

Compte-rendu du Projet Apprentissage non supervisé

Méthode de programmation dynamique de Fisher
sous R

Achraf Benlemkaddem

12/11/2017

Implémentation de la méthode de programmation dynamique de Fisher sous R et son application pour classifier des données simulées et réelle en comparant avec d'autres méthodes de Clustering.

Contenu

Figures.....	1
Abstract.....	2
1) Implémentation de l'algorithme sous R.....	2
2) Application au données simulées	2
a) Méthode du coude.....	2
b) Résultat du pour 4 classes.....	3
c) Comparaison avec K-Means et CAH-Ward.	3
3) Application au données réelles.....	4
a) Résultat de clustering par Fisher	4
b) Comparaison avec les classes d'origines, K-Means et CAH-Wards.....	5
Conclusion générale.....	6

Figures

Figure 1: Application de la méthode du coude par rapport au critère C_k	2
Figure 2: Classification des données simulées par l'algorithme de Fisher.....	3
Figure 3: Comparaison entre les méthodes Fisher, K-Means et CAH-Ward appliquées au données simulées.	4
Figure 4: Classification des données Aiguillage par l'algorithme de Fisher.	5
Figure 5: Comparaison entre les méthodes Fisher, K-Means et CAH-Ward appliquées au données simulées et les classes d'origines.....	5

Abstract

L'objectif de ce projet est d'étudier une méthode de Clustering qui recherche des classes ordonnées dans le temps : la méthode de programmation dynamique de Fisher [1][2]. On procède dans une première partie à l'implémentation de l'algorithme sous R. Ensuite, on l'applique à deux exemples de jeux de données réels et simulés. Puis on compare les résultats à celui de d'autres algorithmes de Clustering.

1) Implémentation de l'algorithme sous R.

Vous trouverez l'implémentation de l'algorithme sous dans le fichier « **FisherDynamicProgramming.R** ».

2) Application au données simulées

Dans cette partie on applique l'algorithme de Fisher pour segmenter des données simulées « **sequencesimu** ».

La détermination du nombre de classes à utiliser se fait à travers la **méthode du coude**.

En fin on compare le résultat du Clustering par Fisher à celui donnée par d'autres méthodes comme KMeans et CAH-Ward.

a) Méthode du coude

On applique l'algorithme de Fisher au données pour plusieurs valeurs de k de 2 à 10 puis on plot le critère C_k pour chaque valeurs de k .

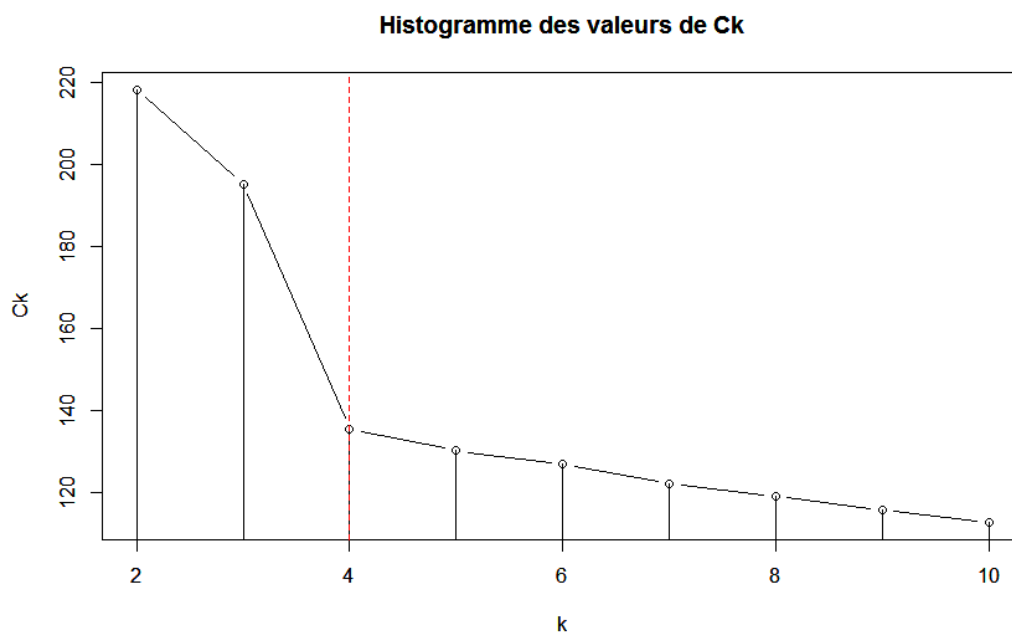


Figure 1: Application de la méthode du coude par rapport au critère C_k

On perçoit une chute forte de C_k juste avant l'indice 4. Donc on choisit $k=4$.

b) Résultat du pour 4 classes

On affiche le résultat de la classification pour $k=4$

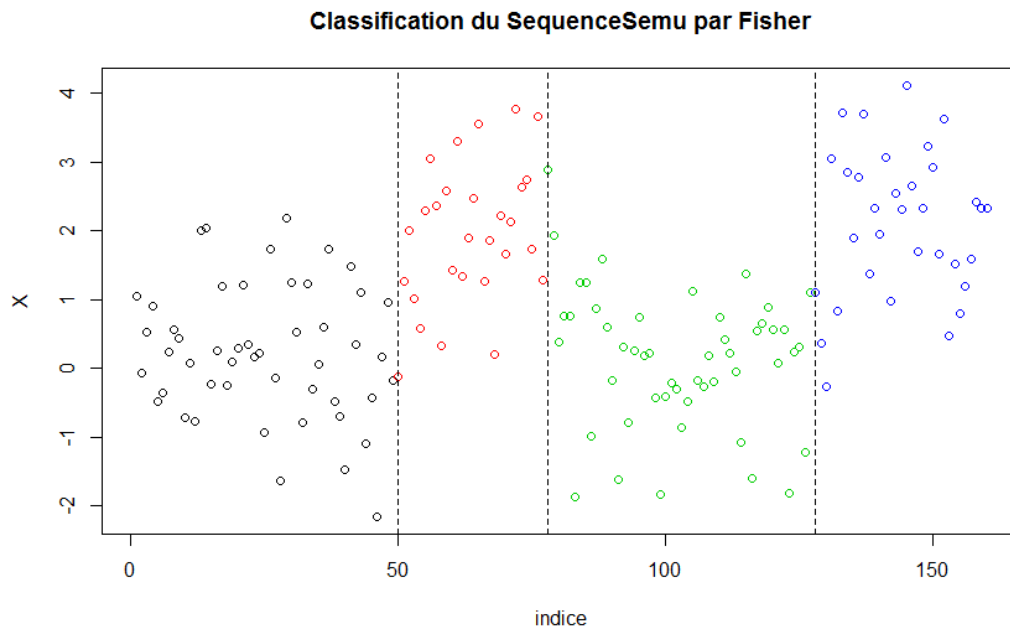


Figure 2: Classification des données simulées par l'algorithme de Fisher

c) Comparaison avec K-Means et CAH-Ward.

En comparant avec K-Means et CAH-Ward (figure 3) on conclue que Fisher donne une meilleur classification sur les données « **sequencesimu** ».

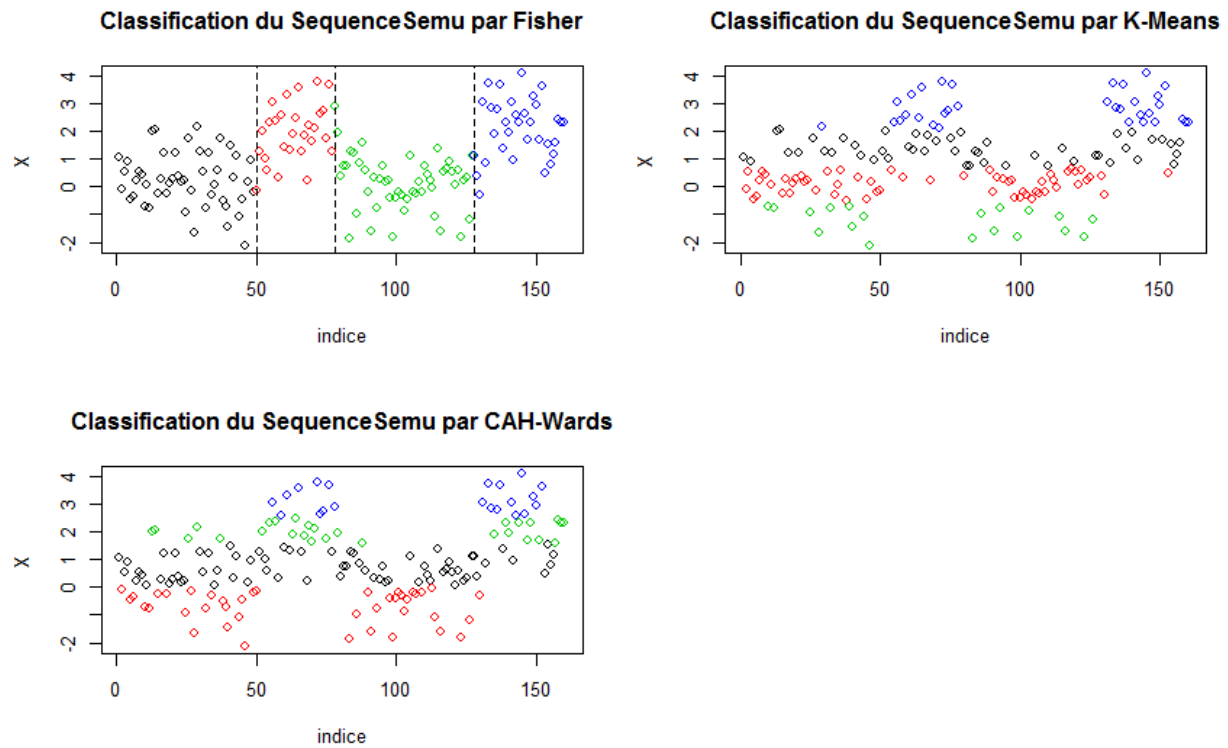


Figure 3: Comparaison entre les méthodes Fisher, K-Means et CAH-Ward appliquées au données simulées.

3) Application au données réelles

Dans cette partie on applique l'algorithme de Fisher à un jeu de données réelles nommé « **Aiguillage** » qui comprend 140 séries temporelles décrites par 553 variables: les 552 premières variables correspondent à l'énergie consommée au cours des mouvements mécaniques d'un système d'aiguillage ferroviaire et la 553e variable correspond à la classe (1 : sans défaut, 2 : défaut mineur, 3 : défaut critique, 4 : panne).

a) Résultat de clustering par Fisher

On remarque que l'application de l'algorithme de Fisher aux données « **Aiguillage** » (figure 4) ne résulte pas en un bon clustering.

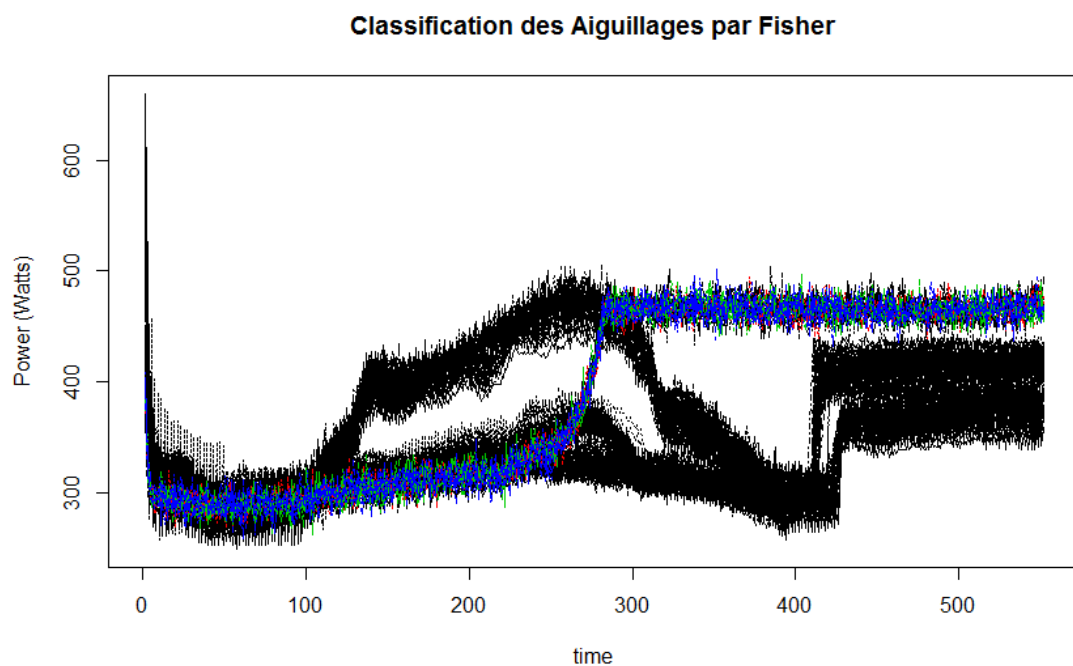


Figure 4: Classification des données Aiguillage par l'algorithme de Fisher.

b) Comparaison avec les classes d'origines, K-Means et CAH-Wards

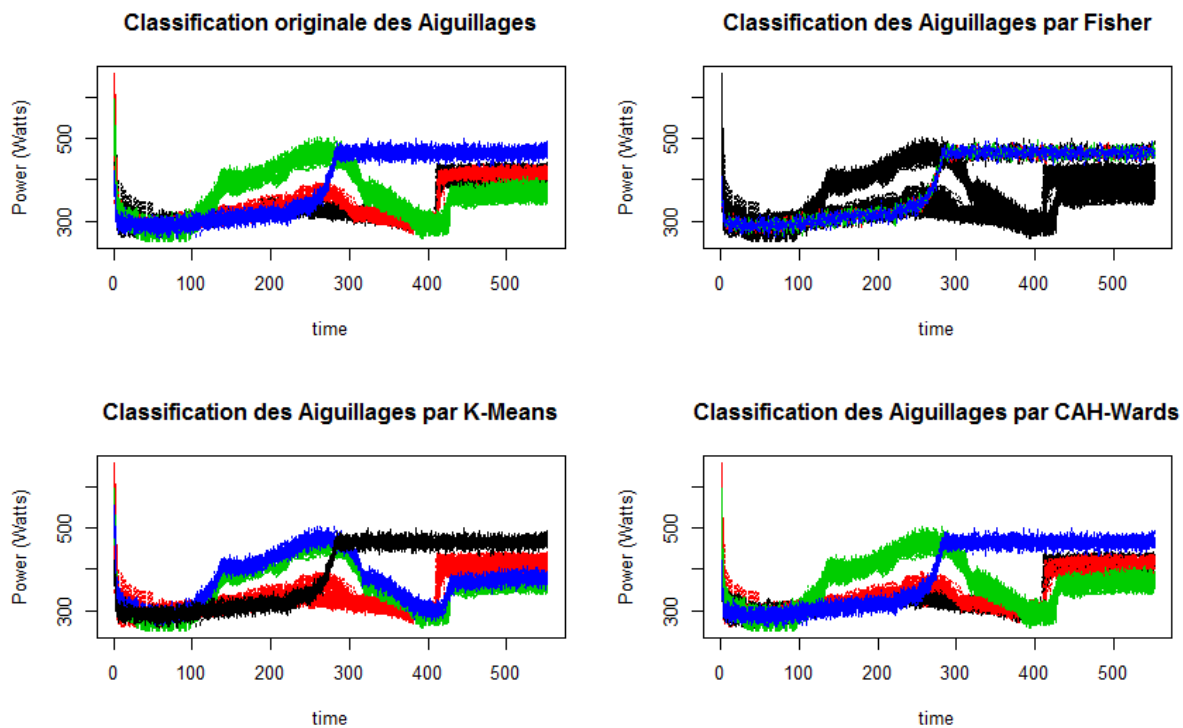


Figure 5: Comparaison entre les méthodes Fisher, K-Means et CAH-Ward appliquées au données simulées et les classes d'origines

Conclusion générale

La *méthode de programmation dynamique de Fisher* à données des résultats très performantes dans le cas des données simulées « **sequencesimu** » pendant qu'elle n'a pas bien performer pour les données « **Aiguillage** ».

L'échec de la *méthode de programmation dynamique de Fisher* dans le cas des données « **Aiguillage** » peut être expliquées partiellement par le volume explosif des variables présentes dans ce jeu de données.