> Automated Variable Weighting in k-Means Type Clustering

(Huang, J.Z.; Ng, M.K.; Hongqiang Rong; Zichen Li. ; 2005)

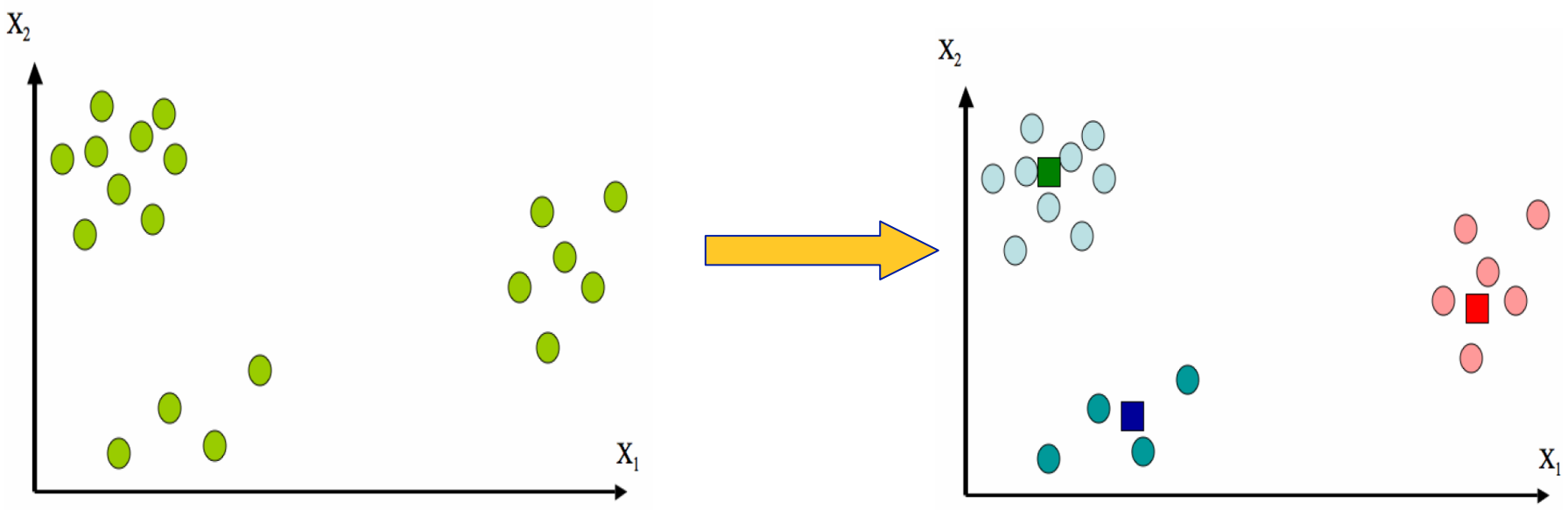
Présentation

Franck.Dernoncourt@gmail.com

28 Septembre 2010

Sommaire

La classification (clustering)



1. La classification

Exemples d'applications de la classification

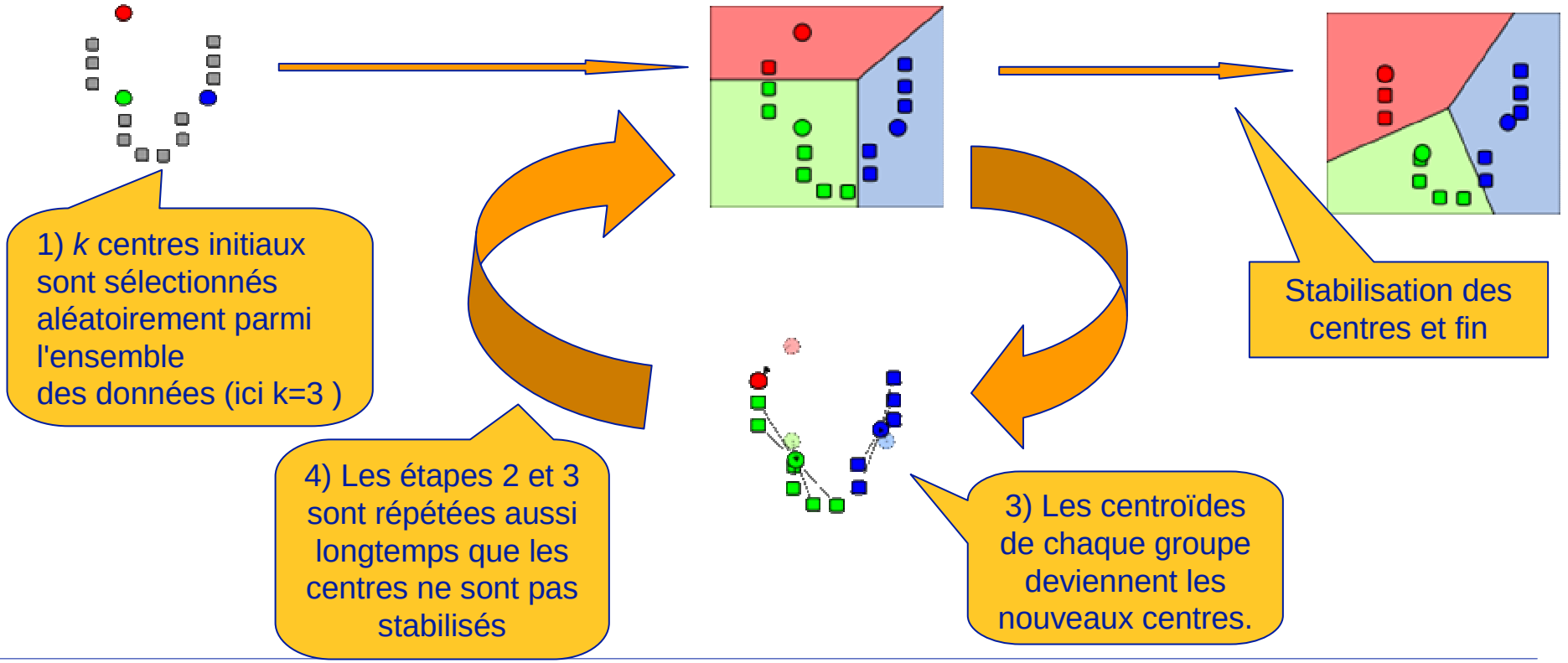
- *Marketing* : trouver des groupes de clients similaires
- *Biologie* : classer des plantes selon leurs caractéristiques
- *Algorithmes évolutionnistes* : améliorer la diversité des croisements.
- …

Sommaire

2. La méthode des K-moyennes

K-moyennes (K-means)

Une méthode non supervisée
MacQueen, 1967



2. La méthode des K-moyennes

Les limites de l'algorithme des K-moyennes :

- Nécessité de l'existence d'une distance
- Choix du nombre de classes
- Influence du choix des centres initiaux sur le résultat
- Sensible au bruit

2. La méthode des K-moyennes

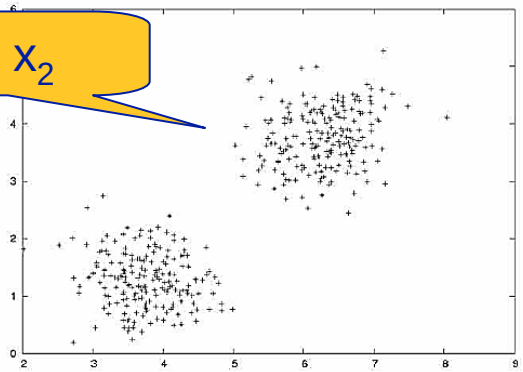
Les limites de l'algorithme des K-moyennes :

- Nécessité de l'existence d'une distance
- Choix du nombre de classes
- Influence du choix des centres initiaux sur le résultat
- **Sensible au bruit**

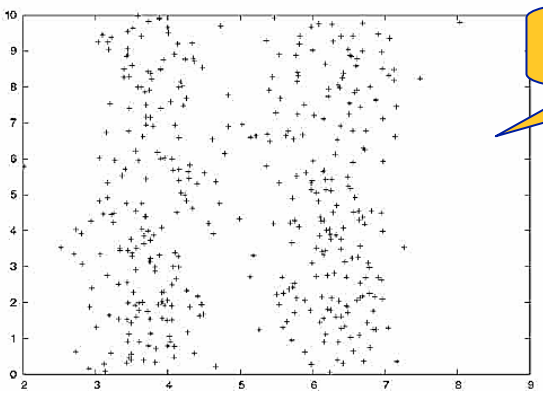
2. La méthode des K-moyennes

Exemple en 3D : x_1 et x_2 sont ok pour la classification, x_3 est du bruit

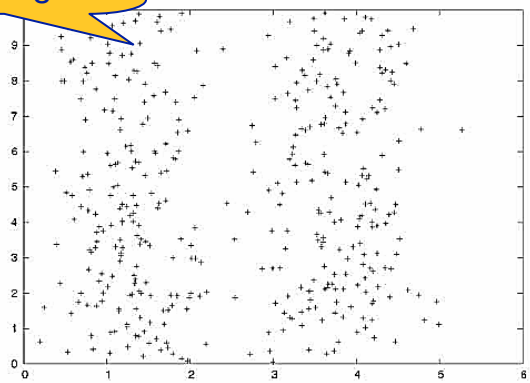
x_1 et x_2



x_1 et x_3



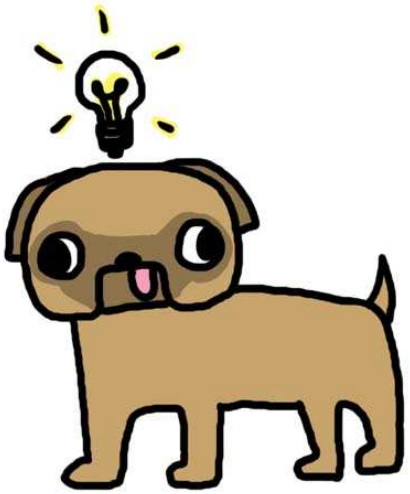
x_2 et x_3



Classification obtenue

Sommaire

K-moyennes avec pondération des poids dynamique
(*Automated Variable Weighting in k-Means Type Clustering – 2005)*

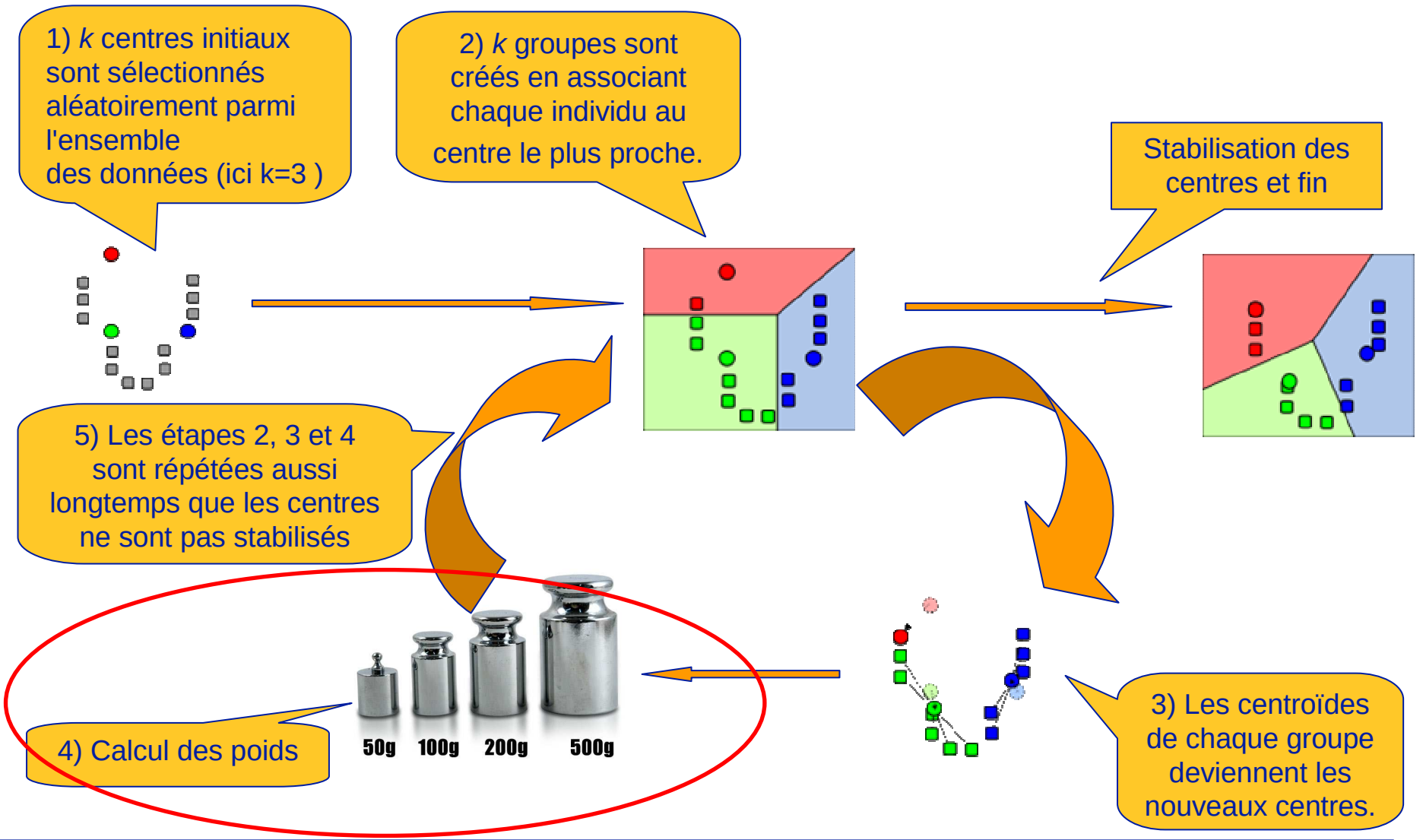


Idée **: Pondérer chaque variable afin de pouvoir donner un poids plus faible aux variables affectées par un bruit important.**

Existant : Modha et Spangler y ont déjà pensé… mais les poids qu'ils utilisaient étaient calculés au tout début de l'algorithme.

Ici, les poids vont être calculés dynamiquement, à chaque itérations de l'algorithme des K-moyennes.

3. K-moyennes avec pondération des poids dynamique



Calcul des poids - Théorème

$$\hat{w}_j = \begin{cases} 0 & \text{si } D_j = 0 \\ \dfrac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & \text{si } D_j \neq 0 \end{cases}$$

avec $$D_j = \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{u}_{i,l} d(x_{i,j}, z_{l,j})$$

Où $u_{i,l}$ signifie que l'objet i est affecté à la classe l

$d(x_{i,j}, z_{l,j})$ est la distance entre les objets x et z

h est le nombre de variables D_j telles que $D_j \neq 0$

$z_{l,j}$ est la valeur de la variable j du centroïde du cluster l

Calcul des poids - Théorème

Idée : Donner un poids faible pour les variables dont la valeur de chacun des individus est en moyenne éloigner des centroïdes

$$\hat{w}_j = \begin{cases} 0 & \text{si } D_j = 0 \\ \dfrac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & \text{si } D_j \neq 0 \end{cases}$$

avec

$$D_j = \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{u}_{i,l} d(x_{i,j}, z_{l,j})$$

Où $u_{i,l}$ signifie que l'objet i est affecté à la classe l

$d(x_{i,j}, z_{l,j})$ est la distance entre les objets x et z

h est le nombre de variables D_j telles que $D_j \neq 0$

$z_{l,j}$ est la valeur de la variable j du centroïde du cluster l

Calcul des poids

Fonction à minimiser :

$$P(\hat{U}, \hat{Z}, W) = \sum_{j=1}^{m} w_j^{\beta} \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{u}_{i,l} d(x_{i,j}, z_{l,j})$$

$$= \sum_{j=1}^{m} w_j^{\beta} D_j,$$

Contrainte :

$$\sum_{j=1}^{m} w_j = 1, \quad 0 \leq w_j \leq 1$$

→ Multiplicateurs de Lagrange !

Calcul des poids

Le Lagrangien :

$$\Psi(W,\alpha) = \sum_{j=1}^{h} w_j^{\beta} D_j + \alpha\left(\sum_{j=1}^{h} w_j - 1\right)$$

On dérive :

$$\frac{\partial \Psi(\hat{W},\hat{\alpha})}{\partial \hat{w}_j} = \beta \hat{w}_j^{\beta-1} D_j + \hat{\alpha} = 0 \quad \text{for } 1 \le j \le h,$$

$$\frac{\partial \Psi(\hat{W},\hat{\alpha})}{\partial \hat{\alpha}} = \sum_{j}^{h} \hat{w}_j - 1 = 0.$$

Calcul des poids

On voit : $$\hat{w}_j = \left(\frac{-\hat{\alpha}}{\beta D_j} \right)^{\frac{1}{\beta-1}} \quad \text{for } 1 \leq j \leq h$$

De plus : $$\sum_{t=1}^{h} \left(\frac{-\hat{\alpha}}{\beta D_t} \right)^{\frac{1}{\beta-1}} = 1 \longrightarrow (-\hat{\alpha})^{\frac{-1}{\beta-1}} = 1 / \left[\sum_{t=1}^{h} \left(\frac{1}{\beta D_t} \right)^{\frac{1}{\beta-1}} \right]$$

D'où : $$\hat{w}_j = \frac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t} \right]^{\frac{1}{\beta-1}}}$$

CQFD !

Calcul des poids - Théorème

Idée : Donner un poids faible pour les variables dont la valeur de chacun des individus est en moyenne éloigner des centroïdes

$$\hat{w}_j = \begin{cases} 0 & \text{si } D_j = 0 \\ \dfrac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & \text{si } D_j \neq 0 \end{cases}$$

avec

$$D_j = \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{u}_{i,l} d(x_{i,j}, z_{l,j})$$

Où $u_{i,l}$ signifie que l'objet i est affecté à la classe l

$d(x_{i,j}, z_{l,j})$ est la distance entre les objets x et z

h est le nombre de variables D_j telles que $D_j \neq 0$

$z_{l,j}$ est la valeur de la variable j du centroïde du cluster l

Sommaire

4. Expériences

Expérience 1 : Avec un jeu de données synthétique

5 Variables, 300 individus :

X_1, X_2, X_3 : Données formant 3 classes nettes

X_4, X_5 : Bruit

Comme nous connaissons les 3 classes, nous allons comparer les résultats obtenus par l'algorithme des K-moyennes standard avec l'algorithme des K-moyennes avec pondération des poids dynamique.

Pour effectuer cette comparaison, nous utiliserons l'indice de Rand et le clustering accuracy qui permettent d'évaluer la performance d'une classification par comparaison avec la classification voulue.

4. Expériences

Indice de Rand :

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

- $a = |S^*|$, où $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i, o_j \in Y_l\}$
- $b = |S^*|$, où $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $c = |S^*|$, où $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $d = |S^*|$, où $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i, o_j \in Y_l\}$

avec $1 \leq i, j \leq n, i \neq j, 1 \leq k, k_1, k_2 \leq r, k_1 \neq k_2, 1 \leq l, l_1, l_2 \leq s, l_1 \neq l_2$.

Clustering accuracy :

$$r = 100 \frac{\sum_{i=1}^{k} a_i}{N}$$

- a_i est le nombre de points affectés à la bonne classe
- N est le nombre total de point

4. Expériences

Indice de Rand :

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

- $a = |S^*|$, où $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i, o_j \in Y_l\}$
- $b = |S^*|$, où $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $c = |S^*|$, où $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $d = |S^*|$, où $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i, o_j \in Y_l\}$

avec $1 \leq i, j \leq n, i \neq j, 1 \leq k, k_1, k_2 \leq r, k_1 \neq k_2, 1 \leq l, l_1, l_2 \leq s, l_1 \neq l_2$.

Clustering accuracy :

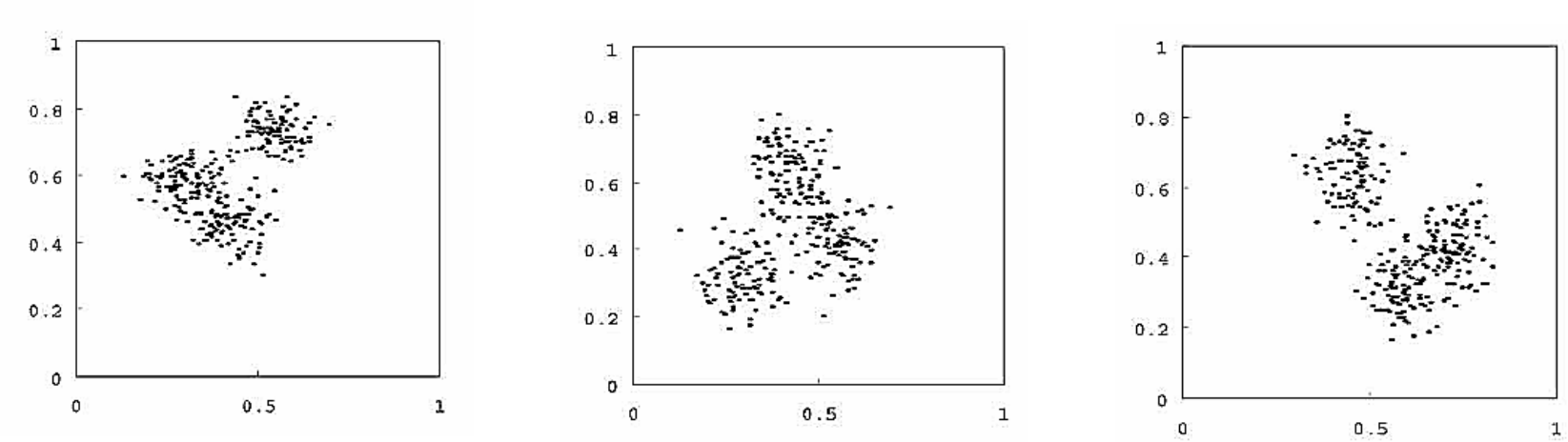
Erratum

$$r = \frac{\sum_{i=1}^{k} a_i}{N}$$

- a_i est le nombre de points affectés à la bonne classe
- N est le nombre total de point

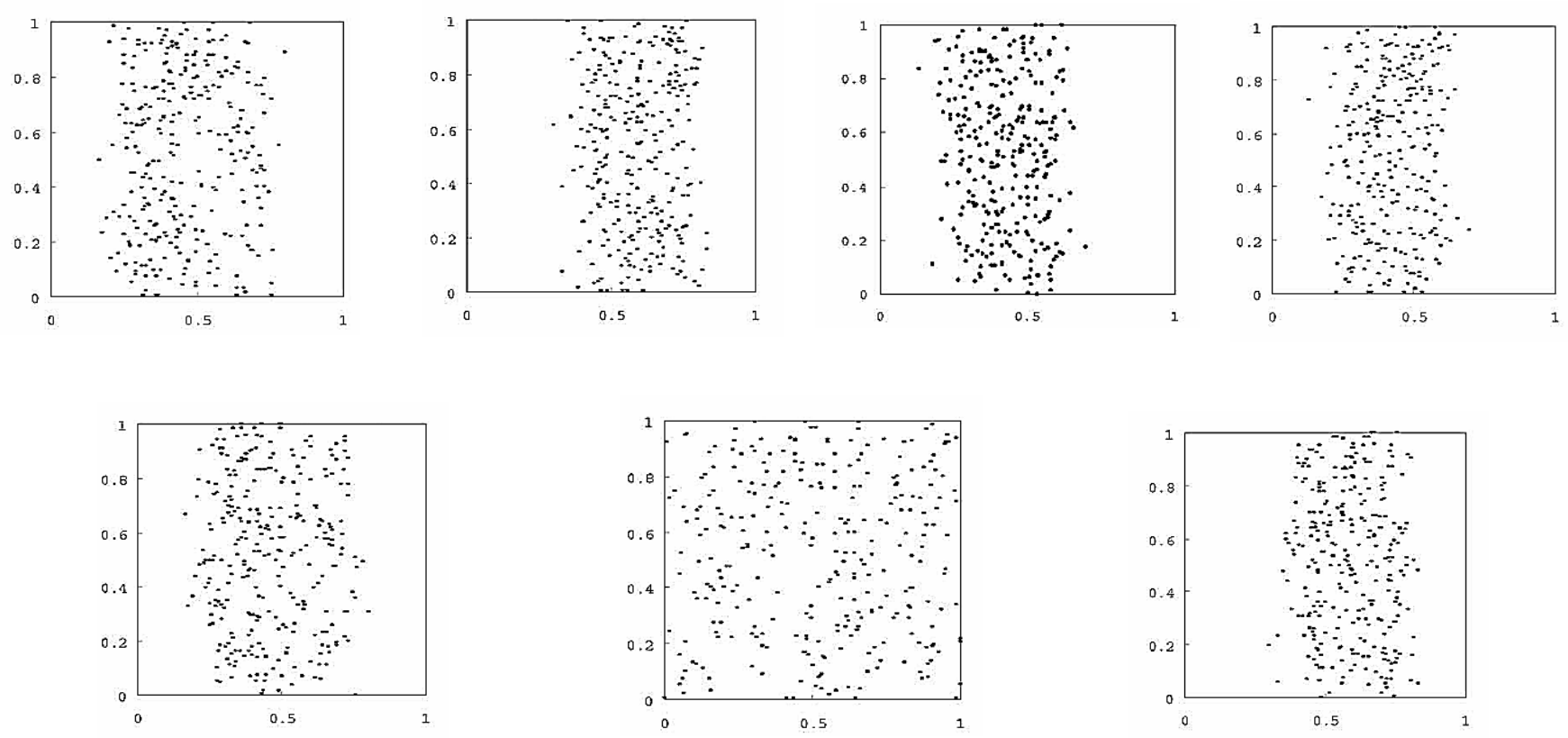
4. Expériences

X_1, X_2, X_3 : Données formant 3 classes nettes



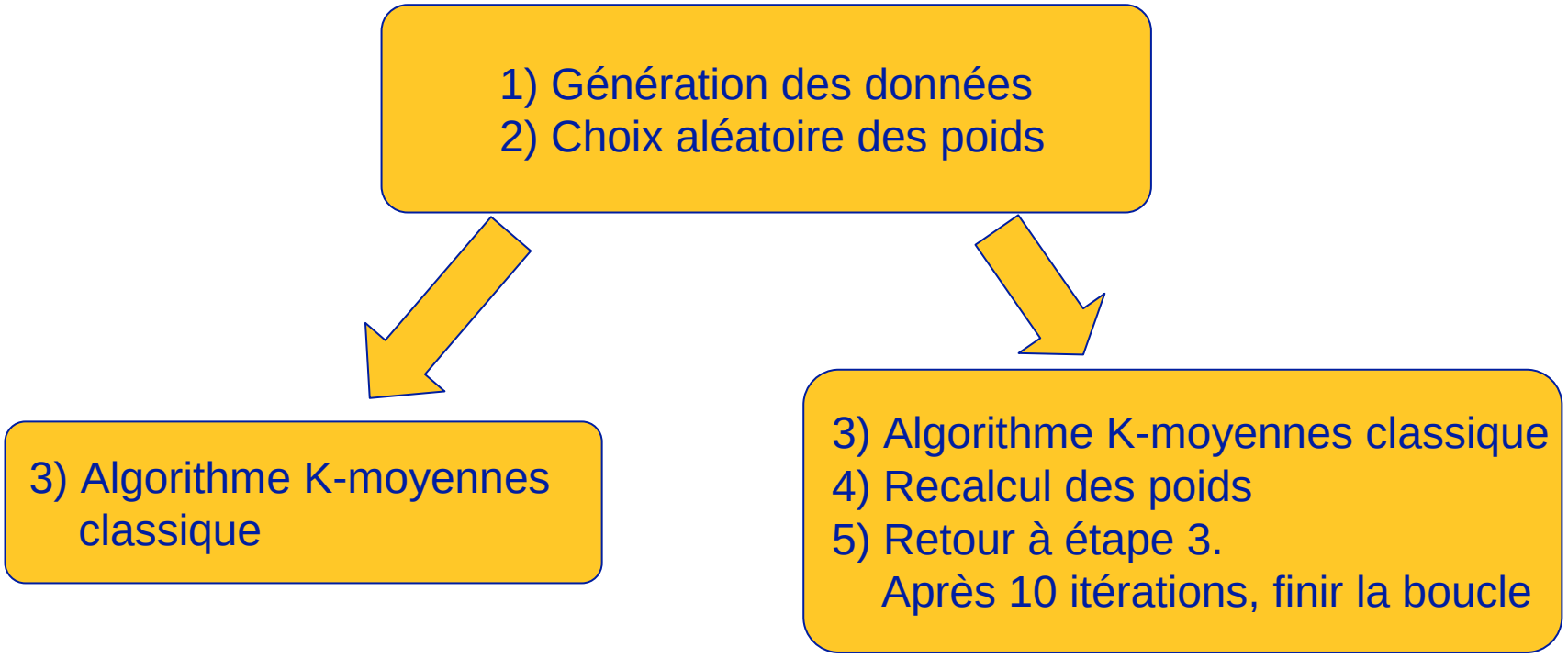
4. Expériences

X_4, X_5 : Bruit



4. Expériences

Expérience :



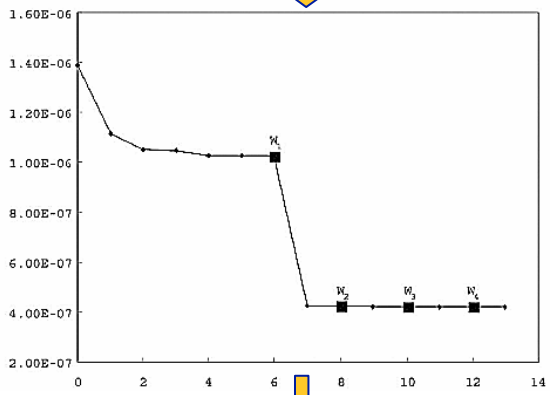
4. Expériences

Résultats

Algorithme K-moyennes classique

Num	weight0	weight1	weight2	weight3	weight4	Rand Index	Accuracy
1	0.2185	0.2845	0.0809	0.2457	0.1704	0.7577	0.7467
2	0.2968	0.3261	0.0982	0.1740	0.1049	0.9738	0.9800
3	0.3637	0.1018	0.1642	0.2899	0.0804	0.7766	0.7967
4	0.2661	0.1881	0.0680	0.2413	0.2365	0.6738	0.6033
5	0.3841	0.1989	0.0841	0.1500	0.1829	0.7795	0.7933
6	0.3337	0.0510	0.0496	0.2351	0.3305	0.6174	0.5367
7	0.3377	0.0285	0.1386	0.0844	0.4109	0.5661	0.4367
8	0.2804	0.2525	0.0821	0.0172	0.3678	0.5663	0.4367
9	0.3569	0.1190	0.0654	0.4327	0.0261	0.5545	0.3767
10	0.2503	0.1202	0.1236	0.3400	0.1658	0.5545	0.3733

K-moyennes avec pondération des poids dynamique



Num	weight0	weight1	weight2	weight3	weight4	Rand Index	Accuracy
1	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
2	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
3	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
4	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
5	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
6	0.3249	0.1362	0.1212	0.0814	0.3362	0.6204	0.5533
7	0.1204	0.0942	0.0850	0.0601	0.6403	0.5721	0.4500
8	0.1204	0.0942	0.0850	0.0601	0.6403	0.5721	0.4500
9	0.1092	0.0826	0.0772	0.6822	0.0487	0.5545	0.3767
10	0.1091	0.0826	0.0772	0.6824	0.0487	0.5545	0.3733

4. Expériences

Résultats

Algorithme K-moyennes classique

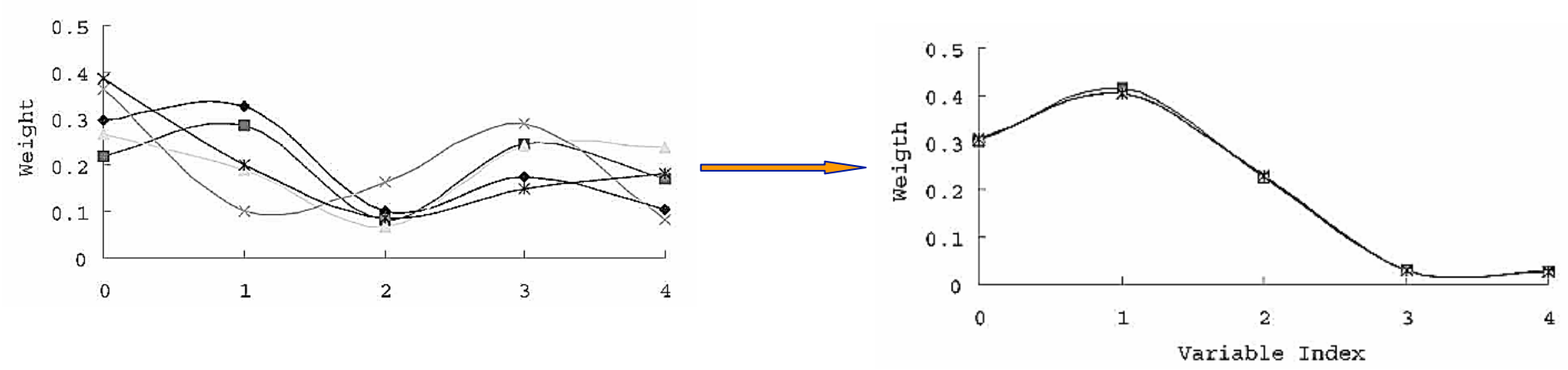
K-moyennes avec pondération des poids dynamique

Problème commun :
La classification dépend
des centroïdes et poids initiaux

4. Expériences

Solution commune :
Faire tourner les algorithmes plusieurs fois
et prendre le meilleur résultat.

Les poids convergent de façon similaire :



4. Expériences

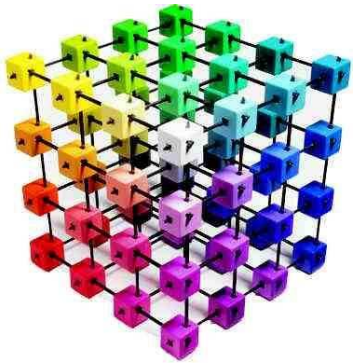
Solution commune :
Faire tourner les algorithmes plusieurs fois
et prendre le meilleur résultat.

Résultats :

Num	No Weights	Fixed Weights	Weights Changed
1	(0.4767, 0.5768)	(0.6764, 0.7317)	(0.8225, 0.8671)
2	(0.4833, 0.5796)	(0.6990, 0.7462)	(0.8453, 0.8809)
3	(0.5267, 0.6052)	(0.6871, 0.7429)	(0.7830, 0.8357)
4	(0.7200, 0.7652)	(0.6880, 0.7448)	(0.7893, 0.8403)
5	(0.7800, 0.7877)	(0.6938, 0.7445)	(0.8682, 0.8963)
6	(0.4764, 0.5780)	(0.6930, 0.7444)	(0.8337, 0.8713)
7	(0.7167, 0.7610)	(0.6960, 0.7479)	(0.7992, 0.8474)
8	(0.7767, 0.7884)	(0.6778, 0.7361)	(0.8003, 0.8478)
9	(0.7800, 0.7877)	(0.7040, 0.7515)	(0.8426, 0.8776)
10	(0.7200, 0.7589)	(0.6740, 0.7379)	(0.7810, 0.8341)
Average	(0.6457, 0.6989)	(0.6889, 0.7428)	(0.8156, 0.8599)

4. Expériences

Expérience 2 : Avec 2 jeux de données réels



Australian Credit Card data : 690 individus, 5 variables quantitatives, 8 variables qualitatives.

Heart Diseases : 270 individus, 6 variables quantitatives, 9 variables qualitatives.

Objectifs :

1) Évaluer l'impact de β, paramètre utilisé dans la formule de calcul des poids.

2) Comparer les résultats obtenus avec les études précédentes réalisées sur ces mêmes jeux de données.

4. Expériences

Résultats

Australian Credit Card data

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.85																	1	1	1	
0.84																				
0.83																	1	1	1	
0.82																4				3
0.81	4	4	6	5	7	13	10	7	12	11			8	46	39	42				13
0.80	32	32	27	23	22	15	19	19	10	16			6	11	18	11				6
0.79	6	6	8	8	7	7	7	8	8	1			2							4
0.78	3	3	3	3	4	2	1	2	2	5			3							2
0.77	7	6	6	6	4	5	5	5	7	5			19				3	3	3	2
0.76	1	2	2	2	4	5	5	6	3								3	3	3	10
0.75									4	8							4	4	4	3
0.74																	4	4	4	3
0.73								1									3	3	3	2
0.72								1												
$\leq$0.71	47	47	48	53	52	53	53	53	54	54	100	100	62	43	43	43	81	81	81	52

4. Expériences

Résultats

Australian Credit Card data

+0.02 de prévision que les études précédentes !

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.85																	1	1	1	
0.84																				
0.83																	1	1	1	
0.82																4				3
0.81	4	4	6	5	7	13	10	7	12	11			8	46	39	42				13
0.80	32	32	27	23	22	15	19	19	10	16			6	11	18	11				6
0.79	6	6	8	8	7	7	7	8	8	1			2							4
0.78	3	3	3	3	4	2	1	2	2	5			3							2
0.77	7	6	6	6	4	5	5	5	7	5			19				3	3	3	2
0.76	1	2	2	2	4	5	5	6	3								3	3	3	10
0.75									4	8							4	4	4	3
0.74																	4	4	4	3
0.73								1									3	3	3	2
0.72								1												
≤0.71	47	47	48	53	52	53	53	53	54	54	100	100	62	43	43	43	81	81	81	52

4. Expériences

Résultats

Australian Credit Card data

+0.02 de prévision que les études précédentes !

Erratum : 0.85 est aussi atteint lorsque β = 8

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=	β=8	β=9	β=1	β=0
0.85																	1	1	1	
0.84																				
0.83																	1	1	1	
0.82																4				3
0.81	4	4	6	5	7	13	10	7	12	11			8	46	39	42				13
0.80	32	32	27	23	22	15	19	19	10	16			6	11	18	11				6
0.79	6	6	8	8	7	7	7	8	8	1			2							4
0.78	3	3	3	3	4	2	1	2	2	5			3							2
0.77	7	6	6	6	4	5	5	5	7	5			19				3	3	3	2
0.76	1	2	2	2	4	5	5	6	3								3	3	3	10
0.75									4	8							4	4	4	3
0.74																	4	4	4	3
0.73								1									3	3	3	2
0.72								1												
≤0.71	47	47	48	53	52	53	53	53	54	54	100	100	62	43	43	43	81	81	81	52

4. Expériences

Résultats

Heart Diseases

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.85																	1	1	1	
0.84						1	3		5											
0.83	2	4	5	6	8	11	13	14	2	13							5	5	5	13
0.82					1				6								4	4	4	
0.81			1	1	1	2	6	50	53	5							2	2	2	
0.80					1	52	72	21	10	44							3	3	3	49
0.79			1	5	63	17			3	14				4	1	8	4	4	4	23
0.78	93	91	88	83	7	9			4	6			4	41	97	91				
0.77					12							1	88	55	2		1	1	1	
0.76												73					3	3	3	
0.75												5					2	2	2	
0.74												2	2				5	5	5	
0.73								1			1	3					7	7	7	
0.72								1				2	4							
≤0.71	5	5	5	5	7	8	9	13	17	18	99	14	2			1	63	63	63	15

4. Expériences

Résultats

Heart Diseases

+0.02 de prévision que les études précédentes !

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.85																	1	1	1	
0.84						1	3		5											
0.83	2	4	5	6	8	11	13	14	2	13							5	5	5	13
0.82					1				6								4	4	4	
0.81			1	1	1	2	6	50	53	5							2	2	2	
0.80					1	52	72	21	10	44							3	3	3	49
0.79			1	5	63	17			3	14				4	1	8	4	4	4	23
0.78	93	91	88	83	7	9			4	6			4	41	97	91				
0.77					12							1	88	55	2		1	1	1	
0.76												73					3	3	3	
0.75												5					2	2	2	
0.74												2	2				5	5	5	
0.73								1			1	3					7	7	7	
0.72								1				2	4							
≤0.71	5	5	5	5	7	8	9	13	17	18	99	14	2			1	63	63	63	15

4. Expériences

Qu'en est-il des poids ?

Credit Card Data				Heart Disease Data			
v_1	0.0130	v_9	0.1670	v_1	0.1176	v_9	0.0122
v_2	0.1652	v_{10}	0.0139	v_2	0.0091	v_{10}	0.1553
v_3	0.1871	v_{11}	0.0088	v_3	0.0069	v_{11}	0.0104
v_4	0.0167	v_{12}	0.0083	v_4	0.1492	v_{12}	0.0070
v_5	0.0167	v_{13}	0.0167	v_5	0.3331	v_{13}	0.0122
v_6	0.0044	$^{**}v_{14}$	0.0044	v_6	0.0123		
v_7	0.0093	$^{**}v_{15}$	0.0021	$^{**}v_7$	0.0064		
v_8	0.5167			v_8	0.1684		

4. Expériences

Qu'en est-il des poids ?

Credit Card Data				Heart Disease Data			
v_1	0.0130	v_9	0.1670	v_1	0.1176	v_9	0.0122
v_2	0.1652	v_{10}	0.0139	v_2	0.0091	v_{10}	0.1553
v_3	0.1871	v_{11}	0.0088	v_3	0.0069	v_{11}	0.0104
v_4	0.0167	v_{12}	0.0083	v_4	0.1492	v_{12}	0.0070
v_5	0.0167	v_{13}	0.0167	v_5	0.3331	v_{13}	0.0122
v_6	0.0044	v_{14}	0.0044	v_6	0.0123		
v_7	0.0093	v_{15}	0.0021	v_7	0.0064		
v_8	0.5167			v_8	0.1684		

4. Expériences

Résultats en supprimant les variables au poids faible

Australian Credit Card data

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.86																		1	1	
0.85																				
0.84																				
0.83																				
0.82																2	4			
0.81	2	1	1	2	1		1						6	32	51	41	45	2	2	
0.80	35	36	40	38	38	37	36	29	28	24			7	33	16	24	19			31
0.79	10	10	6	7	5	4	3	11	10	9			1	1		1				17
0.78	3	3	3	3	4	4	4	3	1				3					1	1	10
0.77	20	20	20	20	20	20	20	21	15	11			29					2	2	10
0.76									10	16			1					4	4	
0.75																		5	5	
0.74															1			2	2	2
0.73								1										4	4	
0.72								1										2	2	
≤0.71	30	30	30	30	32	35	36	36	36	40	100	100	53	34	32	32	32	77	77	30

4. Expériences

Résultats en supprimant les variables au poids faible

Australian Credit Card data

+0.03 de prévision que les études précédentes !

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.86																		1	1	
0.85																				
0.84																				
0.83																				
0.82																2	4			
0.81	2	1	1	2	1		1						6	32	51	41	45	2	2	
0.80	35	36	40	38	38	37	36	29	28	24			7	33	16	24	19			31
0.79	10	10	6	7	5	4	3	11	10	9			1	1		1				17
0.78	3	3	3	3	4	4	4	3	1				3					1	1	10
0.77	20	20	20	20	20	20	20	21	15	11			29					2	2	10
0.76									10	16			1					4	4	
0.75																		5	5	
0.74															1			2	2	2
0.73								1										4	4	
0.72								1										2	2	
≤0.71	30	30	30	30	32	35	36	36	36	40	100	100	53	34	32	32	32	77	77	30

4. Expériences

Résultats en supprimant les variables au poids faible

Heart Diseases

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.84	1									5							4	4	4	
0.83	12	12	11	8	1	1	1	2	5	14							4	4	4	23
0.82	5	7	6	5	14	6	4										2	2	2	
0.81	33	61	48	48	15	26	73		3	3							3	3	3	26
0.80	37	13	28	28	52	49	2	78	72	60							3	3	3	44
0.79	6	1	1	4	14	15	17	17	14	9				2	5	6	5	5	5	
0.78													2	81	92	90	2	2	2	
0.77												5	91	14			8	8	8	
0.76												11	1				3	3	3	
0.75														2			2	2	2	
0.74												71					1	1	1	
0.73																	5	5	5	
0.72																	3	3	3	
≤0.71	6	6	6	7	4	3	3	3	6	9	100	13	6	1	3	4	55	55	55	7

4. Expériences

Résultats en supprimant les variables au poids faible

Heart Diseases

+0.01 de prévision que les études précédentes !

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.84	1									5							4	4	4	
0.83	12	12	11	8	1	1	1	2	5	14							4	4	4	23
0.82	5	7	6	5	14	6	4										2	2	2	
0.81	33	61	48	48	15	26	73		3	3							3	3	3	26
0.80	37	13	28	28	52	49	2	78	72	60							3	3	3	44
0.79	6	1	1	4	14	15	17	17	14	9				2	5	6	5	5	5	
0.78													2	81	92	90	2	2	2	
0.77												5	91	14			8	8	8	
0.76												11	1				3	3	3	
0.75														2			2	2	2	
0.74												71					1	1	1	
0.73																	5	5	5	
0.72																	3	3	3	
≤0.71	6	6	6	7	4	3	3	3	6	9	100	13	6	1	3	4	55	55	55	7

4. Expériences

Résultats en supprimant les variables au poids faible

+0.01 de prévision que les études précédentes !

Erratum : Les résultats sont ici moins bons qu'avant la suppression des variables

Heart Diseases

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.84	1									5							4	4	4	
0.83	12	12	11	8	1	1	1	2	5	14							4	4	4	23
0.82	5	7	6	5	14	6	4										2	2	2	
0.81	33	61	48	48	15	26	73		3	3							3	3	3	26
0.80	37	13	28	28	52	49	2	78	72	60							3	3	3	44
0.79	6	1	1	4	14	15	17	17	14	9				2	5	6	5	5	5	
0.78													2	81	92	90	2	2	2	
0.77												5	91	14			8	8	8	
0.76												11	1				3	3	3	
0.75														2			2	2	2	
0.74												71					1	1	1	
0.73																	5	5	5	
0.72																	3	3	3	
≤0.71	6	6	6	7	4	3	3	3	6	9	100	13	6	1	3	4	55	55	55	7

Sommaire

5. Limites de l'étude

1) Le choix de β semble vraiment empirique.

L'étude constate simplement que selon la valeur de β, les résultats de la classification varient beaucoup, et que le meilleur résultat est meilleur que les résultats obtenus avec d'autres algorithmes des K-moyennes.

Constatation, mais pas interprétation.

5. Limites de l'étude

Résultats en supprimant les variables au poids faible

Heart Diseases

Accuracy	β=-10	β=-9	β=-8	β=-7	β=-6	β=-5	β=-4	β=-3	β=-2	β=-1	β=2	β=3	β=4	β=5	β=6	β=7	β=8	β=9	β=10	β=0
0.84	1									5							4	4	4	
0.83	12	12	11	8	1	1	1	2	5	14							4	4	4	23
0.82	5	7	6	5	14	6	4										2	2	2	
0.81	33	61	48	48	15	26	73		3	3							3	3	3	26
0.80	37	13	28	28	52	49	2	78	72	60							3	3	3	44
0.79	6	1	1	4	14	15	17	17	14	9				2	5	6	5	5	5	
0.78													2	81	92	90	2	2	2	
0.77												5	91	14			8	8	8	
0.76												11	1				3	3	3	
0.75														2			2	2	2	
0.74												71					1	1	1	
0.73																	5	5	5	
0.72																	3	3	3	
≤0.71	6	6	6	7	4	3	3	3	6	9	100	13	6	1	3	4	55	55	55	7

2) Complexité de l'algorithme ?

L'article indique que la complexité est de O(tmnk) avec :

- k est le nombre de classes ;
- m est le nombre de variables ;
- n est le nombre d'individus ;
- t est le nombre d'itérations de l'algorithme.

t ne devrait pas figurer dans O, car en incluant dedans, la complexité de l'algorithme n'est plus du tout estimable.

2) Complexité de l'algorithme ?

Exemple avec deux algorithmes de tri

bubble sort est **O(n²),** n étant le nombre d'éléments à classer.
En utilisant un t indiquant le nombre d'itérations, la complexité de bubble sort pourrait être également être notée **O(t)** (car une itération est O(1)).

quicksort sort est **O(nlogn),** n étant le nombre d'éléments à classer.
En utilisant un t indiquant le nombre d'itérations, la complexité de quicksort pourrait être également être notée **O(t)** (car une itération est O(1)).

Conclusion : En utilisant un t indiquant le nombre d'itérations, cela rend les complexités incomparables entre elles. Même si cela permet d'évaluer la complexité d'une seule itération.

3) Mesure de qualité de l'indice de qualité ?

Cette étude utilise l'indice de Rand et le clustering accuracy.

Nous avons vu les différences entre ses deux indices sont parfois très importantes.

3) Mesure de qualité de l'indice de qualité ?

Cet étude utilise l'indice de Rand et le clustering accuracy.

Nous avons vu les différences entre ses deux indices sont parfois très importants.

Num	weight0	weight1	weight2	weight3	weight4	Rand Index	Accuracy
1	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
2	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
3	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
4	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
5	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
6	0.3249	0.1362	0.1212	0.0814	0.3362	0.6204	0.5533
7	0.1204	0.0942	0.0850	0.0601	0.6403	0.5721	0.4500
8	0.1204	0.0942	0.0850	0.0601	0.6403	0.5721	0.4500
9	0.1092	0.0826	0.0772	0.6822	0.0487	0.5545	0.3767
10	0.1091	0.0826	0.0772	0.6824	0.0487	0.5545	0.3733

3) Mesure de qualité de l'indice de qualité ?

Quid des autres ?

- Critères de Wallace :

$$W_I(C,C') = \frac{N_{11}}{\sum_{k=1}^{K} n_k (n_k - 1)/2}$$

$$W_{II}(C,C') = \frac{N_{11}}{\sum_{k'=1}^{K'} n'_{k'} (n'_{k'} - 1)/2}$$

- Folks et Mallows : $F(C,C') = \sqrt{W_I(C,C') W_{II}(C,C')}$

- Indice de Jacard : $J(C,C) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}}$

Conclusion

1) Une étude très intéressante améliorant un algorithme classique de clarification (K-moyennes)

2) Une approche dynamique des poids semblant bien fondée et répondant à un besoin existant.

3) Les résultats semblent prometteurs mais auraient mérités à être mieux explorés.


franck.dernoncourt@gmail.com