Chỉ số Đánh giá Phân cụm Bayesian (BCVI)

Phạm Tường Duy, Trần Tuấn Đạt

Giảng viên hướng dẫn: TS. Tô Đức Khánh
Trường Đại học Khoa học Tự nhiên

Ngày 14 tháng 8 năm 2025

Mục lục

Giới thiệu

- **Vấn đề:**
  - Xác định số lượng cụm tối ưu trong phân cụm dữ liệu là một thách thức lớn.
  - Các chỉ số truyền thống (CVI) thiếu linh hoạt và không tích hợp được kinh nghiệm người dùng.
- **Nội dung:**
  - Phát triển Bayesian Cluster Validity Index (BCVI) dựa trên phân phối Dirichlet.
  - Tích hợp dữ liệu thực tế và kiến thức tiên nghiệm, nâng cao độ chính xác.
- **Ứng dụng:**
  - **Y tế**: Phát hiện khối u não từ ảnh MRI.
  - **Dữ liệu lớn**: Phân tích dữ liệu quy mô lớn.

Phương pháp nghiên cứu

Mục tiêu:

- Phát triển **Bayesian Cluster Validity Index (BCVI)** kết hợp dữ liệu thực tế và kiến thức tiên nghiệm thông qua phân phối **Dirichlet** và **Dirichlet tổng quát**.

Thuật toán sử dụng:

- **K-Means**: Xác định cụm bằng cách tối ưu hóa tổng bình phương khoảng cách trong cụm.
- **Fuzzy C-Means (FCM)**: Phân cụm mềm, xác định mức độ thành viên của mỗi điểm dữ liệu trong các cụm.

Phương pháp nghiên cứu

2. Các chỉ số đánh giá (CVI)

CVI truyền thống:

- **Hard clustering**: Calinski-Harabasz, Davies-Bouldin, Silhouette,...
- **Soft clustering**: Xie-Beni, KWON2.

BCVI:

- Tích hợp Bayesian framework với các CVI truyền thống.
- Sử dụng kiến thức tiên nghiệm để cải thiện khả năng phát hiện số cụm tối ưu.

Thuật toán K-means

3. Công thức tổng quát

Mục tiêu của thuật toán K-Means là tối ưu hóa tổng bình phương khoảng cách trong cụm:

$$J = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - v_j\|^2$$

- k: Số cụm.
- C_j: Tập các điểm dữ liệu trong cụm thứ j.
- v_j: Trọng tâm của cụm thứ j.
- $\|x - v_j\|$: Khoảng cách Euclidean giữa điểm dữ liệu x và trọng tâm v_j.

Thuật toán Fuzzy C-Means (FCM)

Hàm mục tiêu:

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

Trong đó:

- u_{ij}: Mức độ thành viên của điểm dữ liệu x_i trong cụm j.
- $m > 1$: Tham số mờ, kiểm soát mức độ mờ của phân cụm.
- c: Số cụm.
- v_j: Trọng tâm của cụm j.

Phân phối Dirichlet

Hàm mật độ xác suất:

$$f(x_1, \ldots, x_K|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}, \quad 0 \leq x_k \leq 1, \quad \sum_{k=1}^{K} x_k = 1$$

Trong đó:

- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ với $\alpha_k > 0$: Các tham số của phân phối Dirichlet.

Hàm Beta tổng quát:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

Chỉ số $r_k(x)$

Tỉ lệ điều chỉnh từ chỉ số CVI

$$r_k(\mathrm{x}) = \begin{cases} \dfrac{GI(k) - \min_j GI(j)}{\sum_{i=2}^{K}\left(GI(i) - \min_j GI(j)\right)} & \text{for Condition A,} \\ \dfrac{\max_j GI(j) - GI(k)}{\sum_{i=2}^{K}\left(\max_j GI(j) - GI(i)\right)} & \text{for Condition B.} \end{cases}$$

Trong đó:

- GI(k): là giá trị của CVI tương ứng với k cụm.
- $min_j GI(j)$ và $max_j GI(j)$ lần lượt là các giá trị nhỏ nhất và lớn nhất trong các giá trị CVI mà ta đang xét.

Phương pháp đánh giá BCVI

Công thức tổng quát:

$$BCVI(k) = \frac{\alpha_k + nr_k(x)}{\alpha_0 + n}$$

Trong đó:

- α_k: Tham số tiên nghiệm của cụm k.
- n: Tổng số điểm dữ liệu.
- $r_k(x)$: Giá trị liên quan đến dữ liệu thực tế (ví dụ: số lượng điểm dữ liệu trong cụm k).
- $\alpha_0 = \sum_{k=1}^{K} \alpha_k$: Tổng các tham số tiên nghiệm.

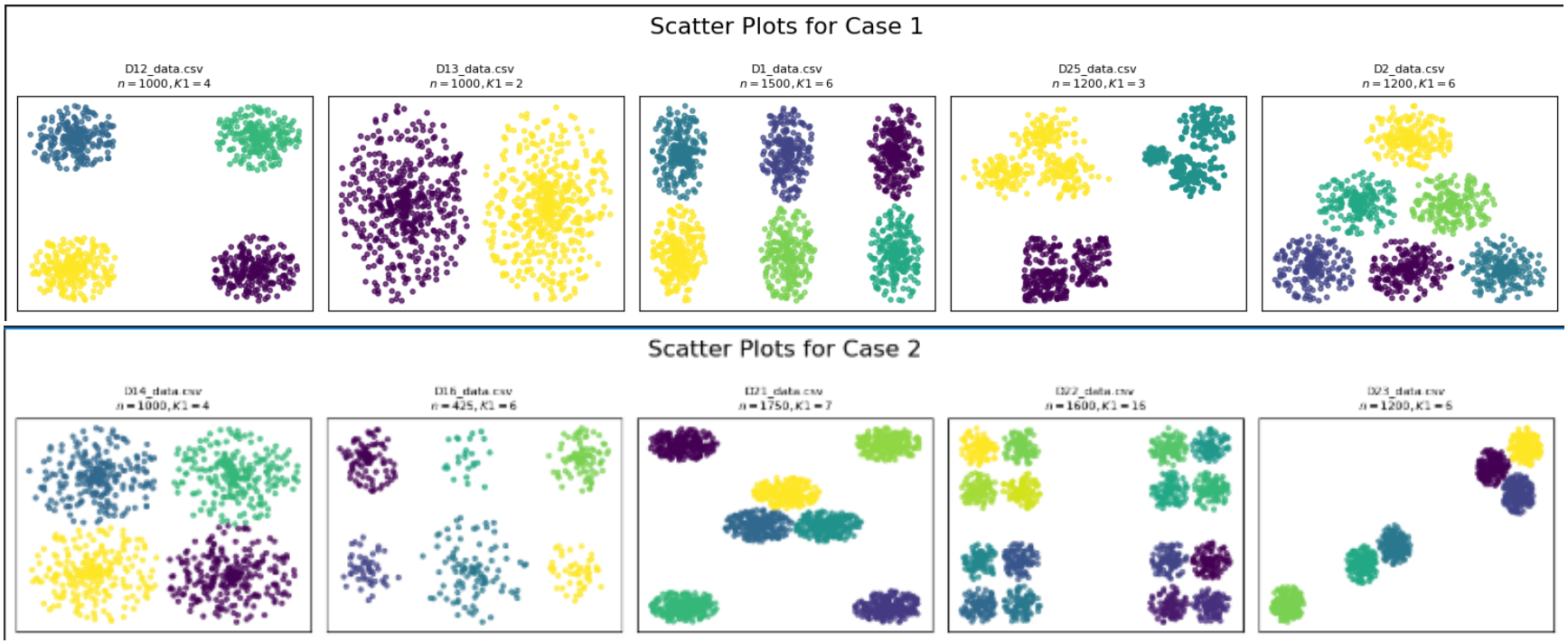
Bộ dữ liệu nhân tạo

Dữ liệu nhân tạo

- **Dữ liệu:** Gồm 25 tập dữ liệu được đánh số từ D1 đến D25.
- **Kích thước:** Đa dạng các kích thước và hình dạng.
- **Đặc trưng:** Chia thành 5 trường hợp.

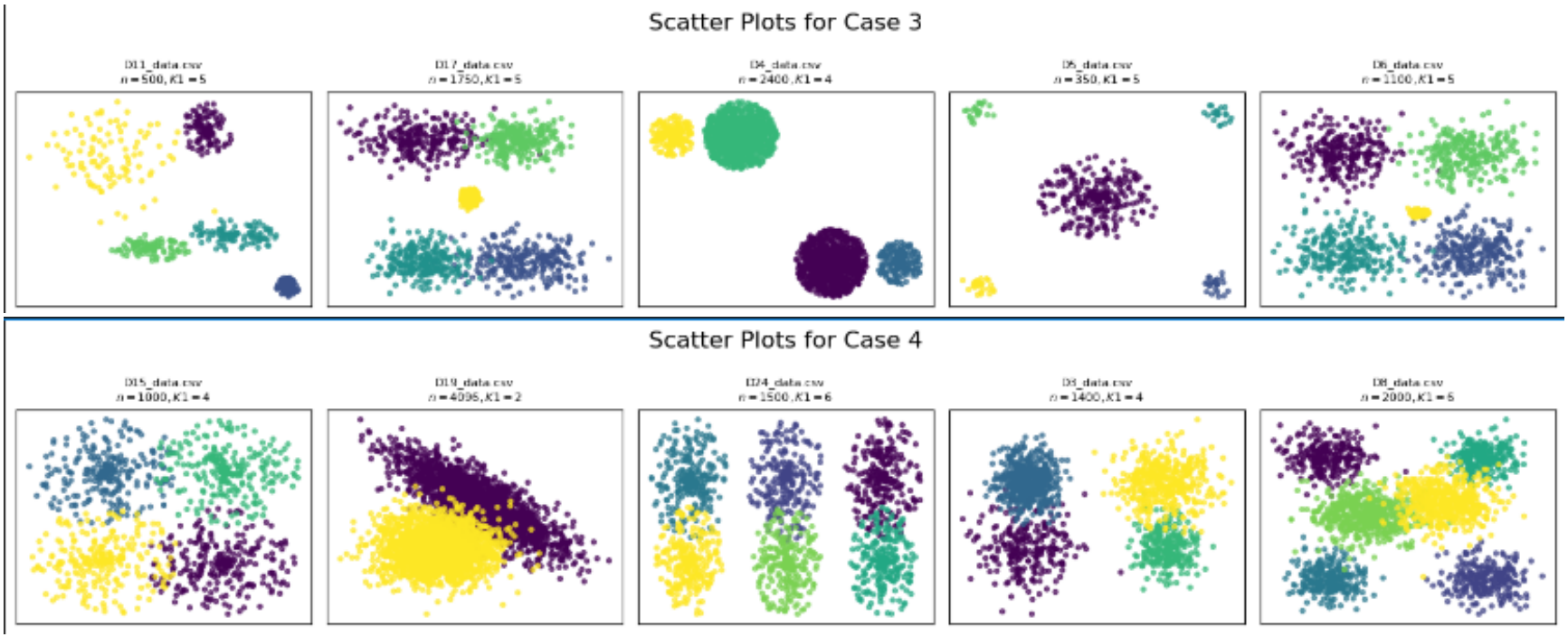
Bộ dữ liệu nhân tạo

Dữ liệu được đánh dấu bằng nhãn



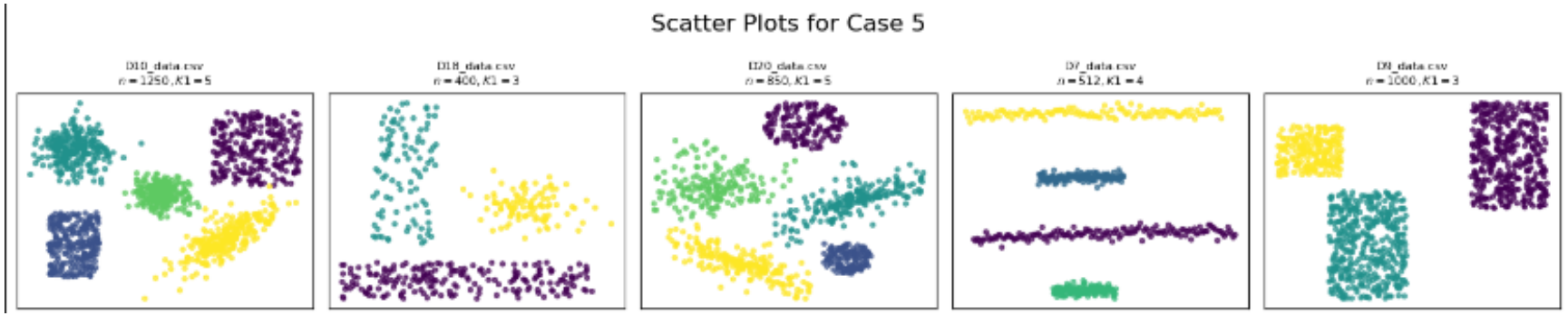
Bộ dữ liệu nhân tạo

Dữ liệu được đánh dấu bằng nhãn



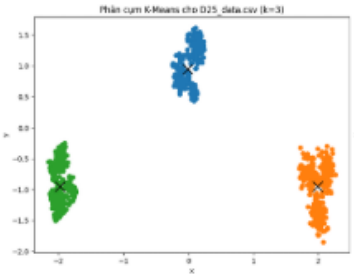
Bộ dữ liệu nhân tạo

Dữ liệu được đánh dấu bằng nhãn

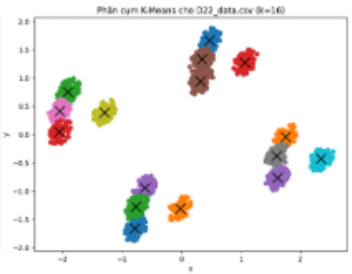


Bộ dữ liệu nhân tạo

Dữ liệu phân cụm bằng K-means

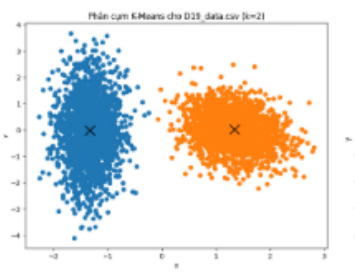


Case 1: D25

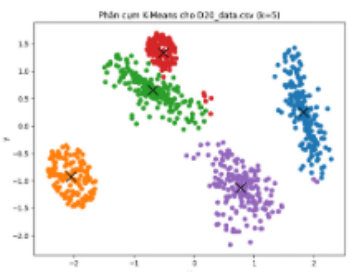


Case 2: D22

Case 3: D6



Case 4: D19



Case 5: D20

Bộ dữ liệu nhân tạo

Dữ liệu phân cụm bằng Fuzzy C-means

Case 1: D25 | Case 2: D22 | Case 3: D6 | Case 4: D19 | Case 5: D20

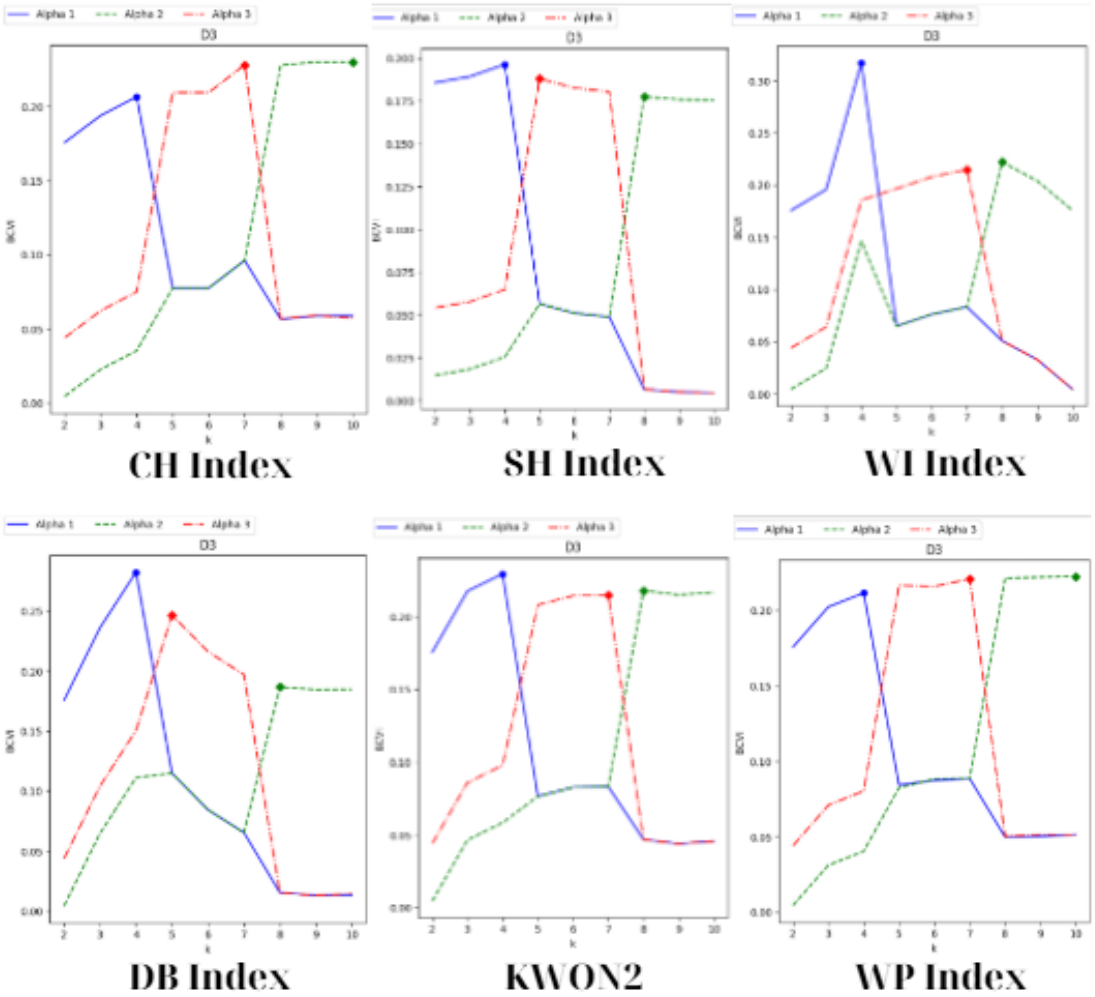
Chỉ số α

Các chỉ số alpha được đưa vào hiệu chỉnh cho BCVI

Data Type	α	Kmax	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Artificial	α_1	10	20	20	20	5	5	5	0.5	0.5	0.5
	α_2		0.5	0.5	0.5	5	5	5	20	20	20
	α_3		5	5	5	20	20	20	0.5	0.5	0.5
Real-world	α_1	10	25	25	2	2	0.5	0.5	0.5	0.5	0.5
	α_2		0.5	0.5	2	2	25	25	25	25	25
	α_3		2	2	25	25	0.5	0.5	0.5	0.5	0.5
MRI	α_1	8	25	25	2	2	0.5	0.5	0.5	-	-
	α_2		0.5	0.5	2	2	25	25	25	-	-
	α_3		2	2	25	25	0.5	0.5	0.5	-	-

Bộ dữ liệu nhân tạo

Bộ dữ liệu nhân tạo sau khi được tính BCVI bằng các chỉ số



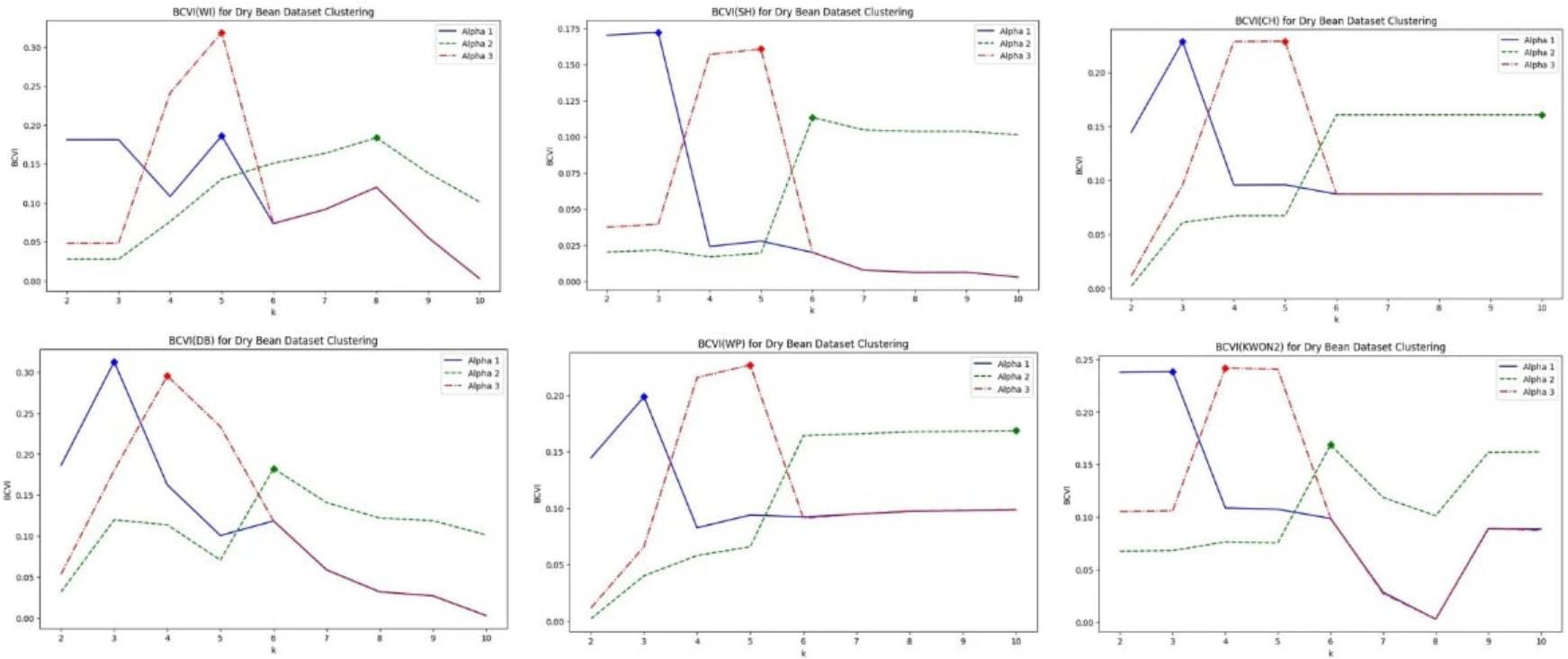
Bộ dữ liệu thực tế

Dữ liệu thực tế

- **Dữ liệu:** Là một bộ dữ liệu về các loại đậu.
- **Kích thước:** (13611, 17)
- **Đặc trưng:** Gồm 16 đặt trưng và 7 loại hạt đậu.

Bộ dữ liệu thực tế

Bộ dữ liệu thực tế sau khi được tính BCVI bằng các chỉ số



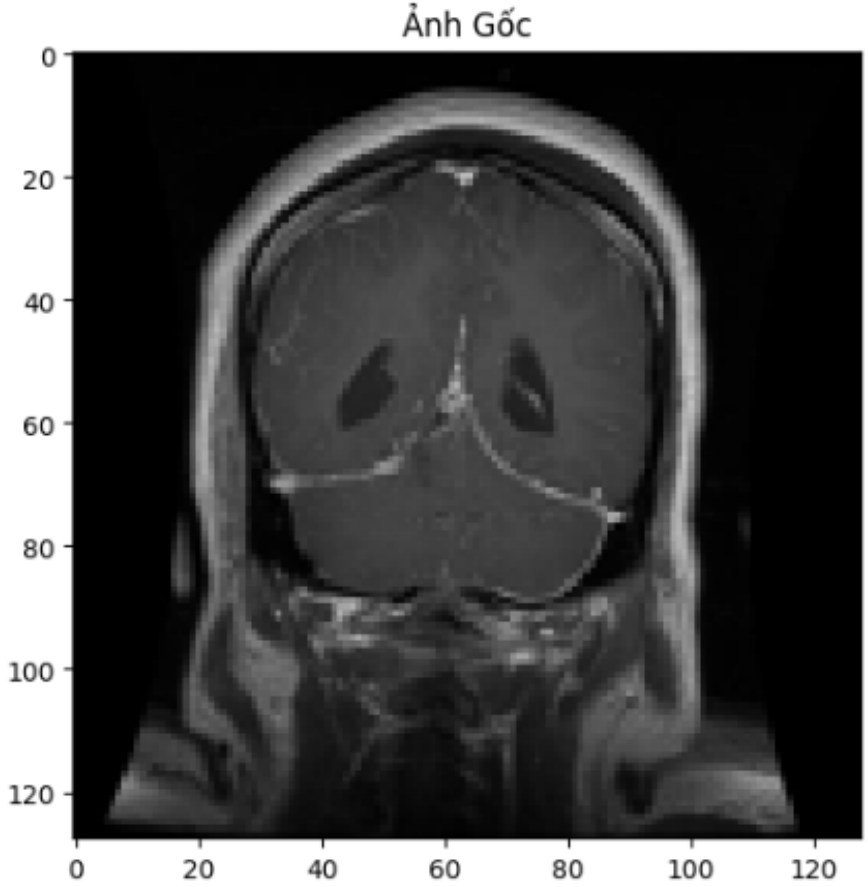
Bộ dữ liệu MRI não

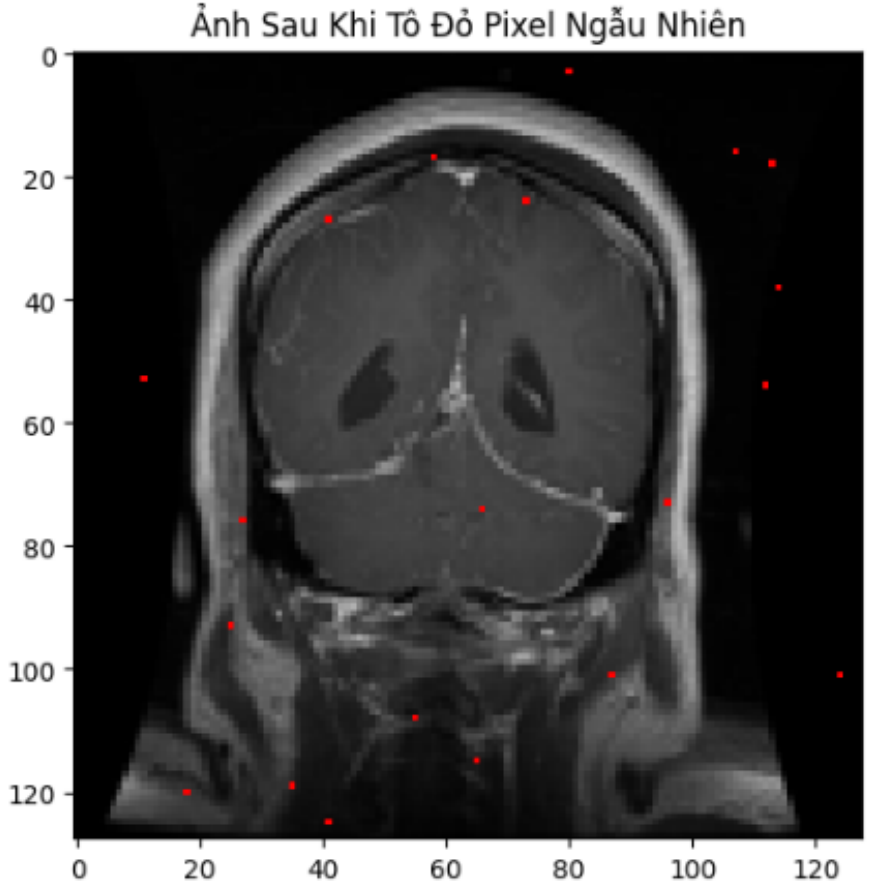
Dữ liệu thực tế MRI não

- **Dữ liệu**: 5712 ảnh(tumor và no tumor)
- **Kích thước**: (512, 512, 3)
- **Đặc trưng**: có dạng RGB (Red,Blue,Green)

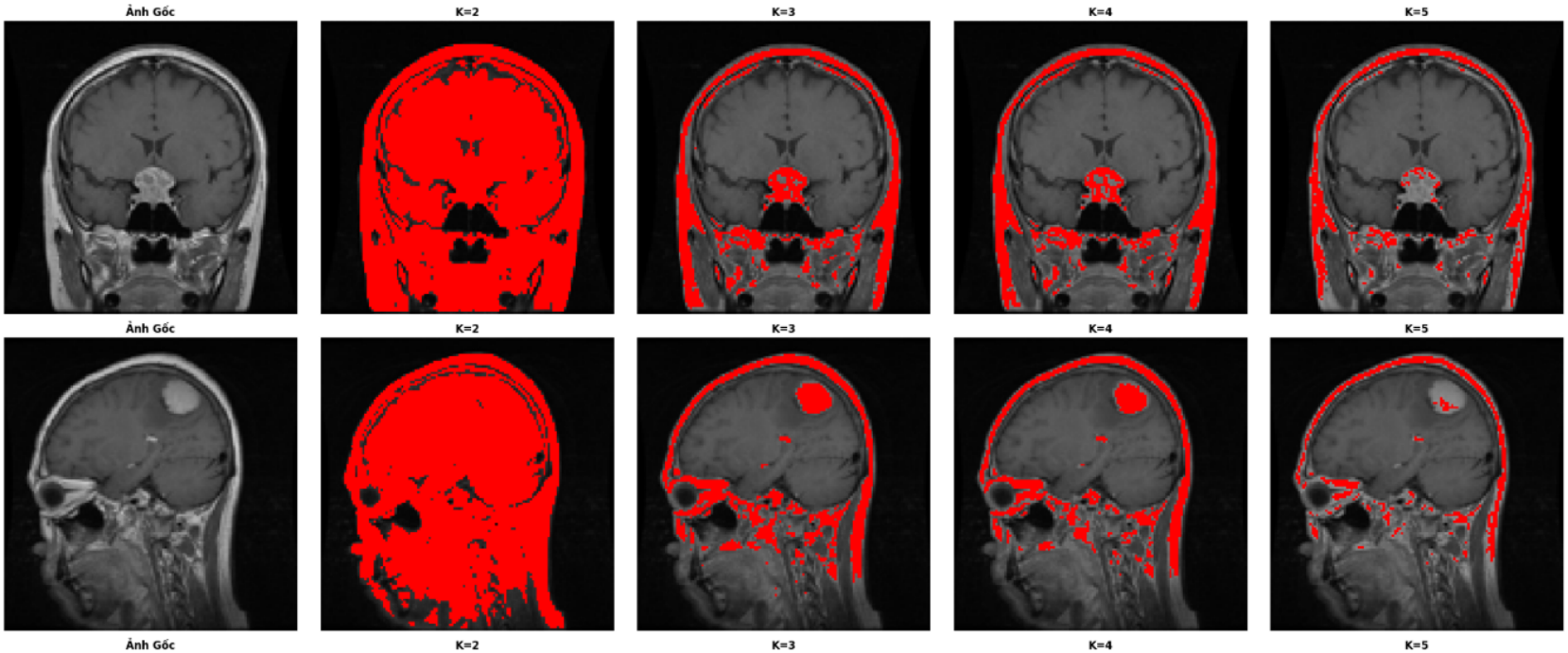
Bộ dữ liệu MRI não

```
Tọa độ (X, Y): (112, 54), Giá trị Pixel (RGB): [0 0 0]
Tọa độ (X, Y): (27, 76), Giá trị Pixel (RGB): [69 69 69]
Tọa độ (X, Y): (96, 73), Giá trị Pixel (RGB): [78 78 78]
Tọa độ (X, Y): (35, 119), Giá trị Pixel (RGB): [35 35 35]
Tọa độ (X, Y): (113, 18), Giá trị Pixel (RGB): [4 4 4]
```

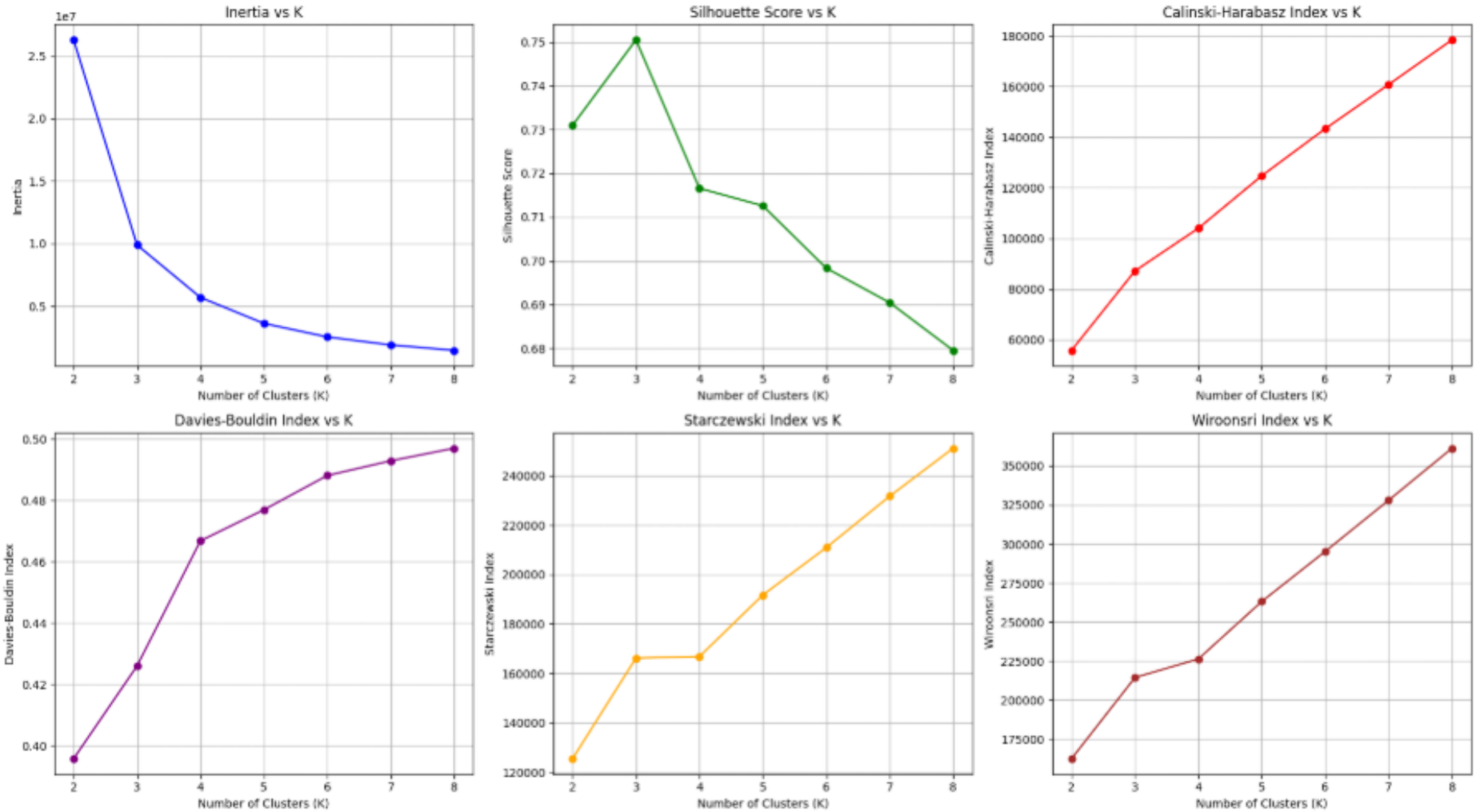




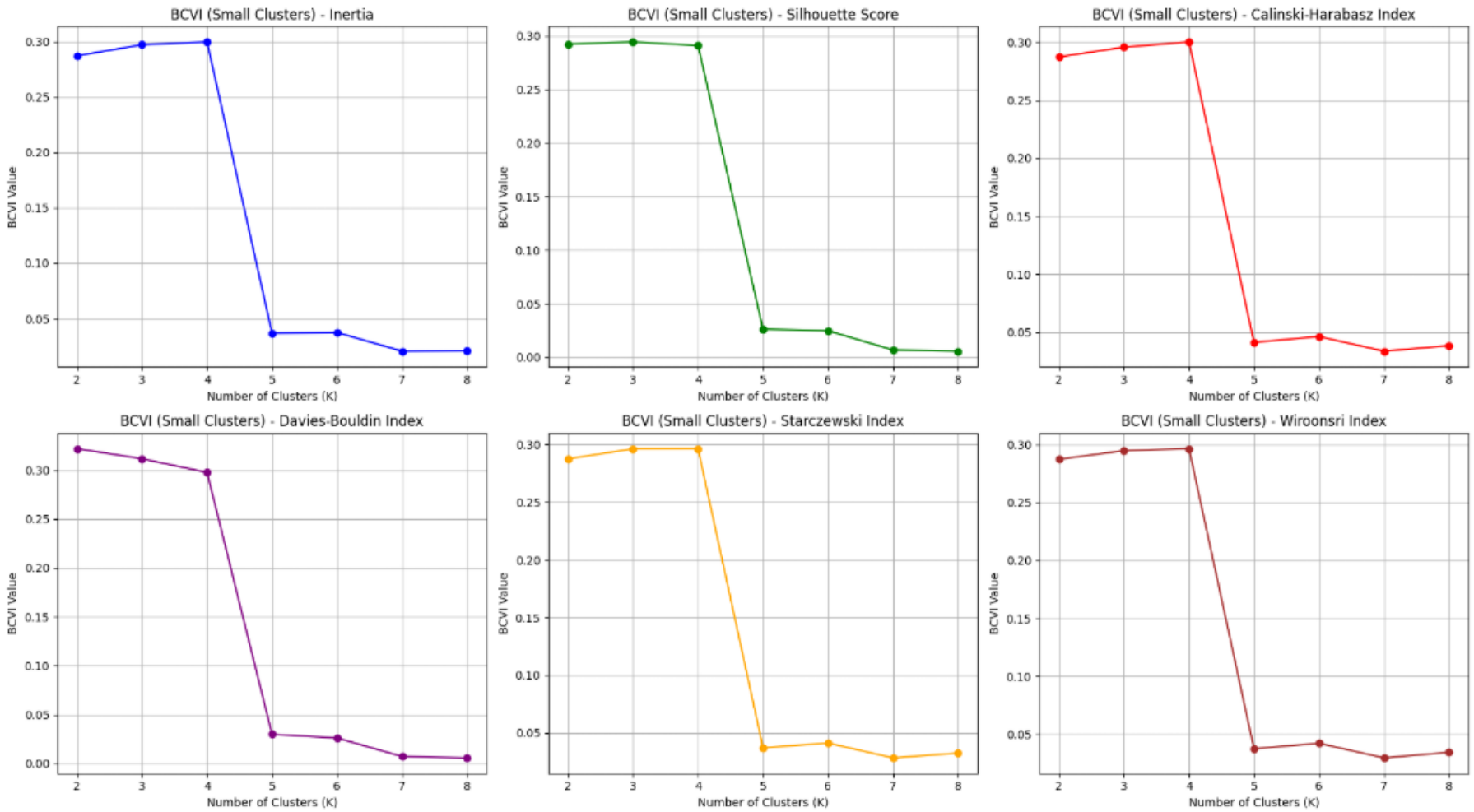
Bộ dữ liệu MRI não



Bộ dữ liệu MRI não



Bộ dữ liệu MRI não



Ứng dụng

- **Y tế:** Hỗ trợ phát hiện khối u não từ ảnh MRI.
- **Marketing:** Phân khúc khách hàng.
- **Dữ liệu lớn:** Phân tích cụm dữ liệu quy mô lớn.

Hạn chế

- Phụ thuộc vào giá trị tiên nghiệm α.
- Tốn tài nguyên tính toán cho dữ liệu lớn.
- Độ chính xác phụ thuộc vào các chỉ số CVI gốc.

Hướng phát triển

- Mở rộng BCVI cho các phân phối khác.
- Áp dụng trên tập dữ liệu lớn và phức tạp hơn.
- Tích hợp BCVI vào hệ thống tự động.

Cảm ơn

Cảm ơn!

Hỏi đáp?