=1}^k MRS_i$$

Índice Davies-Bouldin (DB)

- Dada uma partição $\{C_1, C_2, \dots, C_k\}$, definimos a **similaridade relativa** entre dois grupos, C_i e C_j , como:

$$RS_{i,j} = \frac{E_i + E_j}{d(m_i, m_j)} \qquad E_i = \frac{1}{C_i} \sum_{x \in C_i} (x - z_i)^2$$

Onde:

- $d(m_i, m_j)$ é a distância entre as médias do grupo i e grupo j ;
- E_i é a distância quadrada média dos pontos no i -ésimo grupo para o centroide (média) desse grupo.

Índice Davies-Bouldin (DB)

□ Partição resultante ordenada:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
1	0,43	Masc	Divorciado	0,50	Superior	0,46	Sim	Sim	cluster1
3	1,00	Masc	Viuvo	1,00	Fundamental	1,00	Sim	Nao	cluster1
6	0,57	Masc	Casado	0,75	Medio	0,25	Sim	Nao	cluster1
9	0,37	Masc	Divorciado	0,00	Superior	0,52	Nao	Nao	cluster1
2	0,30	Femi	Solteiro	0,00	Medio	0,29	Sim	Nao	cluster2
7	0,13	Femi	Solteiro	0,25	Superior	0,37	Sim	Nao	cluster2
10	0,00	Masc	Solteiro	0,00	Medio	0,00	Nao	Nao	cluster2
4	0,03	Femi	Casado	0,50	Superior	0,08	Nao	Sim	cluster3
5	0,77	Femi	Casado	0,25	Superior	0,23	Sim	Sim	cluster3
8	0,23	Femi	Casado	0,75	Pos_graduacao	0,73	Sim	Sim	cluster3

Índice Davies-Bouldin (DB)

□ Entendendo o índice:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
1	0,43	Masc	Divorciado	0,50	Superior	0,46	Sim	Sim	cluster1
3	1,00	Masc	Viuvo	1,00	Fundamental	1,00	Sim	Nao	cluster1
6	0,57	Masc	Casado	0,75	Medio	0,25	Sim	Nao	cluster1
9	0,37	Masc	Divorciado	0,00	Superior	0,52	Nao	Nao	cluster1
#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
2	0,30	Femi	Solteiro	0,00	Medio	0,29	Sim	Nao	cluster2
7	0,13	Femi	Solteiro	0,25	Superior	0,37	Sim	Nao	cluster2
10	0,00	Masc	Solteiro	0,00	Medio	0,00	Nao	Nao	cluster2
#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
4	0,03	Femi	Casado	0,50	Superior	0,08	Nao	Sim	cluster3
5	0,77	Femi	Casado	0,25	Superior	0,23	Sim	Sim	cluster3
8	0,23	Femi	Casado	0,75	Pos_graduacao	0,73	Sim	Sim	cluster3

Índice Davies-Bouldin (DB)

- ❑ Calculando o **centróide** do grupo 1:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
1	0,43	Masc	Divorciado	0,50	Superior	0,46	Sim	Sim	cluster1
3	1,00	Masc	Viuvo	1,00	Fundamental	1,00	Sim	Nao	cluster1
6	0,57	Masc	Casado	0,75	Medio	0,25	Sim	Nao	cluster1
9	0,37	Masc	Divorciado	0,00	Superior	0,52	Nao	Nao	cluster1
centróide =>		0,59	Masc	Divorciado	0,56	Superior	0,56	Sim	Nao

Índice Davies-Bouldin (DB)

- Calculando o **centróide** do grupo 2:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
2	0,30	Femi	Solteiro	0,00	Medio	0,29	Sim	Nao	cluster2
7	0,13	Femi	Solteiro	0,25	Superior	0,37	Sim	Nao	cluster2
10	0,00	Masc	Solteiro	0,00	Medio	0,00	Nao	Nao	cluster2

centróide =>	0,14	Femi	Solteiro	0,08	Medio	0,22	Sim	Nao	
------------------------	------	------	----------	------	-------	------	-----	-----	--

Índice Davies-Bouldin (DB)

- Calculando o **centróide** do grupo 3:

#	Idade	Gênero	EC	Filhos	Escola	Renda	Cartão	Imóvel	Grupo
4	0,03	Femi	Casado	0,50	Superior	0,08	Nao	Sim	cluster3
5	0,77	Femi	Casado	0,25	Superior	0,23	Sim	Sim	cluster3
8	0,23	Femi	Casado	0,75	Pos_graduacao	0,73	Sim	Sim	cluster3
centróide =>		0,34	Femi	Casado	0,50	Superior	0,35	Sim	Sim

Índice Davies-Bouldin (DB)

Calculando os E_i :
$$E_i = \frac{1}{C_i} \sum_{x \in C_i} (x - z_i)^2$$

	Atributos									
	1	2	3	4	5	6	7	8		
centróide =>	0,59	Masc	Divorciado	0,56	Superior	0,56	Sim	Nao	Dist.QuaM	E1
	0,16	0,00	0,00	0,06	0,00	0,10	0,00	1,00	1,7490	
	0,41	0,00	1,00	0,44	1,00	0,44	0,00	0,00	10,8077	
	0,02	0,00	1,00	0,19	1,00	0,31	0,00	0,00	6,3378	
	0,22	0,00	0,00	0,56	0,00	0,04	1,00	0,00	3,3215	5,5540
centróide =>	0,14	Femi	Solteiro	0,08	Medio	0,22	Sim	Nao	Dist.QuaM	E2
	0,16	0,00	0,00	0,08	0,00	0,07	0,00	0,00	0,0961	
	0,01	0,00	0,00	0,17	1,00	0,15	0,00	0,00	1,7689	
	0,14	1,00	0,00	0,08	0,00	0,22	1,00	0,00	5,9862	2,6171
centróide =>	0,34	Femi	Casado	0,50	Superior	0,35	Sim	Sim	Dist.QuaM	E3
	0,31	0,00	0,00	0,00	0,00	0,27	1,00	0,00	2,4964	
	0,43	0,00	0,00	0,25	0,00	0,12	0,00	0,00	0,6294	
	0,11	0,00	0,00	0,25	1,00	0,38	0,00	0,00	3,0508	2,0589

Índice Davies-Bouldin (DB)

Calculando os RS_s : $RS_{i,j} = \frac{E_i + E_j}{d(m_i, m_j)}$

1	0,59	Masc	Divorciado	0,56	Superior	0,56	Sim	Nao
2	0,14	Femi	Solteiro	0,08	Medio	0,22	Sim	Nao
3	0,34	Femi	Casado	0,50	Superior	0,35	Sim	Sim



										Soma	RS
[1,2]	0,45	1,00	1,00	0,48	1,00	0,34	0,00	0,00		4,2658	1,9155
[1,3]	0,25	1,00	1,00	0,06	0,00	0,21	0,00	1,00		3,5225	2,1612
[2-3]	0,20	0,00	1,00	0,42	1,00	0,13	0,00	1,00		3,7433	1,2491



$[1-2] = (5,5540 + 2,6171) / 4,2658 = 1,9155$ $[1-3] = (5,5540 + 2,0589) / 3,5225 = 2,1612$ $[2-3] = (2,6171 + 2,0589) / 3,7433 = 1,2491$

Índice Davies-Bouldin (DB)

□ Calculando os MRS_s : $MRS_i = \max_{i,j} \{RS_{i,j}\}$

$$\begin{aligned} [1-2] &= (5,5540 + 2,6171) / 4,2658 = 1,9155 \\ [1-3] &= (5,5540 + 2,0589) / 3,5225 = 2,1612 \\ [2-3] &= (2,6171 + 2,0589) / 3,7433 = 1,2491 \end{aligned}$$



$$\begin{aligned} \text{Máximo entre } [1-2] \text{ e } [1-3] &= [1,9155 : 2,1612] \Rightarrow 2,1612 \\ \text{Máximo entre } [2-1] \text{ e } [2-3] &= [1,9155 : 1,2491] \Rightarrow 1,9155 \\ \text{Máximo entre } [3-1] \text{ e } [3-2] &= [2,1612 : 1,2491] \Rightarrow 2,1612 \end{aligned}$$

Índice Davies-Bouldin (DB)

□ Calculando o **DB**:

$$DB(k) = \frac{1}{k} \sum_{i=1}^k MRS_i$$

$$MRS_1 = 2,1612$$

$$MRS_2 = 1,9155$$

$$MRS_3 = 2,1612$$

$$k = 3$$

$$\text{Média} = (2,1612 + 1,9155 + 2,1612) / 3$$

$$DB = 2,0793$$