Algorithme K-Moyennes

CAFÉ SCIENTIFIQUE

AL-AMRANI Yassine

Plan

- INTRODUCTION (k-moyennes)
- DOMAINES D'APPLICATION
- AVANTAGES ET INCONVENIENTS
- ALGORITHME (classique)
- PROBLEME DE L'ALGORITHME
- LES ALTERNATIVES (les améliorations)
- HYBRIDATIONS

Introduction

Simpson's Family

School Employees

Females

Males

Domaines d'application

Domaine	Forme des données	Clusters
Text mining	Textes Mails	Textes proches Dossiers automatiques
Web mining	Textes et images	Pages web proches
BioInformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image
Web log analysis	Clickstream	Profils utilisateurs

Avantages et Inconvénients

Avantages de l'algorithme :

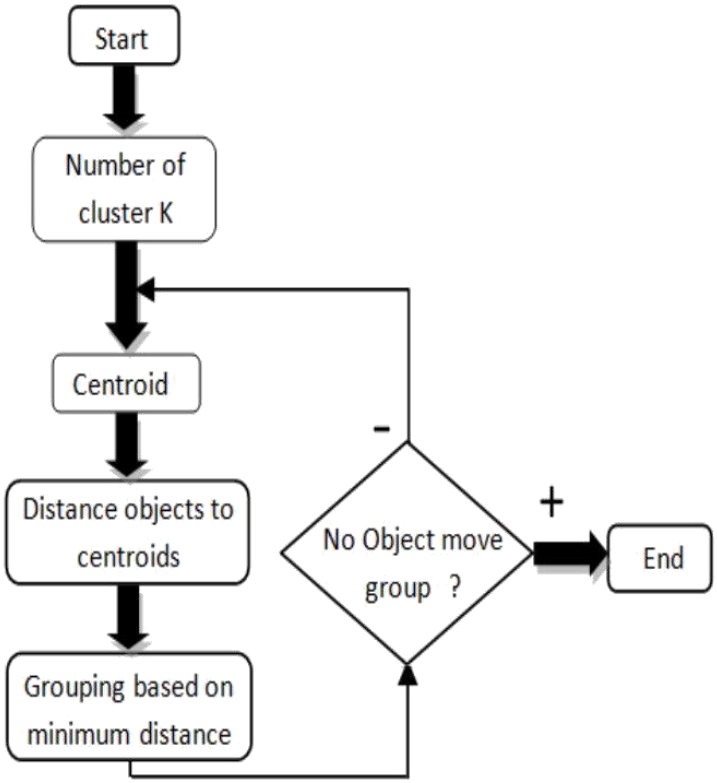
1) L'algorithme de k-means est très populaire du fait qu'il est très facile à comprendre et à mettre en œuvre.
2) Sa simplicité conceptuelle et sa rapidité
3) Applicable à des données de grandes tailles, et aussi à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

Avantages et Inconvénients

Inconvénients de l'algorithme :

1) Le nombre de classe doit être fixé au départ,
2) Le résultat dépend de tirage initial des centres des classes,
3) Les clusters sont construits par rapports à des objets inexistants (les milieux)

Algorithme (classique)



- ✓ Choisir K éléments initiaux "centres" des K groupes
- ✓ Placer les objets dans le groupe de centre le plus proche
- ✓ Recalculer le centre de gravité de chaque groupe
- ✓ Itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe

Algorithme (classique)

Entrée

Ensemble de N données, noté par x

Nombre de groupes souhaité, noté par k

Sortie

Une partition de K groupes $\{C_1, C_2, \ldots Ck\}$

Début

1) Initialisation aléatoire des centres Ck ;

Répéter

2) Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche :

$$\mathbf{x_i} \in c_k \text{ si } \forall_j |\mathbf{x_i} - \mu_k| = \min |\mathbf{x_i} - \mu_j| \qquad (1)$$

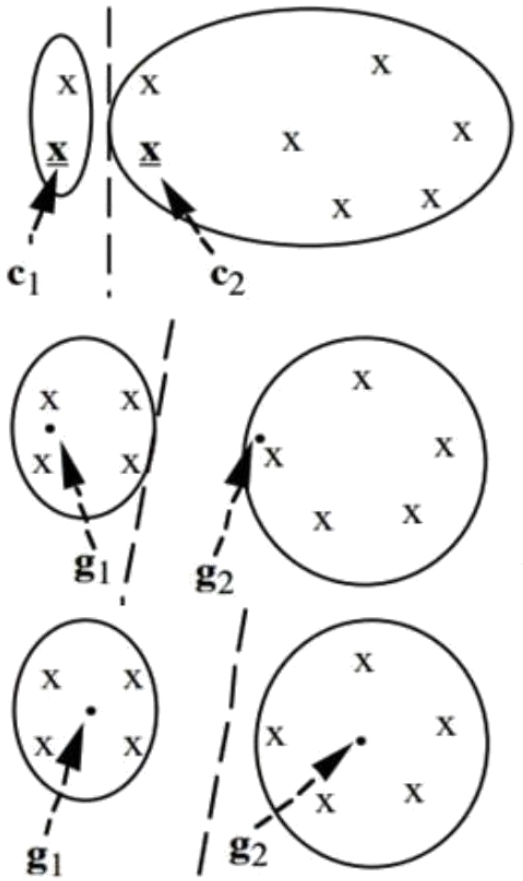
Avec μ_k le centre de la classe K ;

3) Représentation : Calculer les centres associe à la nouvelle partition ;

$$\mu_k = \frac{1}{N} \sum_{x \in c_k} x_i \qquad (2)$$

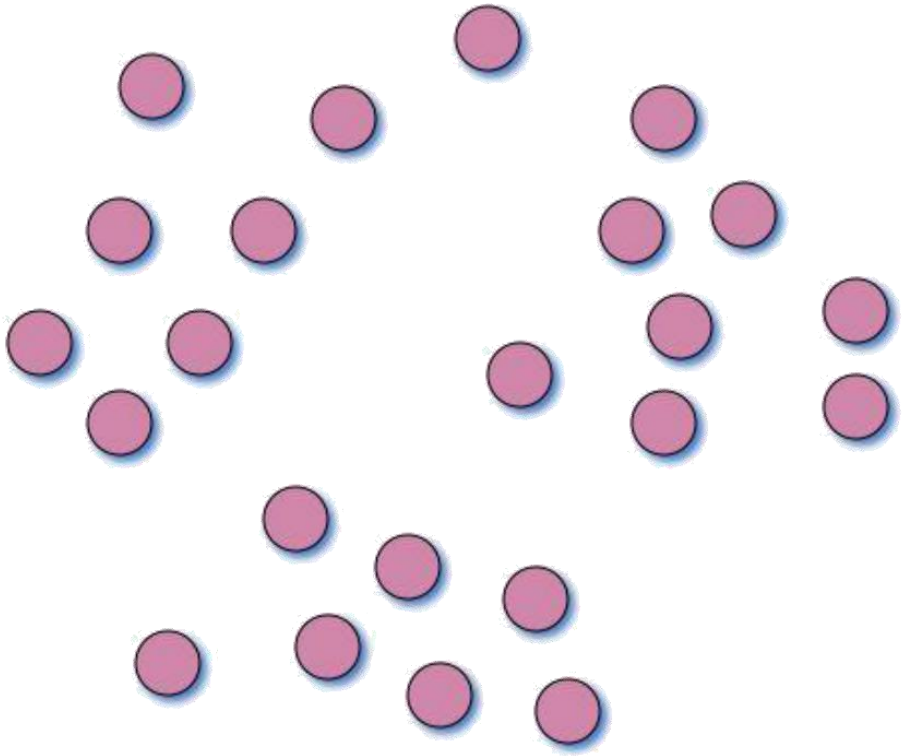
Jusqu'à convergence de l'algorithme vers une partition stable ;

Fin.



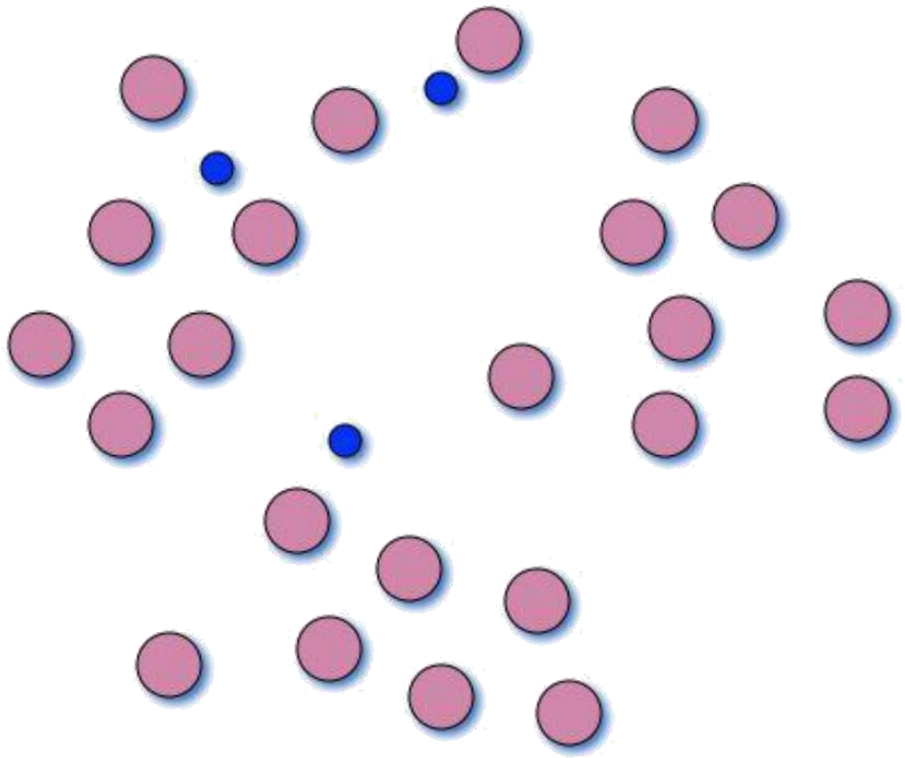
Algorithme (classique)

But: assigner les éléments aux groupes



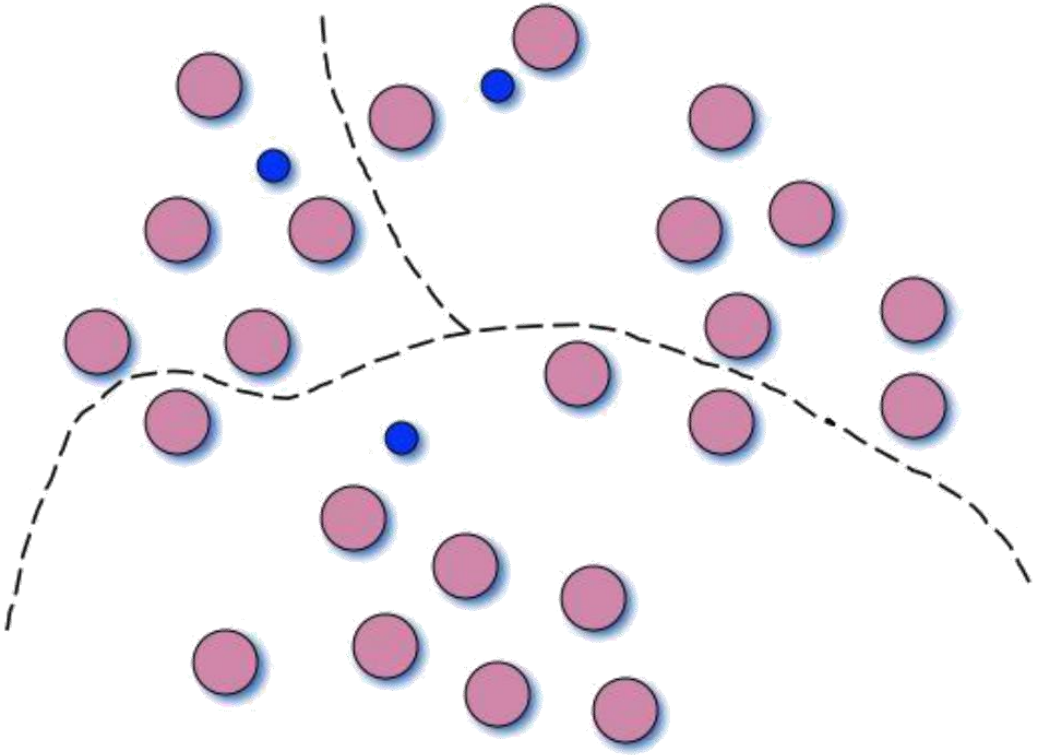
Algorithme (classique)

1: estimer des points K (aléatoirement)



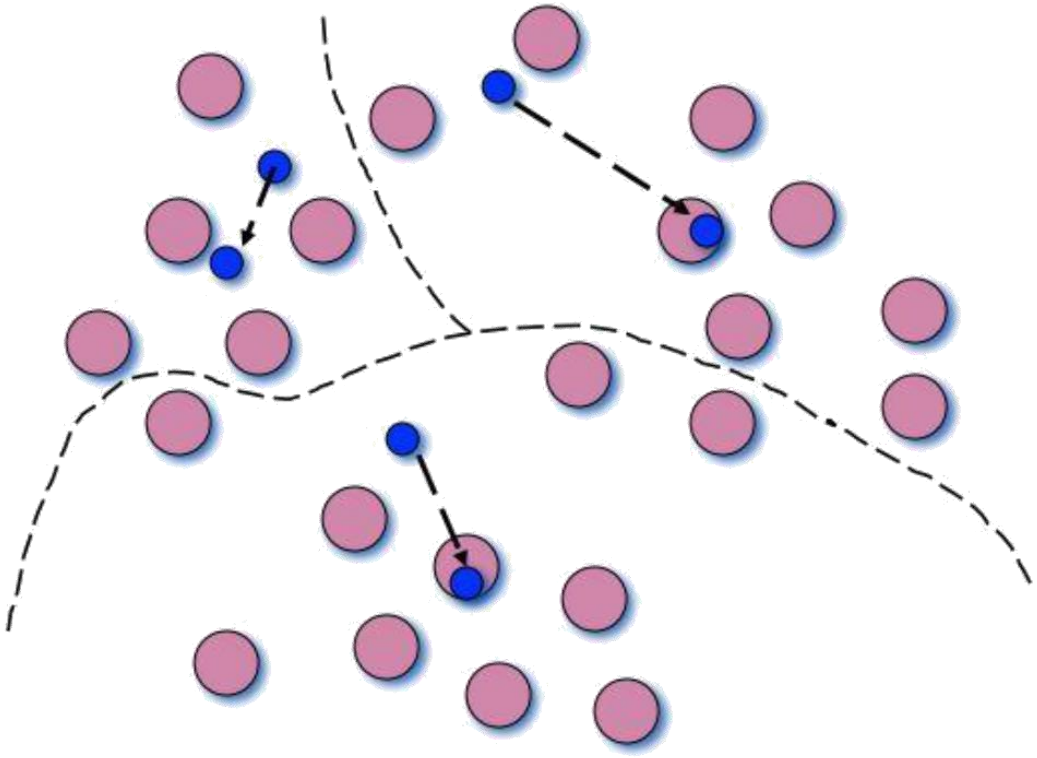
Algorithme (classique)

2: Assigner les éléments à ces groupes



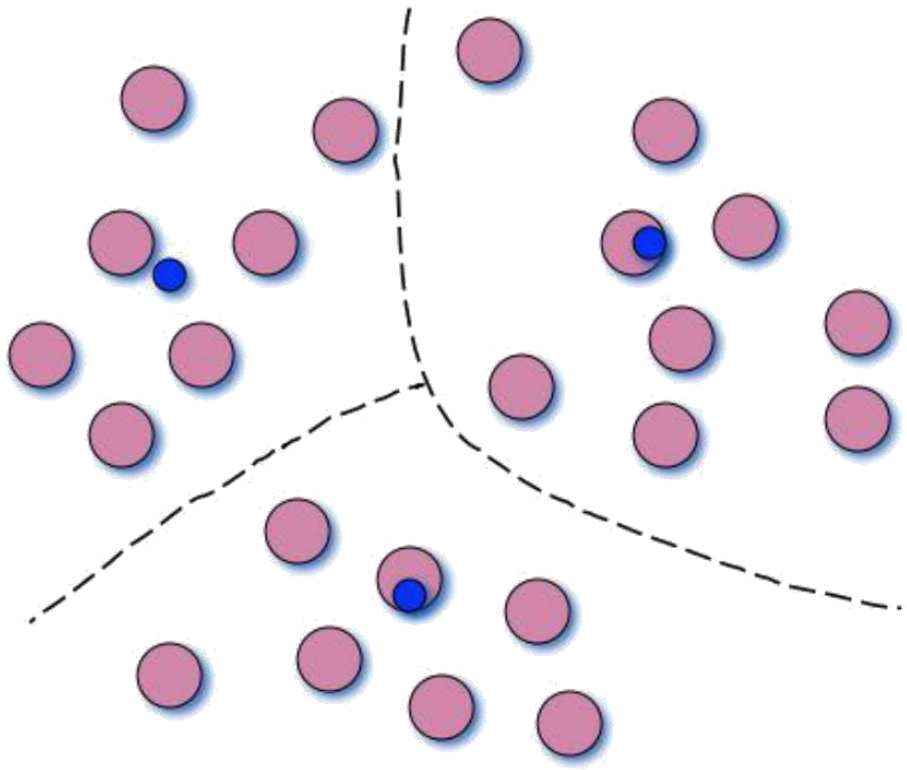
Algorithme (classique)

3: Déplacer les points K vers les centres



Algorithme (classique)

4: Réassigner les éléments et répéter jusqu'à stabilité



Algorithme (classique)

- 27-51-52-33-45-22-28-44-40-38-20-57
- Maximum amplitude = 57 – 20 = 37

	20	22	27	28	33	38	40	44	45	51	52	57
27	0.19	0.14	0.00	0.03	0.16	0.30	0.35	0.46	0.49	0.65	0.68	0.81
51	0.84	0.78	0.65	0.62	0.49	0.35	0.30	0.19	0.16	0.00	0.03	0.16
52	0.86	0.81	0.68	0.65	0.51	0.38	0.32	0.22	0.19	0.03	0.00	0.14
Min	0.19	0.14	0.00	0.03	0.16	0.30	0.30	0.19	0.16	0.00	0.00	0.14
Aff	1	1	1	1	1	1	2	2	2	2	3	3

- Cluster 1 : 20 - 22 - 27 - 28 - 33 - 38
  - Center : 168 / 6 = 28

Algorithme (classique)

- Cluster 2 : 40 - 44 - 45 - 51
  - Center : 180 / 4 = 45
- Cluster 3 : 52 - 57
  - Center : 109 / 2 = 54.5

	20	**22**	**27**	**28**	**33**	**38**	**40**	**44**	**45**	**51**	**52**	**57**
28	0.22	0.16	0.03	0.00	0.14	0.27	0.32	0.43	0.46	0.62	0.65	0.78
45	0.68	0.62	0.49	0.46	0.32	0.19	0.14	0.03	0.00	0.16	0.19	0.32
54.5	0.93	0.88	0.74	0.72	0.58	0.45	0.39	0.28	0.26	0.09	0.07	0.07
Min	0.22	0.16	0.03	0.00	0.14	0.19	0.14	0.03	0.00	0.09	0.07	0.07
Aff	1	1	1	1	1	2	2	2	2	3	3	3

Algorithme (classique)

- Cluster 1: 20 - 22 - 27 - 28 - 33
  - Center = 130 / 5 = 26
- Cluster 2: 38 - 40 - 44 - 45
  - Centrer = 167 / 4 = 41.75
- Cluster 3: 51 - 52 - 57
  - Center = 160 / 3 = 53.33

Problèmes de l'algorithme

Défauts de la méthode :

1) obligation de fixer K.

2) le résultat dépend fortement du choix des centres initiaux.

ne fournit pas nécessairement le résultat optimum

fournit un optimum local qui dépend des centres initiaux.

Les alternatives

Il existe plusieurs versions de l'algorithme k-moyennes, parmi eux on peut citer :

1) Global k-means,

2) Initialisation par le mal classé,

3) L'approche incrémental (ou Modified Fast Global Kmeans),

Les alternatives

Global k-means :

Entrée

Ensemble de N données, notés par x ;

Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes $\{ C_1 , C_2 , \ldots C_k \}$

Début

1) C_1 = Centre de gravité de l'ensemble des données ;

Répéter

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;

3) Trouver l'iéme centre :

Pour chaque donnée x **faire**

3.1) Considère x comme étant le iéme centre ;

3.2) Affecter les données aux plus proche centre ;

3.3) Calculer l'erreur quadratique pour Ci=x ;

$$\mathbf{J} = \sum_{i=1}^{k} \sum_{x_j \in c_i} \| x_j - c_i \|^2$$

Fin faire

3.4) Garder le centre Ci = x qui minimise l'erreur quadratique ;

4) Appliquer le k-means jusqu'à la convergence ;

Jusqu'à obtenir une partition en k groupes ;

Les alternatives

A. Likas et al. / Pattern Recognition 36 (2003) 451–461

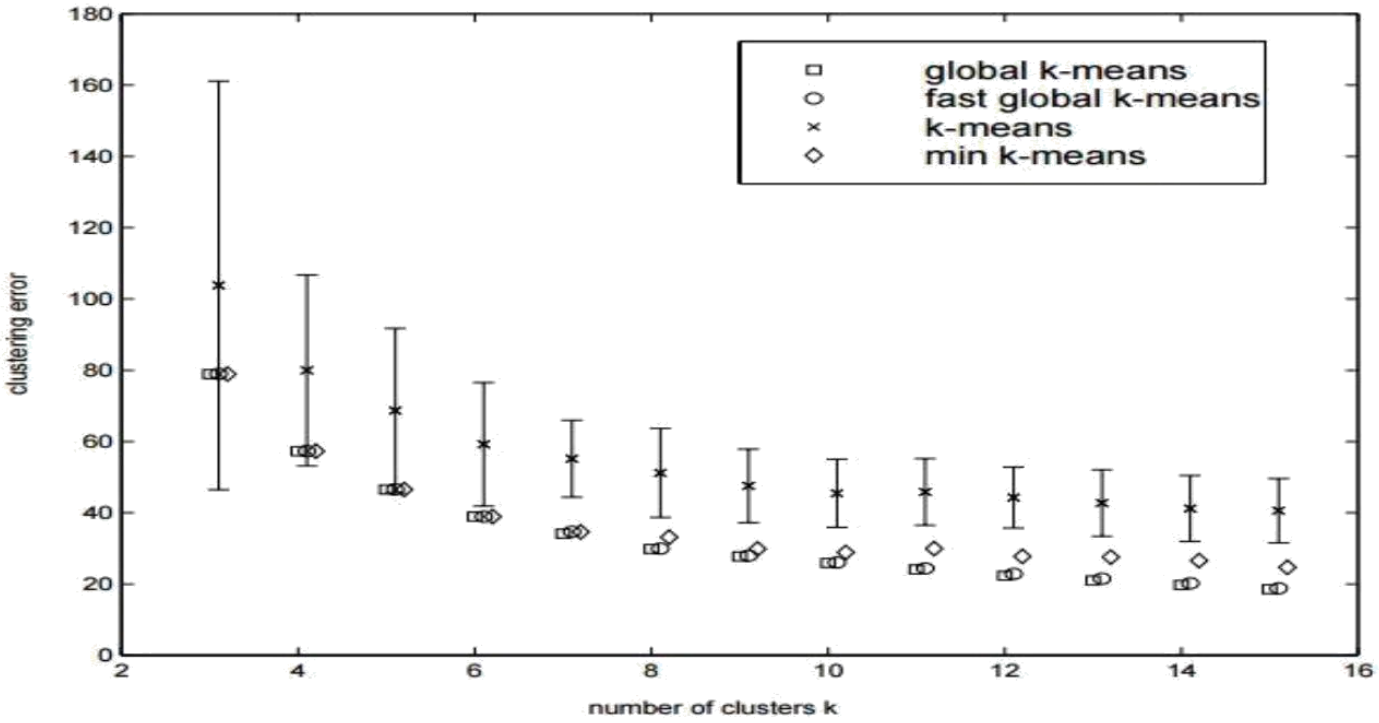


Fig. 2. Performance results for the Iris data set.

Les alternatives

Initialisation par le mal classé :

Début

1) Création d'une matrice de distance

2) Choisir les deux éléments les plus éloignés (ils représentent les deux premiers centres) ;

TANT QUE le nombre de classes souhaité n'est pas atteint **Faire**

3) Affecter les individus aux noyaux disponibles ;

4) Sélectionner un élément mal classé (celui qui possède la plus grande distance de son centre le plus proche) ;

5) Ajouter cet individu à l'ensemble des noyaux ;

6) Augmenter le nombre des noyaux ;

Fin TANTQUE

Fin

Les alternatives

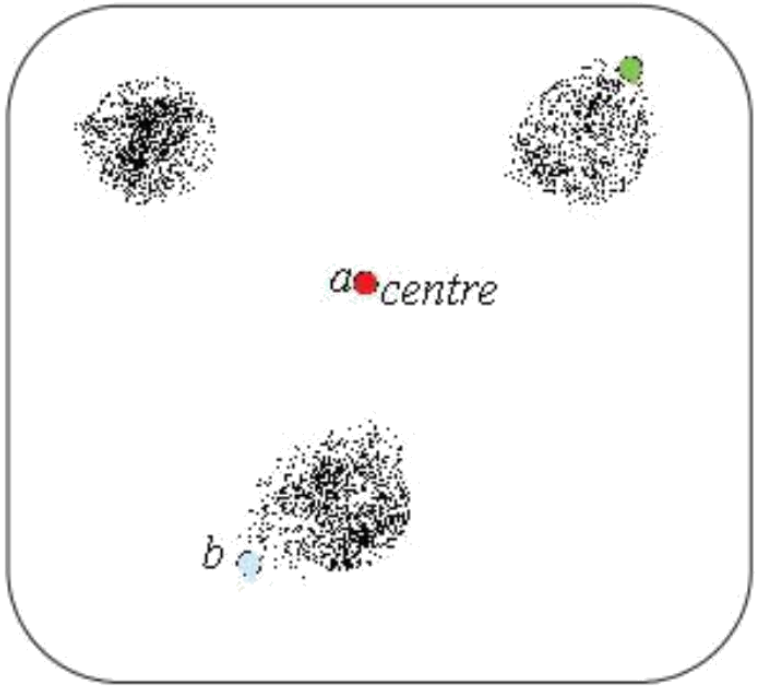


Figure II-1 classification par le principe d' initialisation par le mal classé

(a) Le centre des données en rouge,
(b) le bleu et le vert représentent les deux objets les plus éloigné

Les alternatives

L'approche incrémental :

Entrée

Ensemble de N données, notés par x ;

Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes $\{C_1, C_2, \ldots Ck\}$

Début

1) $c_1 = x_1$;

$c_2 = x_2$; avec $d(x_1, x_2) = \max_{i,j \in \{1,\ldots,N\}, i \neq j}(d(x_i - x_j))$

Répéter

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;

3) Trouver l'$i^{ème}$ centre Ci:

$$C_i = x : x = \max_{i \in [1,n]}(d^i_{k-1})$$

Avec d^i_{k-1} la distance entre x_i et son plus proche centre parmi les k-1 centre

4) Appliquer le k-means jusqu'à la convergence ;

Jusqu'à obtenir une partition en k groupes ;

Fin.

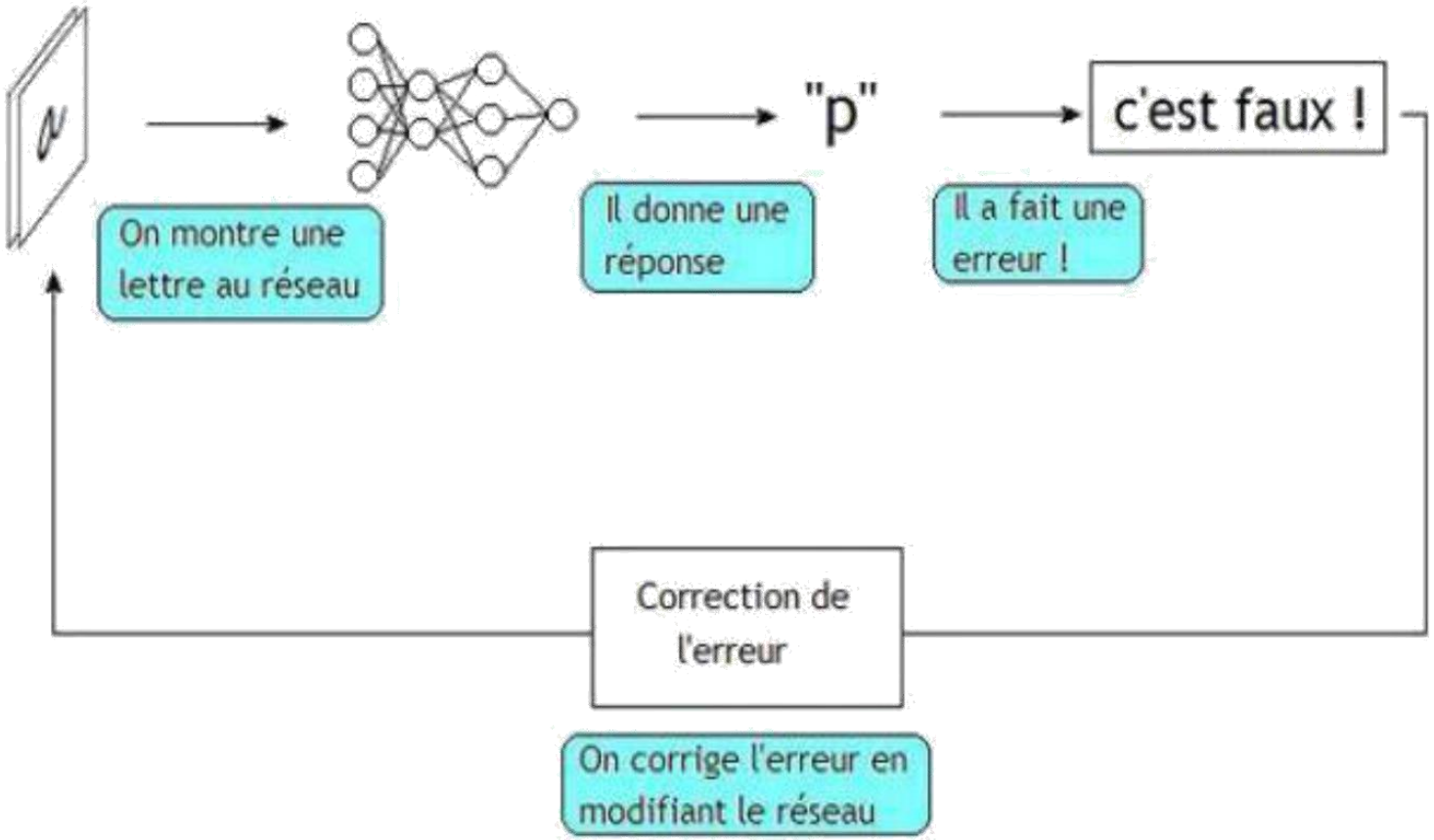
Hybridations

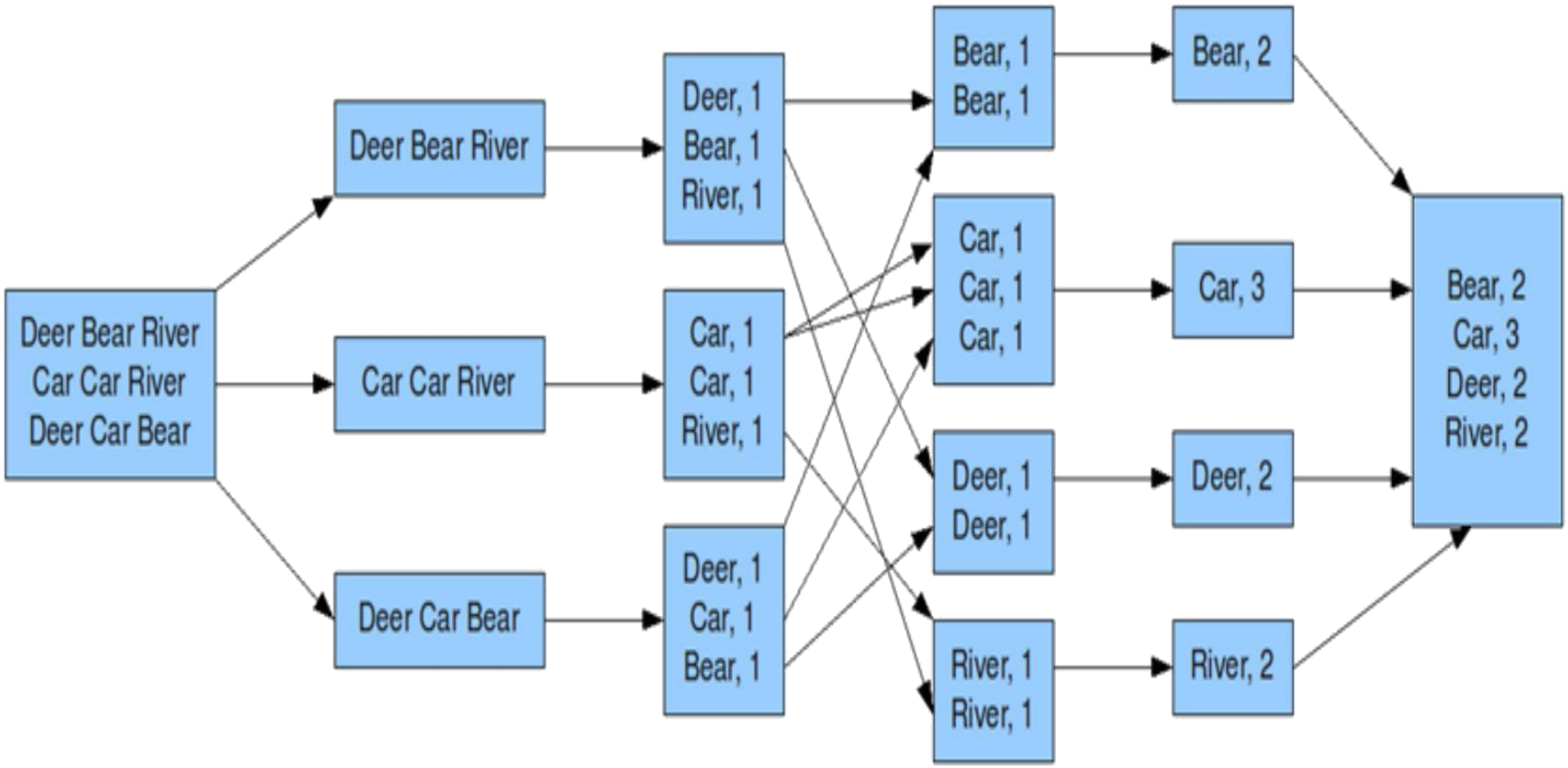
KMSVM : K-Means Support Vector Machine

- ✓ Amélioration du temps de réponse

KMKNN : K-Means for K-Nearest Neighbors

- ✓ Accélération des recherches des plus proches voisins dans des espaces de grande dimension


"p"
c'est faux !
On montre une lettre au réseau
Il donne une réponse
Il a fait une erreur !
Correction de l'erreur
On corrige l'erreur en modifiant le réseau


Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Merci de votre attention

Algorithme K-Moyennes

CAFÉ SCIENTIFIQUE