

الگوریتم Fuzzy c-means و کاربرد آن در پردازش تصاویر پیاده سازی، چالش ها و مقایسه با روش های دیگر

علیرضا عظیمی

خوشه بندی Clustering

خوشه بندی clustering یک روش یادگیری ماشین بدون ناظارت است که داده ها بر اساس شباهت هایشان به گروه هایی (خوشه هایی) تقسیم می کند. بدون اینکه از قبل برچسب یا دسته بندی مشخصی صورت گرفته باشد، هدف یافتن الگو یا ساختار پنهان در داده هاست.

چرا از خوشه بندی استفاده می کنیم؟

- کشف الگوهای پنهان
- تحلیل داده ها
- کاهش پیچیدگی

انواع روش های خوش بندی

۱. خوش بندی مبتنی بر مرکز centroid-based

در این نوع خوش بندی داده به k خوش تقسیم شده، که هر خوش دارای یک مرکز می باشد.

مثل الگوریتم k-means و Fuzzy c-means

کاربرد: ساده و سریع برای داده های عددی.

۲. خوش بندی سلسله مراتبی hierarchical

در این نوع خوش بندی، خوش ها به صورت درختی از پایین به بالا یا از بالا به پایین ساخته می شوند.

کاربرد مناسب برای داده های کوچک و نمایش روابط سلسله مراتبی

۳. خوش بندی مبتنی بر چگالی Density-based

خوش بندی بر اساس نقاط متراکم داده ها شکل می گیرد، مثل الگوریتم DB-scan

کاربرد: شناسایی خوش هایی با شکل های غیر منظم و حذف نویز

۴. خوش بندی مبتنی بر توزیع

در این نوع خوش بندی فرض می شود داده ها از توزیع های آماری خاص تولید شده اند مثل الگوریتم GMM (مدل مخلوط گاوی)

کاربرد: داده پیچیده با توزیع های آماری

۵. خوش بندی مبتنی بر گراف

داده ها به صورت گراف مدل شده و خوش ها بر اساس اتصالات گرافی شناسایی می شوند.

کاربرد: شبکه های اجتماعی

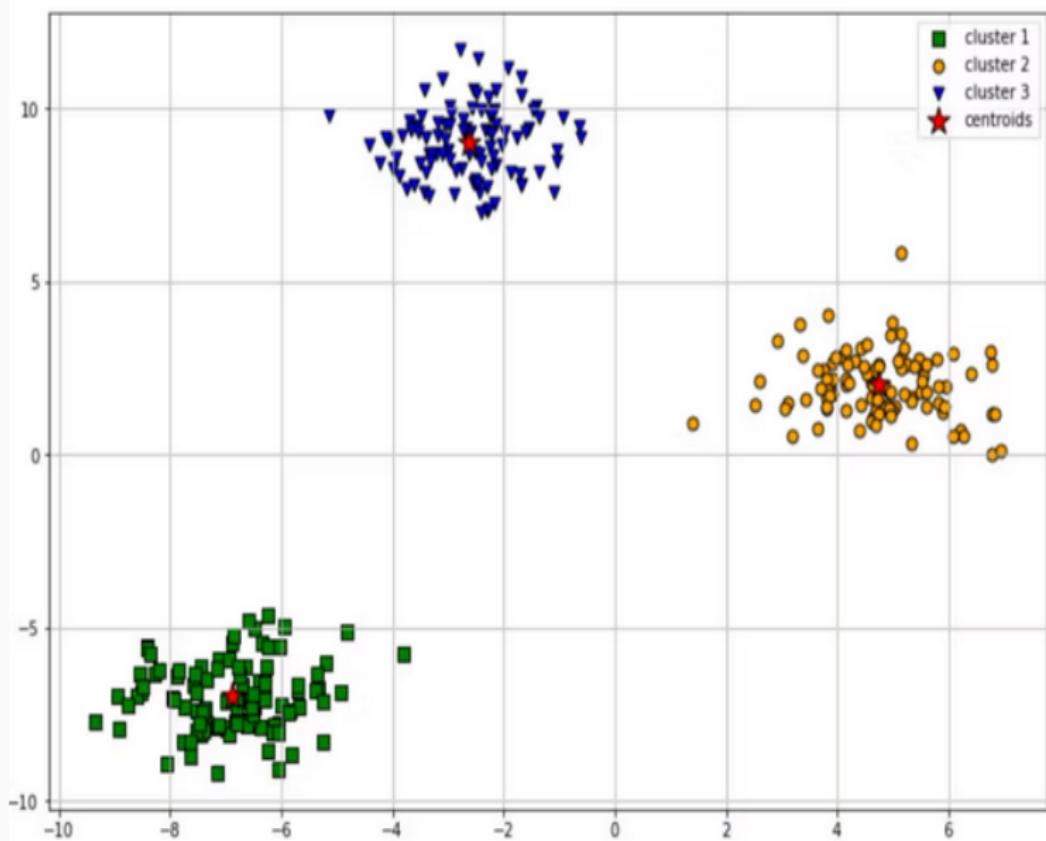
	Centroid-based	Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
	Connectivity-based	Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
	Density-based	Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
	Graph-based	Cluster points based on graph distance	Affinity Propagation Spectral Clustering
	Distribution-based	Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models (GMMs)

K-means Algorithm

الگوریتم k-means برای اولین بار در سال ۱۹۶۷ توسط James Mac Queen به کار گرفته شد.

ایده اصلی این الگوریتم از پردازش سیگنال گرفته شده و هدف آن تقسیم n نقطه (داده) به k خوشه است. در این الگوریتم هر داده متعلق به خوشه‌ای است که داده نزدیک ترین فاصله را با مرکز خوشه داشته باشد.

k-means Algorithm



مزایا و معایب الگوریتم k-means

- مزایا

سریع ترین الگوریتم خوش بندی، کار آمد برای داده هایی با ساختار مشخص.
این الگوریتم در مقیاس بزرگ خوب عمل می کند.

- معایب

به شدت حساس به نویز، انتخاب مراکز اولیه و تعداد خوش ها(K) تاثیر زیادی بر اجرا و نحوه عملکرد الگوریتم دارد و برای خوش های غیر کروی مناسب نیست.

- محدودیت ها

انتساب قطعی نقاط و hard clustering بودن.

حساسیت به نویز و داده های پرت
فرض خوش های متمایز و کروی

Fuzzy c-means

در سال ۱۹۷۳ الگوریتم جدیدی برای خوشه بندی توسط J.C.dunn معرفی شد که هدف آن رفع محدودیت های k-means بود.

الگوریتم fuzzy c-means یک روش خوشه بندی نرم می باشد که بر خلاف k-means که هر داده دقیقا به یک خوشه اختصاص داده می شد، این الگوریتم به هر داده اجازه می شود که با درجات عضویت مختلف (membership) به چندین خوشه متعلق باشد.

1. Initialize Parameters

Specify the number of clusters (c) and the fuzziness parameter(m)

Randomly initialize cluster centers (v_j , for $j = 1, \dots, c$)

Initialize the membership matrix u_{ij} for each data point x_i to the cluster j , satisfying:

$$\sum_{j=1}^c u_{ij} = 1, \quad 0 \leq u_{ij} \leq 1$$

2. Compute Cluster Centers

Update cluster centers (v_j) as the weighted mean of the data points, based on membership degrees.

$$v_j = \frac{\sum_{i=1}^n (u_{ij})^m x_i}{\sum_{i=1}^n (u_{ij})^m}, \forall j = 1, \dots, c$$

where x_i is the data point, u_{ij} is the membership of point i in cluster j and j is the number of clusters

3. Updates Membership Degrees

update membership degrees (u_{ij}) based on Euclidean distance between date points and cluster centers.

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\left(\frac{\|x_i - v_j\|^2}{\|x_i - v_k\|^2} \right) \right)^{\frac{2}{m-1}}}$$

where m is the fuzzifier (fuzzy parameter)

4. Compute The Objective Function

calculate the objective function to minimize:

$$J_m = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m ||x_i - v_j||^2$$

5. The algorithm repeats until one of these conditions reached:

- $\epsilon > \max_{ij} \{ |u_{ij}^{(k+1)} - u_{ij}^{(k)}| \}$ where k is the iteration step.
- maximum iteration reached.

Fuzzy c-means

```
import numpy as np

def fuzzy_c_means(X, n_clusters, m=2, max_iters=100, tol=4-1e):

    n_samples, n_features = X.shape

    U = np.random.rand(n_samples, n_clusters)
    U = U / np.sum(U, axis=1, keepdims=True)

    for iteration in range(max_iters):
        U_old = U.copy()

        Um = U ** m
        centers = np.dot(Um.T, X) / np.sum(Um, axis=0)[:, np.newaxis]

        distances = np.zeros((n_samples, n_clusters))
        for j in range(n_clusters):
            distances[:, j] = np.sum((X - centers[j]) ** 2, axis=1)

        U = np.zeros((n_samples, n_clusters))
        for i in range(n_samples):
            for j in range(n_clusters):
                if distances[i, j] == 0:
                    U[i, j] = 1
                else:
                    U[i, j] = 1 / np.sum((distances[i, j] / distances[i, :]) ** (2 / (m - 1)))

            if np.max(np.abs(U - U_old)) < tol:
                break

    labels = np.argmax(U, axis=1)
    return centers, U, labels
```

مزایا و معایب این الگوریتم

• مزایا

- انعطاف پذیری در خوش بندی
- دقیق بالاتر در داده های پیچیده
- کاربردهای گسترده در حوزه های پردازش تصویر، تحلیل داده های زیستی و تشخیص الگو بیمار

• معایب

- پیچیدگی محاسباتی
- حساسیت به مقدار m
- حساسیت به مقدار دهی اولیه
- نیاز به تعیین تعداد خوش ها قبل از شروع کار

مفهوم فازی در الگوریتم fuzzy c-means

فازی بودن در FCM به درجات عضویت U_{ij} اشاره دارد بخلاف الگوریتم k-means (عضویت ۰ یا ۱) به داده ها اجزه می دهد با وزن های بین ۰ تا ۱ به چندین خوش تعلق داشته باشد. این عضویت در ماتریس U_{ij} ذخیره شده و نشان دهنده میزان تعلق هر نقطه به هر خوش است.

الگوریتم فازی مورد استفاده :

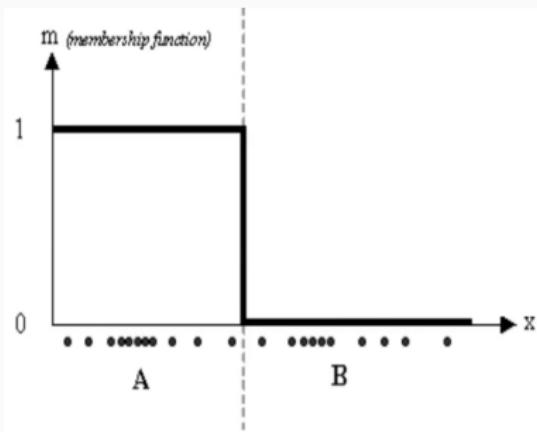
FCM ازتابع عضویت فازی استاندارد استفاده می کند که مبتنی بر فاصله اقلیدسی وزن دار است، نه توابع فازی خاص مثل (مثاشی، ذوزنقه ای و ...)

$$U_{ij} = \frac{1}{\sum_{k=1}^c \left(\left(\frac{\|x_i - v_j\|^2}{\|x_i - v_k\|^2} \right) \right)^{\frac{2}{m-1}}}$$

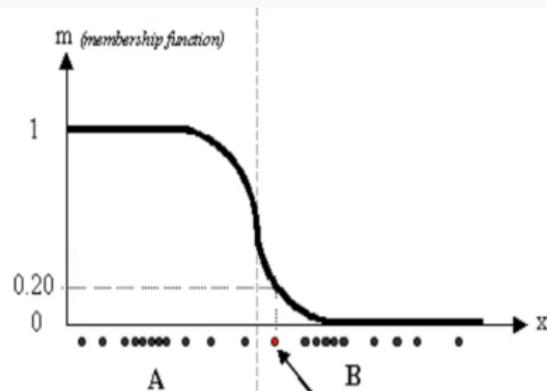
این فرمول عضویت را بر اساس نسبت معکوس فاصله های وزن دار تعیین می کند که یک رویکرد فازی غیر خطی است.

پارامتر m مسئولیت کنترل میزان فازی بودن خوش بندی را بر عهده دارد ($m > 1$)

در واقع اگر m کم باشد، خوش بندی سخت تر می شود و با افزایش پارامتر m خوش بندی به سمت نرم شدن میل می کند.

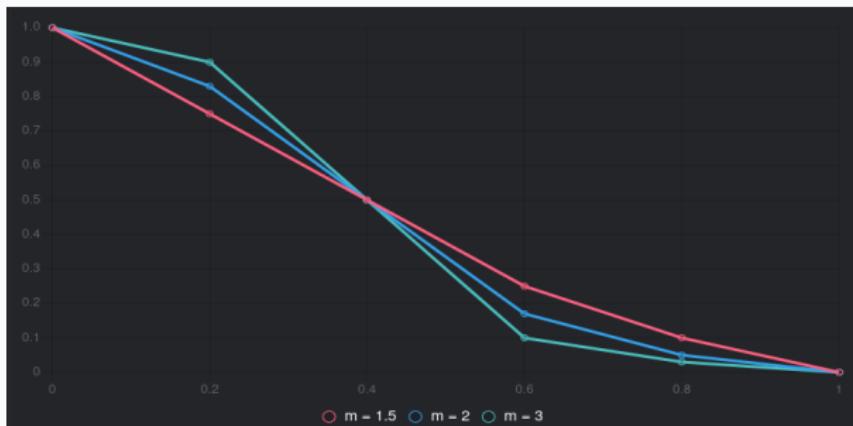


$$U_{cav} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ \dots & \dots \\ 0 & 1 \end{bmatrix}$$



$$U_{cav} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ \dots & \dots \\ 0.9 & 0.1 \end{bmatrix}$$

Fuzzy c-means



Appilcation of Fuzzy c-means

این الگوریتم به دلیل انعطاف پذیری در خوش بندی نرم، در حوزه های مختلفی کاربرد دارد:

۱. پردازش تصویر

- تقسیم بندی تصاویر مثل جداسازی بافت ها در تصاویر پزشکی MRI
- کاهش نویز یا فشردگی تصاویر

۲. تحلیل داده های زیستی

- خوش بندی ژن ها یا پروتئین ها بر اساس الگو.
- تشخیص الگوهای بیماری در داده های پزشکی

۳. بازاریابی و تحلیل داده

- گروه بندی مشتریان بر اساس رفتار خرید با درجات عضویت مختلف
- تحلیل داده های شبکه اجتماعی برای شناسایی جوامع

۴. هوش مصنوعی و یادگیری ماشین

- پیش پردازش داده ها برای بهبود عملکرد مدل های یادگیری ماشین
- ترکیب با الگوریتم های دیگر مثل شبکه های عصبی

FCM خوشه بندی تصاویر با الگوریتم

۱. خواندن تصاویر :

تصویر خاکستری یا رنگی با استفاده از کتابخانه های PIL OpenCV یا خوانده می شود.

$$I(x, y) \rightarrow \text{Intensity or RGB}$$

$$\text{Grayscale} : I \in R^{H \times W}, \text{RGB} : I \in R^{H \times W \times 3}$$

۲. تبدیل به بردار ویژگی :

هر پیکسل به یک بردار ویژگی تبدیل می شود.

تصویر با ابعاد $W \times H$ (ارتفاع و عرض) به یک ماتریس $W \times H = N \times N$ تعداد پیکسل ها و d تعداد ویژگی هاست.

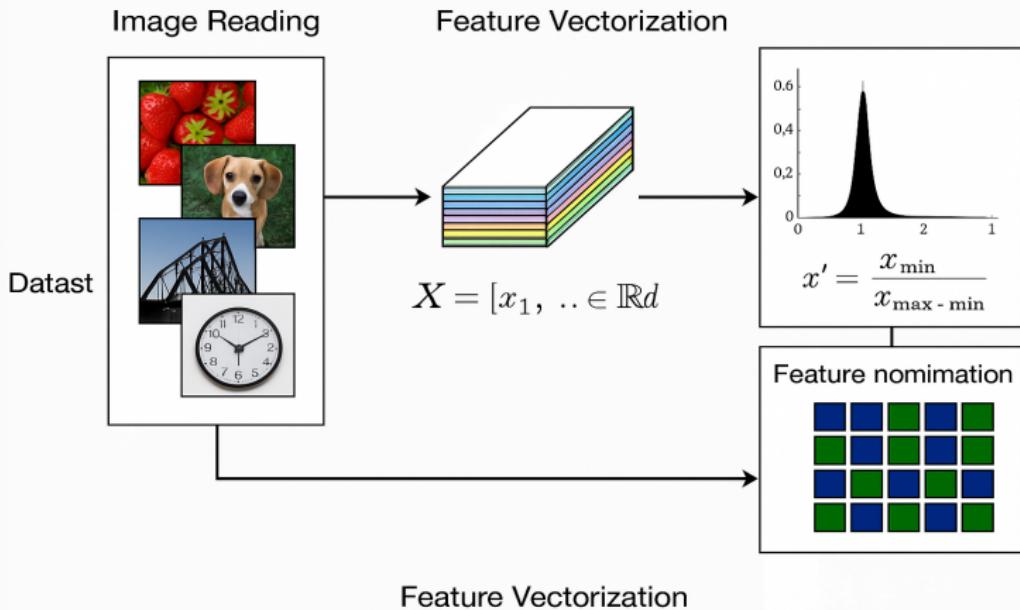
$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \quad x_i \in R^d$$

$$x_i = [R, G, B, x, y]$$

۳. نرمال سازی داده ها:

مقادیر ویژگی ها باید نرمال سازی شود (مثلا در بازه $[0, +\infty]$) تا فاصله اقلیدسی در FCM معنا دار شود.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$



معیار های ارزیابی خوش بندی

Silhoutte Score •

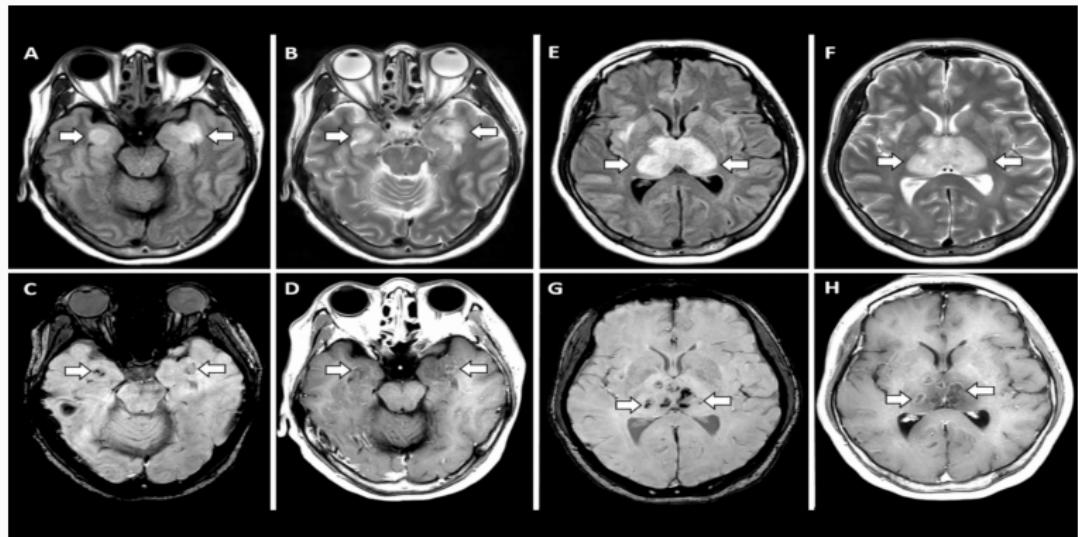
میانگین تفاوت بین فاصله درون خوشه‌ای و فاصله به نزدیک‌ترین خوشه دیگر برای هر نمونه را اندازه‌گیری می‌کند. مقدار آن بین ۱- تا ۱ است؛ عدد بالاتر نشان دهنده‌ی تفکیک بهتر خوشه‌هاست.

Devies-Bouldin Index •

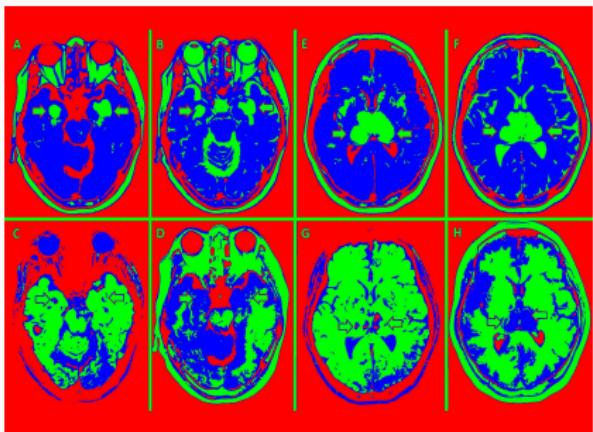
نسبت میانگین پراکندگی در خوشه‌ها به فاصله بین خوشه‌ها را محاسبه می‌کند. مقدار کمتر نشان دهنده خوشبندی بهتر است.

Calinski-Harabasz Index •

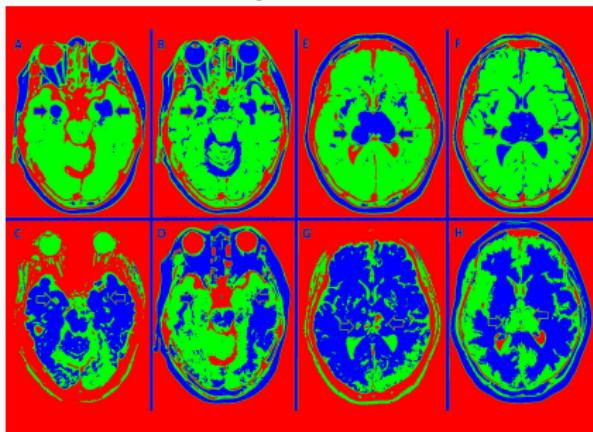
نسبت بین پراکندگی بین خوشه‌ای به پراکندگی درون خوشه‌ای را می‌سنجد. مقدار بیشتر به معنای جداسازی و فشردگی بهتر خوشه‌هاست.



k-means



fuzzy c-means

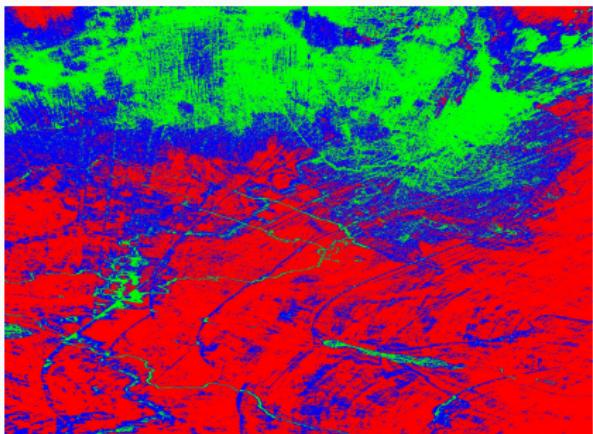


◆ K-Means Evaluation:
Silhouette Score: 0.7158
Davies-Bouldin Index: 0.4644
Calinski-Harabasz Index: 347153.9124

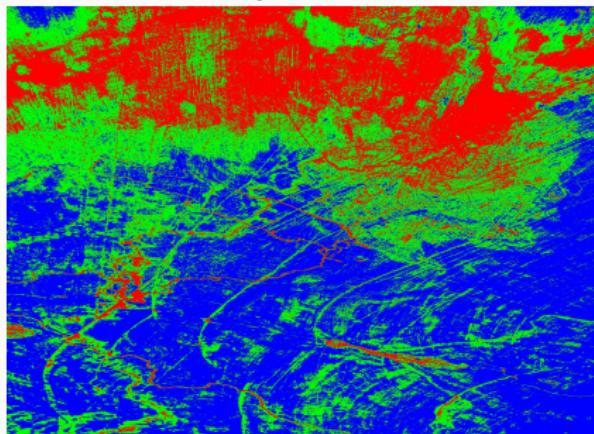
◆ Fuzzy C-Means Evaluation:
Silhouette Score: 0.7128
Davies-Bouldin Index: 0.4679
Calinski-Harabasz Index: 343757.3600



k-means



fuzzy c-means

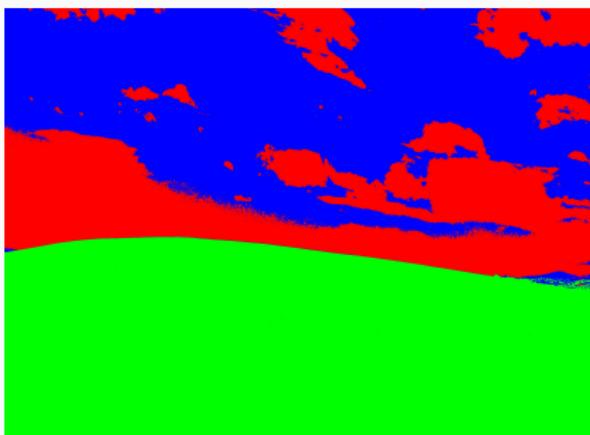


◆ K-Means Evaluation:
Silhouette Score: 0.5766
Davies-Bouldin Index: 0.5301
Calinski-Harabasz Index: 167662.6259

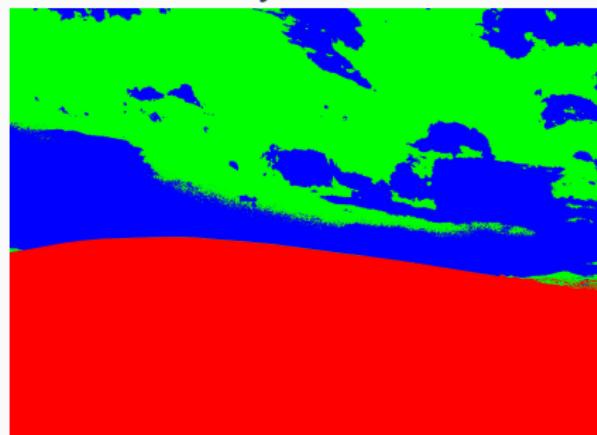
◆ Fuzzy C-Means Evaluation:
Silhouette Score: 0.5740
Davies-Bouldin Index: 0.5331
Calinski-Harabasz Index: 166836.0498



k-means



fuzzy c-means



◆ K-Means Evaluation:
Silhouette Score: 0.5699
Davies-Bouldin Index: 0.5270
Calinski-Harabasz Index: 140992.4491

◆ Fuzzy C-Means Evaluation:
Silhouette Score: 0.5718
Davies-Bouldin Index: 0.5236
Calinski-Harabasz Index: 141593.5424

۱. هر دو الگوریتم بیان شده نقاطی را به صورت رندوم برای تعیین محل اولیه خوشه ها(مراکز خوشه ها) استفاده می کنن. آیا راهی برای بهبود مقدار دهی اولیه مراکز خوشه ها وجود دارد؟
۲. آیا این الگوریتم کاملاً یک الگوریتم بدون نظارت هستش یا نه؟
۳. آیا راهی وجود دارد که بتوانیم با توجه به حجم داده ای که داریم پارامتر فازی ساز m مناسب رو پیدا کنیم و آیا این مقدار می تواند به صورت اتوماتیک با توجه به حجم داده تنظیم شود؟
۴. چرا هر دو الگوریتم بیان شده در حال حاضر منسخ شده اند؟ چه الگوریتم هایی در حال حاضر برای خوشه بندی استفاده می شوند.

Хорошо! Поговорим

Хорошо что поговорим. Хорошо что поговорим чтобы сказать что
хорошо что мы поговорим чтобы сказать что мы поговорим чтобы
мы. Хорошо что поговорим чтобы сказать что мы поговорим чтобы
мы поговорим чтобы сказать что мы (Хорошо что поговорим чтобы сказать
что мы поговорим чтобы сказать что мы поговорим чтобы сказать что мы
поговорим чтобы сказать что мы поговорим чтобы сказать что мы поговорим чтобы
мы)

Хорошо