TP2 - Comparaison des Algorithmes de Clustering : FCM vs K-Means

Thomas Lahely

12 novembre 2024

Résumé

Dans ce rapport, nous allons comparer les performances entre l'algorithme de clustering : Fuzzy C-Means (FCM) et K-Means sur un jeu de données que nous avons créé. Nous introduisons aussi l'indice de Dunn comme valeur pour évaluer la qualité des clusters obtenus. Nous explorons trois scénarios d'initialisation des centres pour analyser l'impact sur la convergence et la qualité des clusters.

1 Introduction

Le clustering est une technique d'apprentissage non supervisé qui vise à regrouper des données similaires. Parmi les algorithmes les plus utilisés, on retrouvele K-Means et le Fuzzy C-Means. K-Means attribue chaque point de données à un seul cluster, tandis que FCM permet une appartenance floue, où chaque point peut appartenir à plusieurs clusters avec un certain degré d'appartenance.

2 Description du Jeu de Données

Nous utilisons un jeu de données généré à l'aide de la fonction `make_blobs` de la bibliothèque `sklearn.datasets`. Ce jeu de données comporte 300 points répartis en trois clusters distincts avec les centres réels suivants :

— Centre 1 : (0, 0)
— Centre 2 : (5, 5)
— Centre 3 : (10, 0)

3 Méthodologie

3.1 Indice de Dunn

L'indice de Dunn est une métrique qui évalue la qualité du clustering en comparant le diamètre maximal des clusters à la distance minimale entre les clusters. Il est défini comme suit :

$$\text{Indice de Dunn} = \frac{\delta}{\Delta}$$

où δ est la distance minimale entre les clusters et Δ est le diamètre maximal des clusters. Un indice de Dunn élevé ou même maximal indique des clusters bien séparés et compacts.

3.2 Scénarios d'Initialisation

Nous avons testé trois scénarios d'initialisation des centres pour les deux algorithmes :

1. **Centres initiaux proches des centres réels** :
   — (0, 0)
   — (5, 5)
   — (10, 0)
2. **Centres initiaux éloignés des centres réels** :
   — (2, 2)
   — (6, 6)
   — (8, -2)
3. **Centres initiaux aléatoires** :

   Générés aléatoirement avec une graine fixe pour assurer la reproductibilité, les centres sont dans l'intervalle [-5, 15] pour chaque dimension et de facon uniforme pour garder la même probabilité pour chaque cluster.

3.3 K-Means

L'algorithme K-Means est implémenté à l'aide de la bibliothèque `scikit-learn`. Les paramètres principaux sont :

— Nombre de clusters (`n_clusters`) : 3
— Initialisation (`init`) : Centres initiaux selon le scénario
— Nombre d'initialisations (`n_init`) : 1
— Nombre maximal d'itérations (`max_iter`) : 10

3.4 Fuzzy C-Means (FCM)

L'algorithme FCM est implémenté à l'aide de la bibliothèque `scikit-fuzzy`. Une matrice de partition initiale est créée pour chaque scénario d'initialisation. Les paramètres principaux sont :

- Nos données transposées (`data`) : X.T
- Nombre de clusters (c) : 3
- Degré de flou (m) : 2
- Critère d'arrêt (`error`) : 0.005
- Nombre maximal d'itérations (`maxiter`) : 10
- Matrice de partition initiale (`init`) : u0

4 Résultats et Discussion

Les résultats des expériences menées avec les algorithmes K-Means et Fuzzy C-Means (FCM) sont présentés ci-dessous. Trois scénarios ont été étudiés, chacun caractérisé par une configuration différente des centres initiaux par rapport aux centres réels. L'indice de Dunn a été utilisé comme valeur principale pour évaluer la qualité des clusters obtenus.

4.1 Scénario 1 : Centres Initiaux Proches des Centres Réels

4.1.1 K-Means

Dans ce scénario, K-Means a démarré avec des centres initiaux proches des centres réels `[[0, 0], [5, 5], [10, 0]]`. L'algorithme a convergé rapidement après quelques itérations. Cependant, l'indice de Dunn obtenu est de **0.0083**, ce qui est relativement faible. Les centres finaux sont légèrement décalés par rapport aux centres réels, ce qui indique une certaine imprécision dans le clustering.

- **Indice de Dunn** : *0.0083 (Valeur faible)*

4.1.2 Fuzzy C-Means

FCM a également démarré avec les mêmes centres initiaux que K-Means. L'algorithme a convergé rapidement, produisant des centres finaux proches des centres réels avec un indice de Dunn de **0.0157**. Cette valeur est plus élevée que celle obtenue par K-Means, suggérant une meilleure qualité de clustering.

- **Indice de Dunn** : *0.0157 (Valeur élevée)*

4.1.3 Comparaison

Dans ce scénario, FCM surpasse K-Means en termes de qualité de clustering, comme en témoigne un indice de Dunn plus élevé. Bien que les centres initiaux soient proches des centres réels, FCM parvient à affiner davantage les clusters, offrant ainsi une meilleure séparation et compacité.

4.2 Scénario 2 : Centres Initiaux Éloignés des Centres Réels

4.2.1 K-Means

Avec des centres initiaux éloignés `[[2, 2], [6, 6], [8, -2]]`, K-Means a nécessité un nombre légèrement plus important d'itérations pour converger. L'indice de Dunn obtenu est de **0.0188**, indiquant une amélioration par rapport au scénario précédent, mais toujours inférieur à celui de FCM.

— **Indice de Dunn** : *0.0188 (Valeur moyenne)*

4.2.2 Fuzzy C-Means

FCM, avec les mêmes centres initiaux, a démontré une meilleure adaptabilité en convergeant vers des centres finaux plus proches des centres réels, obtenant un indice de Dunn de **0.0198**. Cette performance légèrement supérieure à celle de K-Means confirme la robustesse de FCM face à une initialisation moins favorable.

— **Indice de Dunn** : *0.0198 (Valeur élevée)*

4.2.3 Comparaison

Dans ce scénario, FCM continue de surpasser K-Means en termes de qualité de clustering. L'indice de Dunn supérieur de FCM indique une meilleure séparation et compacité des clusters, soulignant sa capacité à mieux gérer des initialisations défavorables par rapport à K-Means.

4.3 Scénario 3 : Centres Initiaux Aléatoires

4.3.1 K-Means

Lorsque les centres initiaux sont choisis de manière aléatoire `[[2.4908, 14.0143], [9.6399, 6.9732], [-1.8796, -1.8801]]`, K-Means a rencontré des difficultés pour converger vers les centres réels. L'indice de Dunn obtenu est de **0.0207**, ce qui est le plus élevé parmi les trois scénarios pour

K-Means, indiquant une bonne qualité de clustering malgré l'initialisation aléatoire.

— **Indice de Dunn** : *0.0207 (Valeur élevée)*

4.3.2 Fuzzy C-Means

FCM, avec les mêmes centres initiaux aléatoires, a obtenu un indice de Dunn de **0.0146**. Bien que ce score soit inférieur à celui de K-Means dans ce scénario particulier, FCM a tout de même produit des clusters de qualité acceptable.

— **Indice de Dunn** : *0.0146 (Valeur moyenne)*

4.3.3 Comparaison

Contrairement aux deux premiers scénarios, dans ce cas spécifique, K-Means a surpassé FCM en termes d'indice de Dunn. Cela suggère que, dans certaines configurations d'initialisation aléatoire, K-Means peut parfois obtenir de meilleurs résultats. Toutefois, cette performance n'est pas systématique, et FCM reste globalement plus robuste face à diverses initialisations.

5 Conclusion

La comparaison entre les algorithmes Fuzzy C-Means et K-Means révèle des différences notables en fonction des configurations des centres initiaux. Globalement, FCM démontre une meilleure robustesse et une qualité de clustering supérieure, particulièrement lorsque les centres initiaux sont proches ou légèrement éloignés des centres réels. Cette performance est attribuable à la capacité de FCM à gérer des degrés d'appartenance flous, ce qui permet une meilleure adaptation aux données.

Cependant, dans le scénario d'initialisation aléatoire, K-Means a montré une performance légèrement supérieure, bien que cela ne soit pas généralisable. Cela souligne que, bien que FCM soit généralement plus robuste, il existe des cas où K-Means peut tout de même offrir des résultats compétitifs.

L'utilisation de l'indice de Dunn a permis une évaluation quantitative précise de la qualité des clusters, mettant en évidence la supériorité de FCM dans la plupart des scénarios. En conclusion, FCM est recommandé pour des applications nécessitant une robustesse face à différentes initialisations, tandis que K-Means peut être utilisé lorsque la rapidité et la simplicité sont prioritaires, surtout dans des contextes où les centres initiaux sont bien choisis.