> Automated Variable Weighting in k-Means Type Clustering

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

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

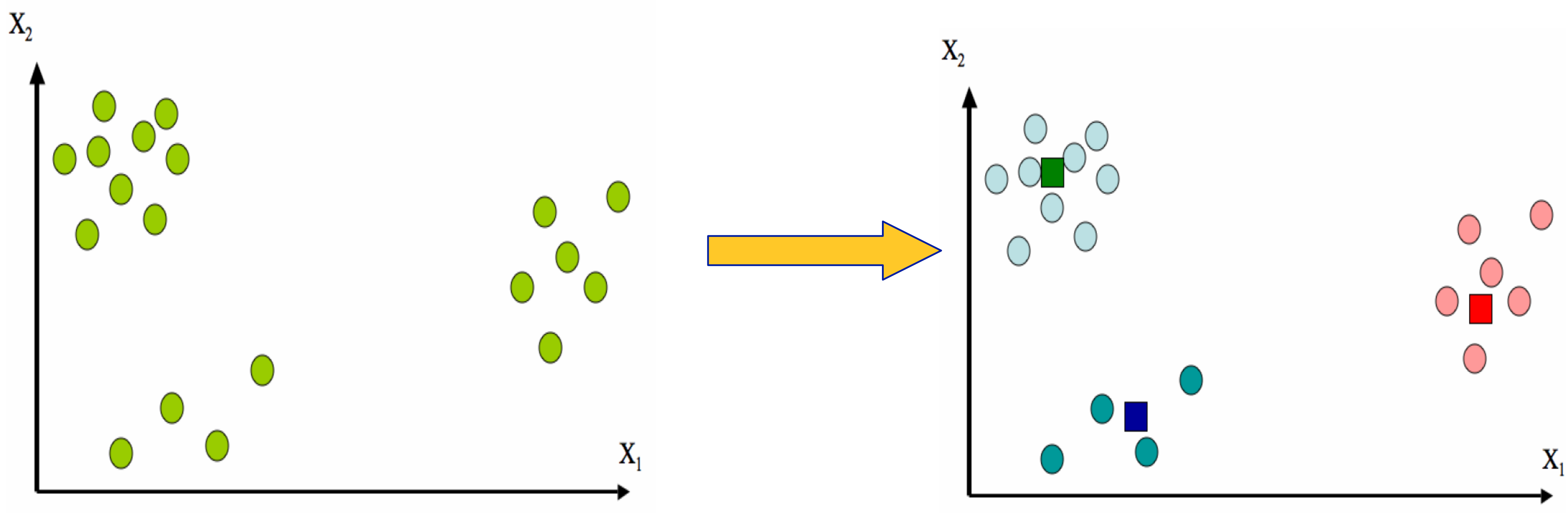
Presentation

NB: this presentation was originally written in French. I have translated it quickly so please don't get too upset if you encounter English mistakes. Instead, blame Gtanslate ;-) and drop me in email. Thanks! Franck.

Franck.Dernoncourt@gmail.com
28 Septembre 2010

Table of Contents

Clustering



Examples of clustering applications

- *Marketing*: find groups of similar customers
- *Biology*: classify plants according to their characteristics
- *Evolutionary algorithms* : improve crossovers' diversity.
- …

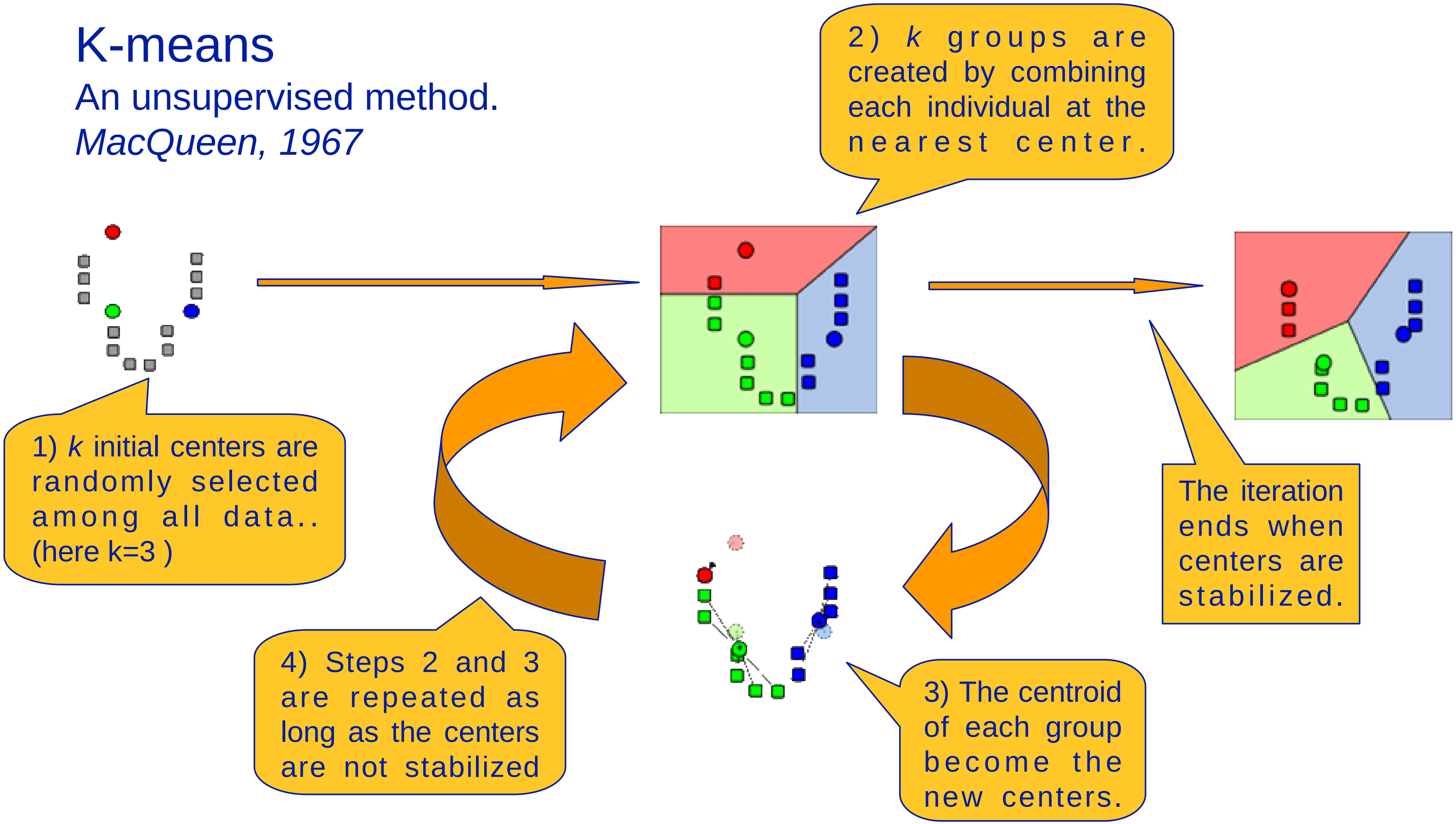
Table of Contents

2. k-means clustering

K-means

An unsupervised method.

MacQueen, 1967



2. k-means clustering

Limitations of the K-means algorithm:

- Require a metric
- Need to guess *a priori* the number of classes
- The choice of initial centers influences the results
- Sensitive to noise

2. k-means clustering

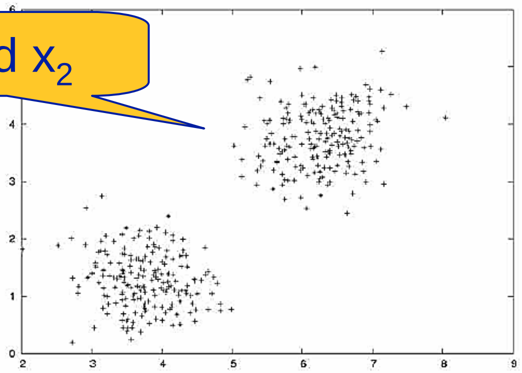
Limitations of the K-means algorithm:

- Require a metric
- Need to guess *a priori* the number of classes
- The choice of initial centers influences the results
- **Sensitive to noise**

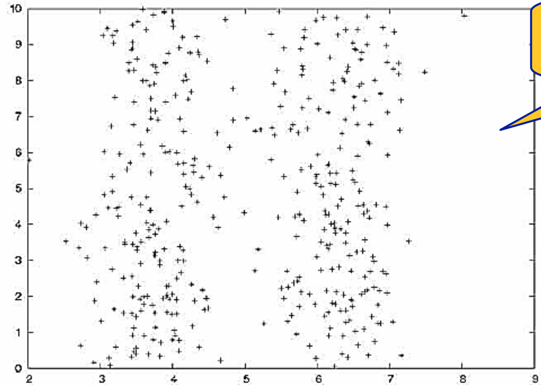
2. k-means clustering

3D example: x_1 and x_2 are "clusterable", x_3 is noise

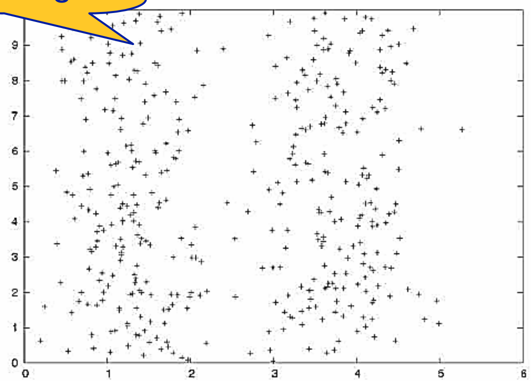
x_1 and x_2



x_1 and x_3



x_2 and x_3



Clustering result

Table of Contents

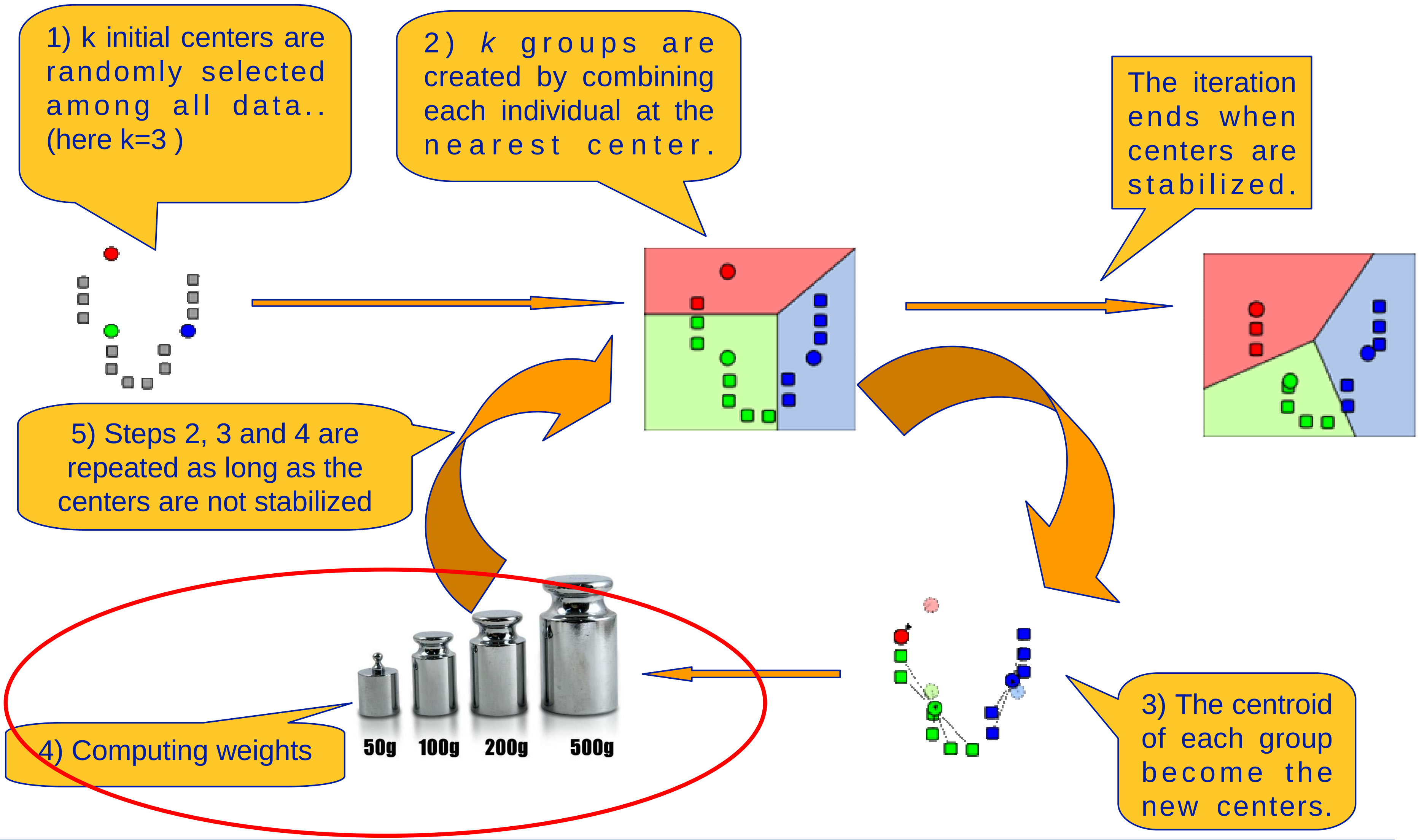
(Automated Variable Weighting in k-Means Type Clustering – 2005)

Idea: Weight each variable in order to give less weight to the variables affected by significant noise.

State of the art: Modha and Spangler already had this idea ... but they calculate the weights at the beginning of the algorithm.

Here, the weights will be calculated dynamically at each iteration of the algorithm K-means.

3. Automated Variable Weighting



Computing weights - Theorem

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

with

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

Where $u_{i,l}$ means that the object i is assigned to the class l

$d(x_{i,j}, z_{l,j})$ is the distance between objects x and z

h is the number of variables D_j such as $D_j \neq 0$

$z_{l,j}$ is the value of the variable j of the centroid of the cluster l

Computing weights - Theorem

Idea: give a low weight for the variables, the value of each individual on average away from the centroid

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

with

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

Where $u_{i,l}$ means that the object i is assigned to the class l

$d(x_{i,j}, z_{l,j})$ is the distance between objects x and z

h is the number of variables D_j such as $D_j \neq 0$

$z_{l,j}$ is the value of the variable j of the centroid of the cluster l

Computing weights

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

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

→ Lagrange multipliers!

Computing weights

The Lagrangian:

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

We get:

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

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

Computing weights

We can see: $\hat{w}_j = \left(\frac{-\hat{\alpha}}{\beta D_j}\right)^{\frac{1}{\beta-1}}$ for $1 \leq j \leq h$

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

Hence:

$$\hat{w}_j = \frac{1}{\sum_{t=1}^{h}\left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}}$$

QED!

Computing weights - Theorem

Idea: give a low weight for the variables, the value of each individual on average away from the centroid

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

with

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

Where $u_{i,l}$ means that the object i is assigned to the class l

$d(x_{i,j}, z_{l,j})$ is the distance between objects x and z

h is the number of variables D_j such as $D_j \neq 0$

$z_{l,j}$ is the value of the variable j of the centroid of the cluster l

Table of Contents

4. Experiments and results

Experiment 1: with a synthetic data set

5 Variables, 300 individuals :

X_1, X_2, X_3 : Data forming 3 clear classes

X_4, X_5 : Noise

As we know the 3 classes, we will compare the results obtained by the K-means algorithm with the standard K-means with dynamic weighting.

To make this comparison, we use the Rand index and the Clustering Accuracy in order to assess the performance of a classification compared to the desired classification.

4. Experiments and results

Rand index:

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

- $a = |S^*|$, 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 is the number of points assigned to the correct class
- N is the total number of points

4. Experiments and results

Rand index:

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

- $a = |S^*|$, 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 :

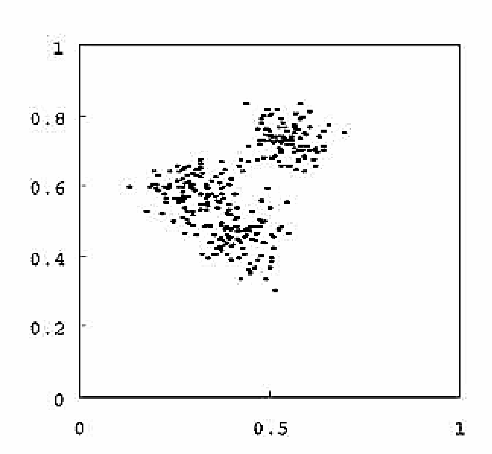
Erratum

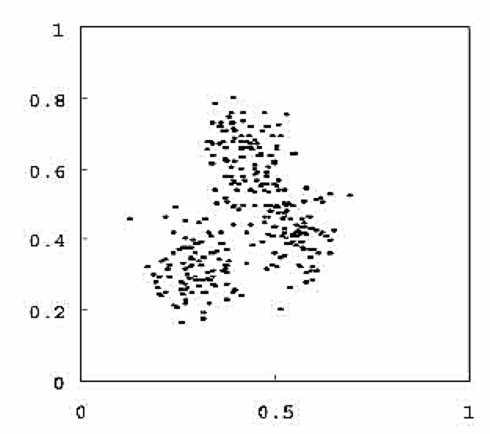
$$r = \cancel{100} \frac{\sum_{i=1}^{k} a_i}{N}$$

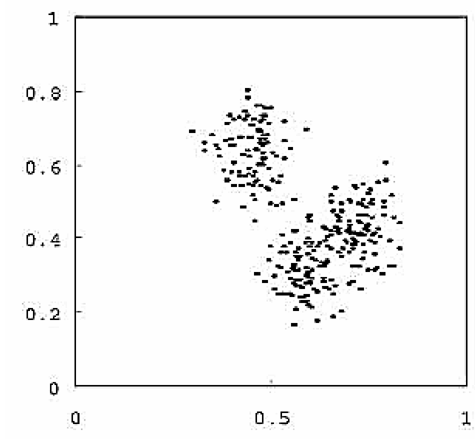
- a_i is the number of points assigned to the correct class
- N is the total number of points

4. Experiments and results

X_1, X_2, X_3: Data forming 3 clear classes

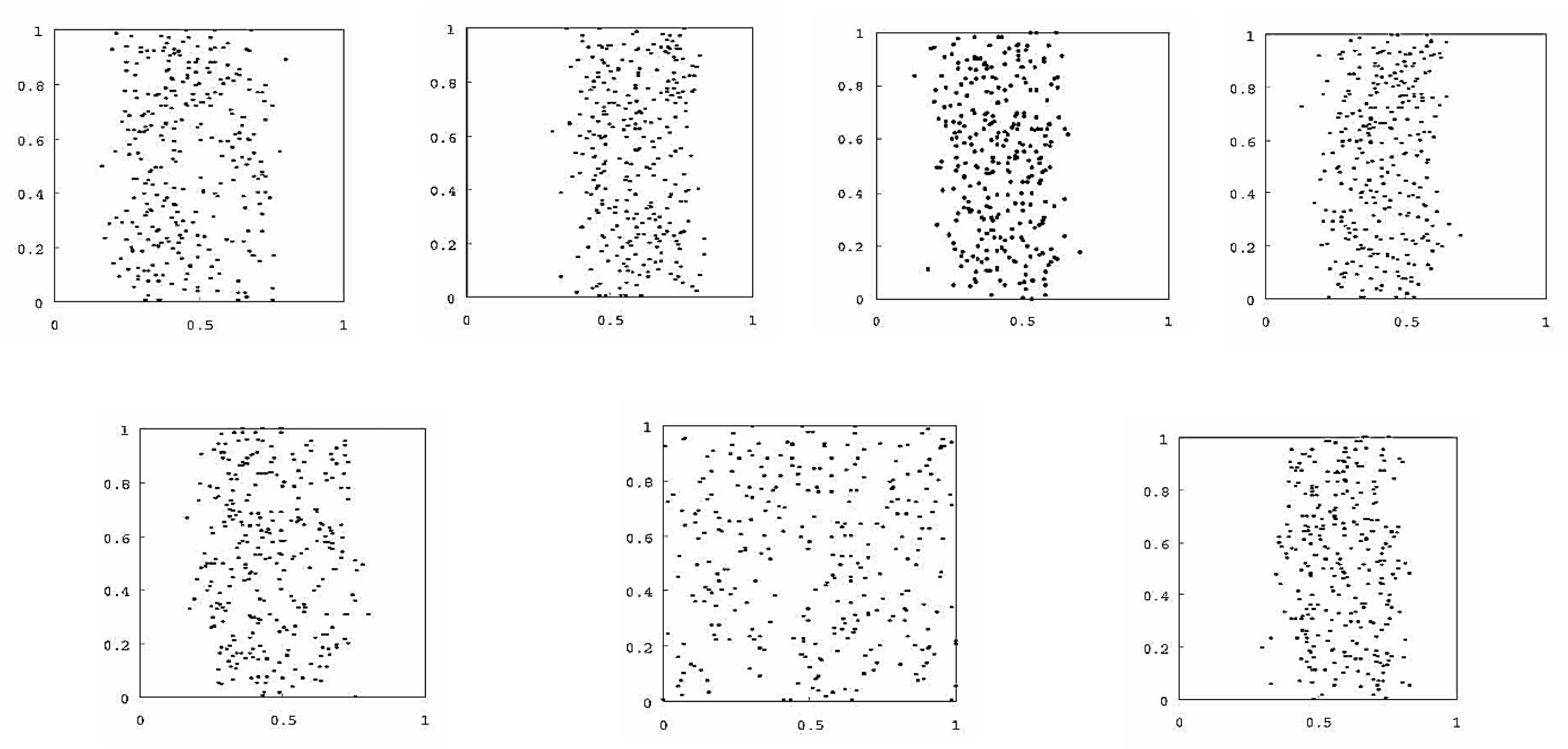






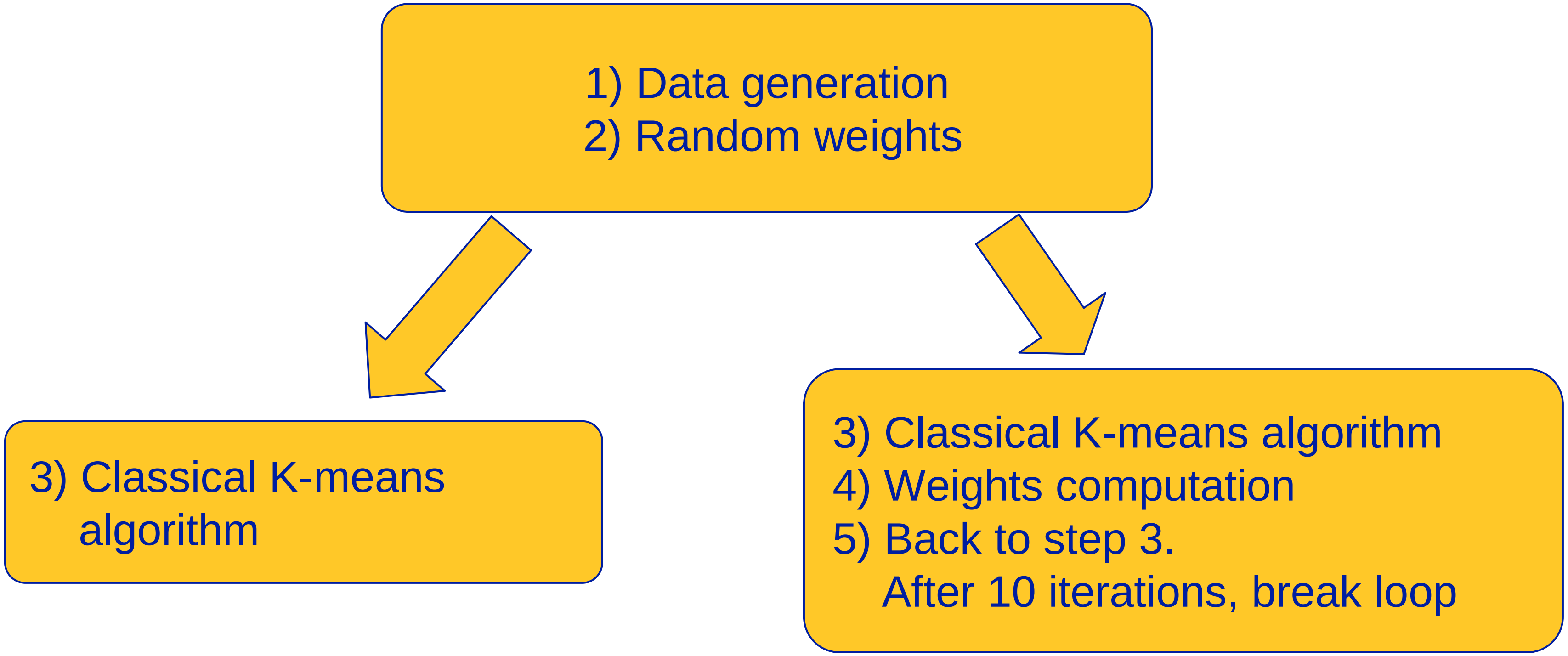
4. Experiments and results

X_4, X_5 :Noise



4. Experiments and results

Experiment:

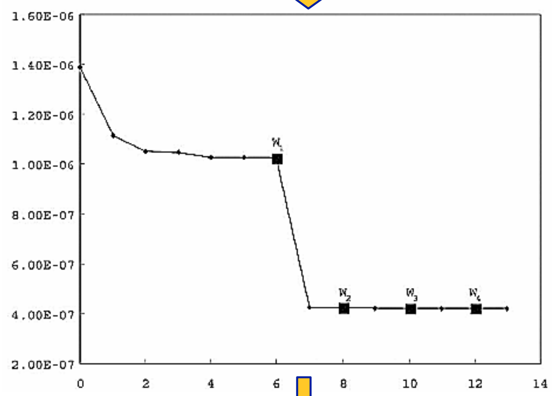


Results

Classical K-means algorithm

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-means algorithm with dynamic weighting



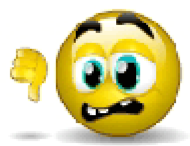
Num	weight0	weight1	weight2	weight3	weight4	Rand Index	Accuracy
1	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
2	0.3021	0.4137	0.2268	0.0301	0.0273	1.0000	1.0000
3	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
4	0.3078	0.4035	0.2310	0.0302	0.0274	0.9956	0.9967
5	0.3078	0.4035	0.2310	0.0302	0.0274	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

Results

Classical K-means algorithm

K-means algorithm with dynamic weighting

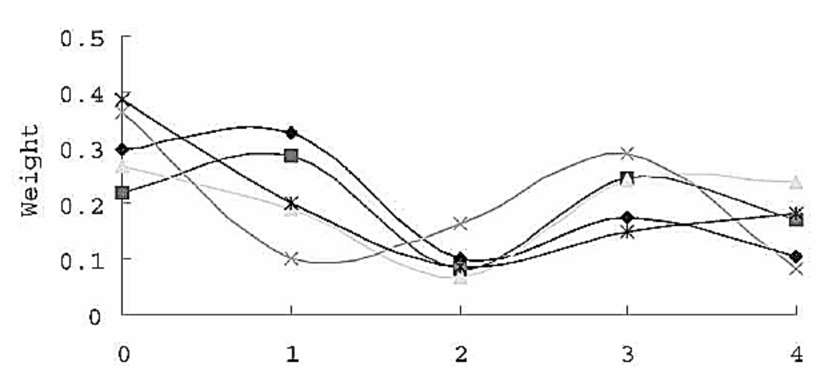
Common issue:
The classification depends on initial centroids and weights

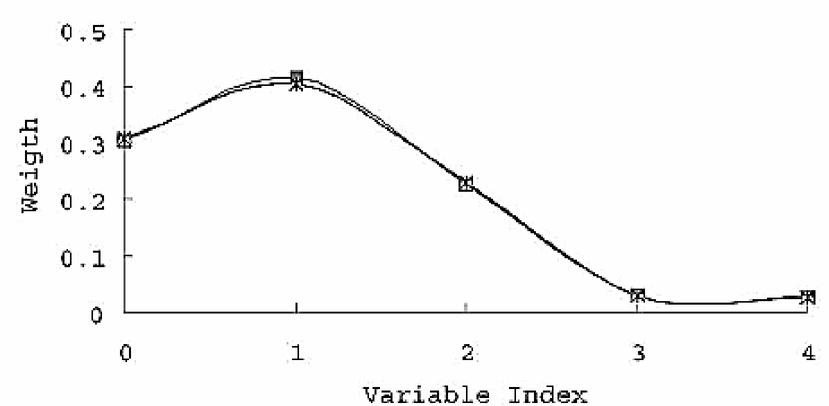


4. Experiments and results

Common solution:
Run the algorithms several times and
take the best result.

The weights converge similarly:





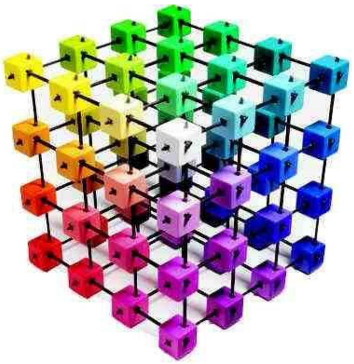
Common solution:
Run the algorithms several times and
take the best result.

Results:

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. Experiments and results

Experiment 2: With two real-world datasets



Australian Credit Card data : 690 individuals, 5 quantitative variables, 8 qualitative variables.

Heart Diseases : 270 individuals, 6 quantitative variables, 9 qualitative variables.

Objectives:

1) Assess the impact of β, the parameter used in the formula calculation of weight.

2) Compare the results with previous studies performed on these data sets.

4. Experiments and results

Results

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												
≤0.71	47	47	48	53	52	53	53	53	54	54	100	100	62	43	43	43	81	81	81	52

Results

Australian Credit Card data

+0.02 increase in prediction compared to previous studies!

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. Experiments and results

Results

Australian Credit Card data

+0.02 increase in prediction compared to previous studies!

Erratum : 0.85 is also reached when 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. Experiments and results

Results

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

Results

Heart Diseases

+0.02 increase in prediction compared to previous studies!

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

How about weights?

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		

How about weights?

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	$\cancel{{**}v_{14}}$	$\cancel{0.0044}$	v_6	0.0123		
v_7	0.0093	$\cancel{{**}v_{15}}$	$\cancel{0.0021}$	$\cancel{{**}v_7}$	$\cancel{0.0064}$		
v_8	0.5167			v_8	0.1684		

Results after removal of the least significant variables

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	
$\leq$0.71	30	30	30	30	32	35	36	36	36	40	100	100	53	34	32	32	32	77	77	30

4. Experiments and results

Results after removal of the least significant variables

Australian Credit Card data

+0.03 increase in prediction compared to previous studies!

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	
$\leq$0.71	30	30	30	30	32	35	36	36	36	40	100	100	53	34	32	32	32	77	77	30

4. Experiments and results

Results after removal of the least significant 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	
$\leq$0.71	6	6	6	7	4	3	3	3	6	9	100	13	6	1	3	4	55	55	55	7

Results after removal of the least significant variables

Heart Diseases

+0.01 increase in prediction compared to previous studies!

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. Experiments and results

Results after removal of the least significant variables

Erratum: The results here are worse than before the removal of variables

Heart Diseases

+0.01 increase in prediction compared to previous studies!

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

Table of Contents

5. Limitations of the study

1) The choice of β does seem empirically.

The study found that depending on the value of β, the classification results vary widely, and that the best result is better than the results obtained with other algorithms K-means.

Observation, but not interpretation.

5. Limitations of the study

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) Algorithm complexity analysis?

The article says that the complexity is O (tmnk) with:

- k is the number of classes,
- m is the number of variables;
- n is the number of individuals;
- t is the number of iterations of the algorithm.

t should not be included in O, since by including in it, the complexity of the algorithm is not at all assessable.

2) Algorithm complexity analysis?

Example with two sorting algorithms

Bubble sort is **O(n²)**, where n is the number of items to classify. Using a t indicating the number of iterations, the complexity of bubble sort could be also be denoted **O(t)** (since an iteration is O(1)).

Quicksort sort is **O(nlogn)** where n is the number of items to classify. Using a t indicating the number of iterations, the complexity of quicksort sort also be denoted **O(t)** (since an iteration is O(1)).

Conclusion: Using a t indicating the number of iterations makes complexities incomparable . Even if it is used to evaluate the complexity of a single iteration

3) Measurement of quality of the quality index?

This study uses the Rand index and Clustering Accuracy.

We saw that the differences between the two indices are sometimes very important.

3) Measurement of quality of the quality index?

This study uses the Rand index and Clustering Accuracy.

We saw that the differences between the two indices are sometimes very important.

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) Measurement of quality of the quality index?

How about the other quality indexes?

- 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) A very interesting study improving a classical algorithm for clustering (K-means)

2) A dynamic-weighting approach which seems well founded and which meets an existing need.

3) The results look promising but deserved to be better explored.


franck.dernoncourt@gmail.com