TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - TIN HỌC

Báo Cáo

Seminar Khoa Học Dữ Liệu

Đề tài:

A Bayesian cluster validity index(Clustering)

Sinh viên thực hiện:	Trần Tuấn Đạt	21280079
	Phạm Tường Duy	21280011
Giảng viên hướng dẫn:	TS. Tô Đức Khánh	

Hồ Chí Minh, 9-2024

LỜI NÓI ĐẦU

Phần này dùng để trình bày Đề tài seminar nghiên cứu cách đánh giá tính hợp lý của số lượng cụm trong phân cụm dữ liệu, tập trung vào các bài toán phân cụm hình ảnh và dữ liệu thực tế. Mục tiêu là phát triển chỉ số Bayesian Cluster Validity Index (BCVI), dựa trên phân phối tiên nghiệm Dirichlet và Dirichlet tổng quát, cho phép người dùng linh hoạt tùy chỉnh theo nhu cầu ứng dụng. Hai phương pháp phân cụm K-Means và Fuzzy C-Means được sử dụng để đánh giá hiệu quả của BCVI, giúp xác định số cụm tối ưu so với các phương pháp truyền thống. Các thí nghiệm được thực hiện trên dữ liệu thời gian thực, dữ liệu số học và hình ảnh MRI não. Đồng thời, seminar so sánh sự khác biệt giữa kết quả nghiên cứu với các phương pháp trước đây nhằm kiểm tra độ chính xác và hiệu quả của chỉ số đề xuất. Lời cảm ơn tổ chức và cá nhân góp

LỜI CAM ĐOAN

Chúng tôi, Trần Tuấn Đạt, Phạm Tường Duy, sinh viên lớp 21KDL, dưới sự hướng dẫn của giảng viên Tô Đức Khánh, xin cam đoan rằng toàn bộ nội dung được trình bày trong đồ án **"Bayesian Clustering Validity Index"** là kết quả của quá trình tìm hiểu và nghiên cứu nghiêm túc của chúng tôi.

Các dữ liệu, kết quả thực nghiệm và các nhận định trong đồ án là hoàn toàn trung thực, phản ánh đúng những kết quả đo đạc và phân tích thực tế trong quá trình thực hiện. Tất cả các thông tin, số liệu trích dẫn từ các nguồn tài liệu tham khảo đều đã được ghi rõ ràng và đầy đủ, tuân thủ đúng các quy định về sở hữu trí tuệ.

Chúng tôi xin chịu hoàn toàn trách nhiệm về nội dung của đồ án này. Nếu có bất kỳ sai sót hay gian lận nào, chúng tôi hoàn toàn chịu trách nhiệm trước nhà trường và pháp luật.

Contents

DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Trong phần này, chúng tôi giới thiệu ngắn gọn về chỉ số kí hiệu viết tắt được sử dụng trong các phần tiếp theo. Cho $n,k,p \in \mathbb{N}$ và ký hiệu $[n] = \{1,2,\ldots,n\}$. Thiết lập các ký hiệu sau được sử dụng trong công trình này. Đối với $i \in [n]$ và $j \in [k]$, ký hiệu:

1. $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$: Điểm dữ liệu.
2. K: Số lượng cụm thực.
3. C_j: Tập các điểm dữ liệu trong cụm thứ j.
4. v_j: Trọng tâm của cụm thứ j.
5. v_0: Trọng tâm của toàn bộ tập dữ liệu.
6. $\bar{v}$: Trọng tâm của tất cả các v_j.
7. $\mu = (\mu_{ij})$: Ma trận độ thành viên nơi μ_{ij} biểu thị mức độ mà một điểm dữ liệu x_i thuộc về C_j.
8. $||x-y||$: Khoảng cách Euclidean giữa x và y.
9. $\text{Corr}(\cdot,\cdot)$: Hệ số tương quan. Trong công trình này, chúng tôi chỉ xem xét hệ số tương quan Pearson.

TÓM TẮT ĐỒ ÁN

1.1 Mục tiêu nghiên cứu

Báo cáo tập trung phát triển chỉ số đánh giá tính hợp lý của số lượng cụm trong bài toán phân cụm dữ liệu, đặc biệt là phân cụm hình ảnh và dữ liệu thực tế. Chỉ số được đề xuất là **Bayesian Cluster Validity Index (BCVI)**, dựa trên phân phối tiên nghiệm Dirichlet và Dirichlet tổng quát, giúp xác định số cụm tối ưu.

1.2 Phương pháp và thuật toán sử dụng

- **Thuật toán phân cụm:** K-Means và Fuzzy C-Means (FCM).
- **Các chỉ số đánh giá truyền thống:**
  - Calinski-Harabasz
  - Davies-Bouldin
  - Silhouette
- **BCVI:** Kết hợp kiến thức tiên nghiệm với dữ liệu thực tế, sử dụng phân phối Dirichlet và Dirichlet tổng quát để xác định số cụm tối ưu.

Ứng dụng thực nghiệm

Dữ liệu sử dụng:

- Dữ liệu nhân tạo (Gaussian, Uniform).
- Dữ liệu thực tế: Dry Bean (từ UCI Machine Learning Repository).
- Ảnh MRI não dùng cho bài toán phát hiện khối u não.

Các bước thực nghiệm:

- Tiền xử lý ảnh chuẩn hóa kích thước (128x128 pixels).
- Áp dụng K-Means và FCM để phân cụm.
- Đánh giá các chỉ số phân cụm với và không có BCVI.
- Thử nghiệm nhiều giá trị K và các chỉ số CVI.

Kết quả nổi bật

- BCVI giúp cải thiện khả năng xác định số lượng cụm tối ưu và cho kết quả chính xác hơn so với các chỉ số CVI truyền thống.
- Khi áp dụng vào bài toán phát hiện khối u não, BCVI xác định số cụm tối ưu là $K = 4$.
- Độ chính xác của phân cụm dữ liệu thực tế đạt trên 75%.

CHƯƠNG 1. CHƯƠNG MỞ ĐẦU

Phân cụm (Clustering) là một công cụ học không giám sát phổ biến trong thống kê và học máy, dùng để chia các quan sát thành các nhóm có hành vi tương đồng. Các thuật toán phân cụm phổ biến bao gồm: K-means, Fuzzy C-means (FCM), phân cụm phân cấp (Hierarchical Clustering), và phân cụm dựa trên mật độ (DBSCAN).

Trong quá trình phân cụm, một bước quan trọng là đánh giá khuynh hướng phân cụm (clustering tendency) nhằm xác định xem tập dữ liệu có thực sự chứa các cụm hay không và tìm số lượng cụm tối ưu. Các chỉ số xác thực cụm (Cluster Validity Index - CVI) thường được sử dụng trong bước này, chẳng hạn như Calinski-Harabasz, Davies-Bouldin, Silhouette và gần đây là Wiroonsri Index (WI).

Bài báo này tập trung vào việc phát triển một chỉ số mới có tên Bayesian Cluster Validity Index (BCVI) dựa trên các CVI hiện có, sử dụng phân phối tiên nghiệm Dirichlet và Dirichlet tổng quát. BCVI cho phép tích hợp kiến thức chuyên môn của người dùng trong việc lựa chọn số lượng cụm tối ưu, đồng thời có thể áp dụng cho cả các thuật toán phân cụm cứng (K-means) và phân cụm mềm (FCM).

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 K-means

K-means (MacQueen, 1967) là một thuật toán phân cụm đơn giản nhưng hiệu quả. Nó hoạt động bằng cách chia một tập dữ liệu thành k cụm, trong đó k là một tham số được người dùng xác định. Thuật toán bắt đầu bằng việc khởi tạo ngẫu nhiên các trọng tâm cụm. Sau đó, mỗi điểm dữ liệu được gán cho trọng tâm gần nhất và cập nhật các trọng tâm dựa trên các điểm dữ liệu mới được gán. Quá trình lặp lại này tiếp tục cho đến khi các trọng tâm cụm hội tụ. Mục tiêu của K-means là giảm thiểu tổng bình phương trong cụm, được biểu diễn như sau:

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

Chi tiết các bước:

1. **Khởi tạo:** Chọn K centroid ban đầu một cách ngẫu nhiên từ tập dữ liệu.
2. **Phân cụm:** Gán mỗi điểm dữ liệu x_i vào cụm có trọng tâm gần nhất.

   Euclidean:

   $$d(x_i, v_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - v_{jl})^2}$$

   Trong đó:

   - $d(x_i, v_j)$ là khoảng cách Euclidean giữa điểm dữ liệu x_i và centroid v_j.
   - p là số chiều của dữ liệu.
3. **Cập nhật centroid:** Tính lại trọng tâm của mỗi cụm dựa trên trung bình của các điểm dữ liệu trong cụm đó:

   $$v_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

   Trong đó:

- v_j là centroid mới của cụm j.
- $|C_j|$ là số lượng điểm dữ liệu trong cụm j.
- C_j là tập hợp các điểm dữ liệu thuộc cụm C_j.

4. **Lặp lại:** Lặp lại bước 2 và bước 3 cho đến khi không có sự thay đổi nào về việc phân cụm hoặc số lần lặp đạt đến giới hạn tối đa.

2.1.1 Giảm thiểu chi phí trong K-means Clustering

Trong thuật toán K-means Clustering, mục tiêu là tìm một cách phân chia dữ liệu sao cho tổng bình phương khoảng cách trong cụm (WCSS - Within-Cluster Sum of Squares) là nhỏ nhất:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Trong đó:

- k: Số cụm.
- C_i: Tập hợp các điểm dữ liệu thuộc cụm thứ i.
- μ_i: Centroid của cụm thứ i.
- x: Một điểm dữ liệu thuộc cụm C_i.
- $\|x - \mu_i\|^2$: Khoảng cách bình phương từ điểm x đến centroid μ_i.

Ý nghĩa của WCSS:

- Đây là tổng khoảng cách bình phương từ mỗi điểm dữ liệu đến trung tâm của cụm mà nó thuộc về, WCSS càng nhỏ thì phân cụm càng tốt.
- Trong mỗi cụm, WCSS đo lường sự phân tán của các điểm dữ liệu xung quanh centroid của cụm đó.

2.2 Fuzzy C-means

FCM, được giới thiệu bởi Dunn (1973) và sau đó được cải tiến bởi Bezdek và các cộng sự (1984), là một kỹ thuật phân cụm được sử dụng để nhóm các điểm dữ

liệu tương tự vào các cụm do người dùng chỉ định là c. Mỗi điểm dữ liệu được gán một mức độ thành viên, biểu thị mức độ thuộc về từng cụm. Mục tiêu của FCM là giảm thiểu hàm mục tiêu sau:

$$\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{m}\left\|x_i - v_j\right\|^2,$$

trong đó $m > 1$ biểu thị tham số mờ. Việc tối ưu hóa (1) bắt đầu bằng việc khởi tạo ngẫu nhiên các trọng tâm v_j. Lặp lại, các mức độ thành viên được cập nhật theo công thức:

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

và các trọng tâm được cập nhật như sau:

$$v_j = \frac{\sum_{i=1}^{n}\mu_{ij}^{m}x_i}{\sum_{i=1}^{n}\mu_{ij}^{m}},$$

cho $i \in [n]$ và $j \in [c]$. Quá trình lặp này tiếp tục cho đến khi đạt được sự hội tụ.

CHƯƠNG 3. Các chỉ số index

3.1 Các chỉ số trong phân cụm cứng

3.1.1 Chỉ số Calinski-Harabasz

Chỉ số Calinski-Harabasz (Caliński và Harabasz, 1974) được định nghĩa như sau:

$$\mathrm{CH}(k) = \frac{n-k}{k-1} \frac{\sum_{j=1}^{k} |C_j| \|v_j - v_0\|^2}{\sum_{j=1}^{k} \sum_{x \in C_j} \|x - v_j\|^2}$$

Ý nghĩa:

- n là số điểm dữ liệu, k là số cụm được xác định.
- $|C_j|$ là số lượng điểm trong cụm j, v_j là tâm của cụm j, và v_0 là tâm của toàn bộ tập dữ liệu.
- Tử số đại diện cho khoảng cách giữa các tâm cụm và tâm của toàn bộ dữ liệu (khoảng cách giữa cụm), còn mẫu số là tổng khoảng cách từ mỗi điểm dữ liệu đến tâm cụm của nó (độ gần nhau trong mỗi cụm).

Sử dụng: Giá trị lớn nhất của $\mathrm{CH}(k)$ chỉ ra sự phân chia cụm tối ưu, do nó phản ánh sự phân tách tốt giữa các cụm và độ gần nhau trong mỗi cụm.

3.1.2 Chỉ số xác thực cụm dựa trên các điểm gần nhất (CVNN)

Chỉ số CVNN (Liu et al., 2013) được định nghĩa như sau:

$$\mathrm{CVNN}(k, NN) = \frac{\mathrm{Sep}(k, NN)}{\max_{K_{\min} \le k \le K_{\max}} \mathrm{Sep}(k, NN)} + \frac{\mathrm{Com}(k)}{\max_{K_{\min} \le k \le K_{\max}} \mathrm{Com}(k)},$$

Ý nghĩa:

- $\mathrm{Sep}(k, NN)$ là sự phân tách giữa các cụm, được tính bằng cách xem xét các điểm gần nhất không thuộc cùng cụm:

$$\mathrm{Sep}(k, NN) = \max_{j \in [k]} \frac{1}{|C_j|} \sum_{x \in C_j} \frac{q_x}{NN},$$

với NN là số lượng điểm gần nhất, q_x là số điểm gần nhất của x không thuộc cụm của nó.

- $\mathrm{Com}(k)$ đo lường độ gọn của cụm:

$$\mathrm{Com}(k) = \sum_{j=1}^{k} \left[2\frac{\sum_{x,y\in C_j} \|x-y\|}{|C_j|(|C_j|-1)} \right],$$

với x và y là hai điểm khác nhau trong cụm j.

Sử dụng: Giá trị nhỏ nhất của $\mathrm{CVNN}(k)$ chỉ ra cấu trúc cụm tốt nhất, vì nó cân nhắc sự phân tách và độ gọn của cụm.

3.1.3 Chỉ số Davies và Bouldin

Chỉ số Davies-Bouldin (Davies và Bouldin, 1979) được định nghĩa như sau:

$$\mathrm{DB}(k) = \frac{1}{k}\sum_{i=1}^{k} R_{i,\mathrm{qt}},$$

với:

$$R_{i,\mathrm{qt}} = \max_{j\in[k]\setminus\{i\}} \left\{ \frac{S_{i,q}+S_{j,q}}{M_{ij,t}} \right\},$$

$$S_{i,q} = \left(\frac{1}{|C_i|}\sum_{x\in C_i} \|x-v_i\|^q \right)^{1/q},$$

và

$$M_{ij,t} = \left(\sum_{s=1}^{p} |v_{is}-v_{js}|^t \right)^{1/t}.$$

Ý nghĩa:

- $S_{i,q}$ là độ rải rác của cụm i với tham số q.
- $M_{ij,t}$ là khoảng cách giữa tâm của cụm i và j với tham số t.

Sử dụng: Giá trị nhỏ nhất của $\mathrm{DB}(k)$ chỉ ra sự phân chia cụm tốt nhất, vì nó tối thiểu hóa sự chồng chéo giữa các cụm.

3.1.4 Chỉ số Silhouette

Với $i \in [n]$, $l \in [k]$, và $x_i \in C_l$, ta có:

$$a(i) = \frac{1}{|C_l|-1}\sum_{y\in C_l} \|x_i-y\| \quad \text{và} \quad b(i) = \min_{r\neq l} \frac{1}{|C_r|}\sum_{y\in C_r} \|x_i-y\|.$$

Giá trị Silhouette của một điểm dữ liệu x_j được định nghĩa:

$$s(j) = \begin{cases} \frac{b(j)-a(j)}{\max\{a(j),b(j)\}} & \text{nếu } |C_j| > 1 \\ 0 & \text{nếu } |C_j| = 1 \end{cases}$$

Chỉ số Silhouette (Rousseeuw, 1987; Kaufman và Rousseeuw, 2009) là:

$$\mathrm{SH}(k) = \frac{1}{n}\sum_{i=1}^{n} s(i).$$

Ý nghĩa:

- $a(i)$ là khoảng cách trung bình từ điểm x_i đến các điểm khác trong cùng cụm.
- $b(i)$ là khoảng cách trung bình nhỏ nhất từ điểm x_i đến các điểm trong một cụm khác.
- $s(i)$ đo lường mức độ phù hợp của một điểm trong cụm của nó so với các cụm khác.

Sử dụng: Giá trị lớn nhất của $\mathrm{SH}(k)$ chỉ ra sự phân chia cụm tối ưu, vì nó phản ánh rằng các điểm nằm trong cụm của mình một cách hợp lý hơn so với các cụm khác.

Chỉ số đánh giá phân cụm dựa trên khoảng cách đến hàng xóm gần nhất

Chỉ số CVNN (Liu và các cộng sự, 2013) là một biện pháp đánh giá nội bộ được định nghĩa như sau:

$$\mathrm{CVNN}(k,NN) = \frac{\mathrm{Sep}(k,NN)}{\max_{K\min\le k\le K\max}\mathrm{Sep}(k,NN)} + \frac{\mathrm{Com}(k)}{\max_{K\min\le k\le K\max}\mathrm{Com}(k)},$$

trong đó Sep và Com được định nghĩa như sau:

- Sự tách biệt giữa các cụm:

$$\mathrm{Sep}(k,NN) = \max_{j\in[k]} \frac{1}{\left|C_j\right|}\sum_{x\in C_j}\frac{q_x}{NN},$$

trong đó NN là số lượng hàng xóm gần nhất được nhập vào và q_x biểu thị số lượng hàng xóm gần nhất của x nằm ngoài cụm của nó.

- Tính gọn gàng trong cụm:

$$\text{Com}(k) = \sum_{j=1}^{k} \left[2\frac{\sum_{x,y\in C_j} \|x-y\|}{|C_j|\left(|C_j|-1\right)} \right],$$

trong đó x và y là hai đối tượng khác nhau trong C_j.

Giá trị nhỏ nhất của $\text{CVNN}(k)$ biểu thị một phân chia tối ưu hợp lệ.

3.1.5 Chỉ số Starczewski

Chỉ số Starczewski (Starczewski, 2017) được định nghĩa như sau:

$$\text{STR}(k) = [E(k) - E(k-1)][D(k+1) - D(k)],$$

với

$$D(k) = \frac{\max_{i,j\in[k]} \|v_i - v_j\|}{\min_{i,j\in[k]} \|v_i - v_j\|},$$

và

$$E(k) = \frac{\sum_{i=1}^{n} \|x_i - v_0\|}{\sum_{j=1}^{k} \sum_{x\in C_j} \|x - v_j\|}.$$

Ý nghĩa:

- $D(k)$ đại diện cho tỷ lệ giữa khoảng cách xa nhất và gần nhất giữa các tâm cụm, phản ánh sự phân tách của các cụm.
- $E(k)$ là tỷ lệ giữa tổng khoảng cách từ các điểm dữ liệu đến tâm của toàn bộ dữ liệu và tổng khoảng cách từ các điểm đến tâm của từng cụm, đo lường độ gọn của cụm.

Sử dụng: Giá trị lớn nhất của $\text{STR}(k)$ chỉ ra sự phân chia cụm tối ưu, vì nó phản ánh sự cải thiện từ cụm $k-1$ đến cụm k và sự giảm sút từ cụm k đến cụm $k+1$.

3.1.6 Chỉ số Wiroonsri

Chỉ số Wiroonsri (Wiroonsri, 2024) được định nghĩa như sau:

Với $m \in \{2,3,\ldots,n-1\}$ và $k \in \{2,3,\ldots,m\}$:

Trường hợp 1: $\max_{2\le l\le m} \mathrm{NCI1}(k) < +\infty$

$$\mathrm{NCI}_m(k) = \begin{cases} \min_{2\le l\le m}\{\mathrm{NCI1}(l) \mid \mathrm{NCI1}(l) > -\infty\} & \text{nếu } \mathrm{NCI1}(k) = -\infty \\ \mathrm{NCI1}(k) & \text{còn lại,} \end{cases}$$

Trường hợp 2: $\max_{2\le l\le m} \mathrm{NCI1}(k) = +\infty$

$$\mathrm{NCI}_m(k) = \begin{cases} \min\limits_{2\le l\le m}\{\mathrm{NCI1}(l) \mid \mathrm{NCI1}(l) > -\infty\} + \mathrm{NCI2}(k) & \text{nếu } \mathrm{NCI1}(k) = -\infty \\ \max\limits_{2\le l\le m}\{\mathrm{NCI1}(l) \mid \mathrm{NCI1}(l) < +\infty\} + \mathrm{NCI2}(k) & \text{nếu } \mathrm{NCI1}(k) = +\infty \\ \mathrm{NCI1}(k) + \mathrm{NCI2}(k) & \text{còn lại,} \end{cases}$$

trong đó:

$$\mathrm{NCI1}(k) = \frac{(\mathrm{NC}(k) - \mathrm{NC}(k-1))(1-\mathrm{NC}(k))}{\max\{0, (\mathrm{NC}(k+1) - \mathrm{NC}(k))(1-\mathrm{NC}(k-1))\}},$$

và

$$\mathrm{NCI2}(k) = \frac{\mathrm{NC}(k) - \mathrm{NC}(k-1)}{1-\mathrm{NC}(k-1)} - \frac{\mathrm{NC}(k+1) - \mathrm{NC}(k)}{1-\mathrm{NC}(k)},$$

với $\mathrm{NC} = \mathrm{Corr}(\vec{d}, \vec{c}(k))$, $\mathrm{NC}(1) = \frac{\mathrm{SD}(\vec{d}_v)}{\max \vec{d}_v - \min \vec{d}_v}$. Lưu ý rằng ta để:

$$\vec{d}_v = (\|x_i - v_0\|)_{i\in[n]},$$

$$\vec{d} = (\|x_i - x_j\|)_{i,j\in[n]},$$

là vector của độ dài $\binom{n}{2}$ chứa khoảng cách của tất cả các cặp điểm dữ liệu, và

$$\vec{c}(k) = (\|v_i(k) - v_j(k)\|)_{i,j\in[n]},$$

là vector của cùng độ dài chứa khoảng cách của tất cả các cặp tâm cụm tương ứng của các cụm mà hai điểm đang nằm trong.

Ý nghĩa:

- NC là hệ số tương quan giữa khoảng cách các điểm dữ liệu và khoảng cách các tâm cụm.

- NCI1 và NCI2 giúp xác định các đỉnh cục bộ của chỉ số, cho phép tìm ra số lượng cụm tối ưu và các lựa chọn thứ cấp.

Sử dụng: Giá trị lớn nhất của $\mathrm{WI}(k)$ chỉ ra sự phân chia cụm tối ưu, do nó cân nhắc cả sự phân tách và độ gọn của cụm, đồng thời cung cấp thông tin về các lựa chọn không tối ưu nhất.

3.2 Các chỉ số trong phân cụm mềm

3.2.1 Chỉ số KWON2

Chỉ số KWON2 (Kwon et al., 2021) được định nghĩa như sau:

$$\mathrm{KWON2}(k) = \frac{w_1 \left[w_2 \sum_{j=1}^{k} \sum_{i=1}^{n} \mu_{ij}^{2\sqrt{\frac{2}{2}}} \left\| x_i - v_j \right\|^2 + \frac{\sum_{j=1}^{k} \left\| v_j - v_0 \right\|^2}{\max_j \left\| v_j - v_0 \right\|^2} + w_3 \right]}{\min_{i \neq j} \left\| v_i - v_j \right\|^2 + \frac{1}{k} + \frac{1}{k^{m-1}}},$$

- **Ý nghĩa thành phần:** - $w_1 = \frac{n-k+1}{n}$: Trọng số giảm dần khi số cụm k tăng, giúp điều chỉnh mức độ phức tạp của phân cụm. - $w_2 = \left(\frac{k}{k-1}\right)^{\sqrt{2}}$: Điều chỉnh độ tập trung của các điểm trong cụm. - $w_3 = \frac{nk}{(n-k+1)^2}$: Trọng số phụ thuộc vào số điểm dữ liệu và số cụm. - μ_{ij}: Mức độ thành viên của điểm x_i trong cụm j. - $\|x_i - v_j\|$: Khoảng cách từ điểm x_i đến tâm cụm v_j. - $\|v_j - v_0\|$: Khoảng cách từ tâm cụm j đến tâm của toàn bộ dữ liệu. - $\min_{i \neq j} \|v_i - v_j\|$: Khoảng cách nhỏ nhất giữa các tâm cụm khác nhau, đánh giá sự tách biệt giữa các cụm.

Cách sử dụng: - Giá trị nhỏ nhất của $\mathrm{KWON2}(k)$ chỉ ra số lượng cụm tối ưu, phản ánh sự tập trung cao và sự tách biệt tốt giữa các cụm.

3.2.2 Chỉ số Wiroonsri và Preedasawakul

Chỉ số WP (Wiroonsri và Preedasawakul, 2023a) được định nghĩa theo ba trường hợp. Với $m \in \{2,3,\ldots,n-1\}$ và $k \in \{2,3,\ldots,m\}$,

Trường hợp 1: $\max_{2 \leq l \leq p} \mathrm{WPCI1}(k) < +\infty$ và tồn tại $l \in [p] \backslash \{1\}$ sao cho $|\mathrm{WPCI1}(l)| < \infty$.

$$\mathrm{WP}_p(k) = \begin{cases} \min_{2 \leq l \leq p} \{\mathrm{WPCI1}(l) \mid \mathrm{WPCI1}(l) > -\infty\} & \text{nếu } \mathrm{WPCI1}(k) = -\infty \\ \mathrm{WPCI1}(k) & \text{trong trường hợp khác,} \end{cases}$$

Trường hợp 2:

$$\max_{2\le l\le p} \text{WPCI1}(k) = +\infty \quad \text{và tồn tại } l \in \{2,3,\ldots,p\} \text{ sao cho } |\text{WPCI1}(l)| < \infty.$$

$$\text{WP}_p(k) = \begin{cases} \min\limits_{2\le l\le p} \{\text{WPCI1}(l) \mid \text{WPCI1}(l) > -\infty\} + \text{WPCI2}(k) & \text{nếu WPCI1}(k) = -\infty \\ \max\limits_{2\le l\le p} \{\text{WPCI1}(l) \mid \text{WPCI1}(l) < +\infty\} + \text{WPCI2}(k) & \text{nếu WPCI1}(k) = +\infty \\ \text{WPCI1}(k) + \text{WPCI2}(k) & \text{trong trường hợp khác.} \end{cases}$$

Trường hợp 3: $\forall l \in \{2,3,\ldots,p\}, |\text{WPCI1}(l)| = +\infty.$

$$\text{WP}_p(k) = \text{WPCI2}(k),$$

- **Ý nghĩa thành phần:** - $\text{WPCI1}(k)$ và $\text{WPCI2}(k)$: Được định nghĩa tương tự như (2) và (3), nhưng thay NC bằng WPC, nơi $\text{WPC}(k) = \text{Corr}(\vec{d}, \vec{v}(k))$, $\text{WPC}(1) = \frac{\text{SD}(\vec{d}_k)}{\max \vec{d}_k - \min \vec{d}_k}$, $\vec{d}_v$ và $\vec{d}$ là như trong (5) và (4) tương ứng. - $o_i(k,\gamma) = \frac{\sum_{j=1}^k \mu_{ij}^k v_j}{\sum_{j=1}^k \mu_{ij}^k}$: Tâm mềm của điểm x_i khi xem xét số cụm k. - $\vec{v}(k) = \left(\left\|o_i(k,\gamma) - o_j(k,\gamma)\right\|\right)_{i,j\in[n]}$: Vector khoảng cách giữa các tâm mềm.

Cách sử dụng: - Giá trị lớn nhất của $\text{WP}(k)$ chỉ ra số lượng cụm tối ưu, cung cấp cả thông tin về các lựa chọn phụ để người dùng có thể xem xét.

3.2.3 Chỉ số Xie và Beni

Chỉ số XB (Xie và Beni, 1991) được định nghĩa như sau:

$$\text{XB}(k) = \frac{\sum_{j=1}^k \sum_{i=1}^n \mu_{ij}^2 \|x_i - v_j\|^2}{n \cdot \min_{j\neq l} \left\{\|v_j - v_l\|^2\right\}}.$$

- **Ý nghĩa thành phần:** - μ_{ij}^2: Bình phương mức độ thành viên, tăng trọng lượng cho các điểm gần tâm cụm hơn. - $\|x_i - v_j\|$: Khoảng cách từ điểm dữ liệu đến tâm cụm. - $\min_{j\neq l} \|v_j - v_l\|^2$: Khoảng cách nhỏ nhất giữa các tâm cụm, đánh giá sự tách biệt của các cụm.

Cách sử dụng: - Giá trị nhỏ nhất của $\text{XB}(k)$ chỉ ra số lượng cụm tối ưu, thể hiện sự tập trung cao và sự tách biệt tốt giữa các cụm.

CHƯƠNG 4. Giới thiệu về thuật toán BCVI

4.1 Giới thiệu

BCVI sử dụng các phân phối xác suất (Dirichlet hoặc Generalized Dirichlet) để mô hình hóa xác suất của các số lượng cụm khác nhau, cho phép tích hợp cả dữ liệu thực tế và kiến thức chuyên môn từ người dùng. Thuật toán này bao gồm các bước sau:

1. Xác định các chỉ số hiệu lực cụm ban đầu: Thuật toán bắt đầu với việc chọn một hoặc nhiều chỉ số hiệu lực cụm truyền thống (ví dụ: Calinski-Harabasz, Silhouette, Davies-Bouldin, v.v.) để làm cơ sở cho việc đánh giá.

2. Áp dụng phân phối Dirichlet hoặc Dirichlet tổng quát: Các phân phối này được sử dụng để xác định xác suất tiên nghiệm (prior) cho số lượng cụm dựa trên kinh nghiệm hoặc kiến thức chuyên môn của người dùng.

3. Tính toán phân phối hậu nghiệm (posterior): Dựa trên dữ liệu thu thập được và phân phối tiên nghiệm đã xác định, thuật toán tính toán phân phối hậu nghiệm để đưa ra xác suất cho các số lượng cụm có thể xảy ra.

4. Lựa chọn số lượng cụm tối ưu: Số cụm tối ưu hoặc các cực đại cục bộ được chọn dựa trên các xác suất hậu nghiệm, cho phép người dùng linh hoạt trong việc chọn số cụm phù hợp nhất với ứng dụng của họ.

4.2 Phân phối Dirichlet

4.2.1 Cơ Sở Lý Thuyết

Dirichlet Prior: Chúng tôi sử dụng phân phối Dirichlet với các tham số $\alpha = (\alpha_2, \alpha_3, ..., \alpha_K)$ để thiết lập xác suất tiên nghiệm cho mỗi số lượng cụm k. Hàm mật độ xác suất của Dirichlet là:

$$f(x_1, \ldots, x_K \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1},$$

trong đó $B(\boldsymbol{\alpha})$ là hàm beta đa biến, được định nghĩa bởi:

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

Generalized Dirichlet Prior: Để mô tả mối quan hệ phức tạp hơn giữa các cụm, chúng tôi sử dụng phân phối Generalized Dirichlet với tham số α và β. Hàm mật độ xác suất của GD là:

$$f(x_1,\ldots,x_{K-1} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{k=1}^{K-1} \frac{x_k^{\alpha_k-1}(1-x_1-\cdots-x_k)^{\gamma_k}}{B(\alpha_k,\beta_k)},$$

với $\gamma_k = \beta_k - \alpha_{k+1} - \beta_{k+1}$ cho $k = 1,2,...,K-2$, và $\gamma_{K-1} = \beta_{K-1} - 1$.

4.3 Thuật Toán BCVI

- **Chọn CVI cơ bản:** Chúng tôi chọn các chỉ số phân cụm đã có như Calinski-Harabasz hay Davies-Bouldin.
- **Xác định Phân Phối Tiên Nghiệm:** Tham số α (và β trong trường hợp GD) được điều chỉnh dựa trên kiến thức của người dùng.
- **Tính Tỷ Lệ R:** Dựa trên chỉ số CVI gốc, tính toán tỷ lệ điều chỉnh $r_k(\mathbf{x})$ như sau:

$$r_k(\mathbf{x}) = \begin{cases} \frac{GI(k)-\min_j GI(j)}{2\max_j GI(j)-\min_j GI(j)} & \text{cho Điều kiện A,} \\ \frac{\max_j GI(j)-GI(k)}{\sum_{i=2}^{K}(\max_j GI(j)-GI(i))} & \text{cho Điều kiện B,} \end{cases}$$

với GI là chỉ số phân cụm gốc.

- **Cập Nhật Phân Phối Hậu Nghiệm:** Sử dụng dữ liệu để cập nhật phân phối tiên nghiệm thành hậu nghiệm. Với Dirichlet:

$$\pi(\mathbf{p} \mid \mathbf{x}) = \frac{1}{B(\alpha + n\mathbf{r}(\mathbf{x}))} \prod_{k=2}^{K} p_k^{\alpha_k + nr_k(\mathbf{x})-1},$$

Với Generalized Dirichlet:

$$\pi(\mathbf{p} \mid \mathbf{x}) = \prod_{k=2}^{K-1} \frac{p_k^{\alpha'_k}(1-p_2-\cdots-p_k)^{\gamma'_k}}{B(\alpha'_k,\beta'_k)},$$

với $\alpha'_k = \alpha_k + nr_k(\mathbf{x})$ và $\beta'_k = \beta_k + \sum_{i=k+1}^{K} nr_i(\mathbf{x})$.

- **Tính BCVI:**

$$\text{BCVI}(k) = \mathbb{E}[p_k \mid \mathbf{x}]$$

BCVI cho mỗi số lượng cụm k được tính bằng kỳ vọng của xác suất hậu nghiệm p_k của việc có đúng k cụm trong tập dữ liệu khi đã biết dữ liệu $\mathbf{x}$. Cụ thể, với phân phối tiên nghiệm Dirichlet:

$$\mathbb{E}[p_k \mid \mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\sum_{j=2}^{K}(\alpha_j + nr_j(\mathbf{x}))},$$

với phân phối Generalized Dirichlet:

$$\mathbb{E}[p_k \mid \mathbf{x}] = \begin{cases} \frac{\alpha'_k}{\alpha'_k+\beta'_k}\prod_{i<k}\frac{\beta'_i}{\alpha'_i+\beta'_i} & \text{cho } k < K, \\ \prod_{i=2}^{K-1}\frac{\beta'_i}{\alpha'_i+\beta'_i} & \text{cho } k = K, \end{cases}$$

trong đó $\alpha'_k = \alpha_k + nr_k(\mathbf{x})$ và $\beta'_k = \beta_k + \sum_{i=k+1}^{K} nr_i(\mathbf{x})$.

Điều này có nghĩa là BCVI tính toán xác suất mà số lượng cụm k là đúng dựa trên cả dữ liệu và kiến thức tiên nghiệm, cho phép người dùng quyết định số lượng cụm cuối cùng một cách hiệu quả.

4.4 Tính Chất

- **Tính Linh Hoạt:** BCVI cho phép điều chỉnh theo kiến thức tiên nghiệm.
- **Tính Nghiệm Hậu:** Kết hợp dữ liệu với kiến thức trước để đưa ra quyết định.
- **Khả Năng Cung Cấp Lựa Chọn Phụ:** Phát hiện các đỉnh cục bộ trong phân phối hậu nghiệm.
- **Độ Chính Xác:** Cải thiện việc xác định số lượng cụm.

4.5 Ứng Dụng

- **Y tế:** Xử lý hình ảnh y tế như MRI.
- **Marketing:** Phân khúc khách hàng.
- **Nghiên cứu Khoa Học:** Từ sinh học đến xã hội học.
- **Big Data:** Xử lý dữ liệu lớn.

4.6 Kết Luận

BCVI mang lại một cách tiếp cận mới trong phân cụm, kết hợp kiến thức chuyên môn với dữ liệu thực tế, cung cấp một công cụ mạnh mẽ và linh hoạt cho việc phân tích cụm trong nhiều lĩnh vực. Nghiên cứu tiếp theo có thể bao gồm việc mở rộng BCVI cho các loại phân phối tiên nghiệm khác và kiểm tra hiệu suất trên các loại dữ liệu phức tạp hơn.

CHƯƠNG 5. Kết quả nghiên cứu

Trong phần này thuật toán K-Means được áp dụng vào việc phân cụm các tập dữ liệu nhân tạo bằng phân phối Gaussian và phân phối đều và bộ dữ liệu được đánh số thứ tự từ D1 đến D25. Bộ dữ liệu ở đây khác với bộ dữ liệu trong bài báo BCVI mà bài Seminar đang nghiên cứu. Một điểm khác biệt so với bài báo là bộ dữ liệu này được sử dụng chung cho cả hai phương pháp phân cụm là K-mean và cả FCM tuy nhiên thì bộ dữ liệu này vẫn có độ chính xác giữa việc phân cụm và nhãn có sẵn của dữ liệu là trên 75% không có sự khác biệt so với bộ dữ liệu gốc của bài báo. Bộ dữ liệu Artificial được lấy từ một trang web GitHub (bởi O-PREEDASAWAKUL 2023) và bộ dữ liệu Real-World được thêm vào phần nghiên cứu này là Dry Bean từ UCI (M. Koklu, Ilker Ali Özkan, 2020). Bộ dữ liệu MRI là bộ dữ liệu gốc từ bài báo BCVI đã được nghiên cứu trước đó và được sử dụng lại để đánh giá. Sự khác biệt chính của dữ liệu MRI là kích thước của các ảnh sẽ lớn hơn với kích thước trước đó và được đồng bộ về cùng một kích thước là 128x128 pixels trước khi áp dụng các chỉ số đánh giá và thuật toán BCVI vào bộ dữ liệu.

Artificial Datasets			**Real-world Datasets**		
Data	**Kmeans Acc**	**Fuzzy C-means Acc**	**Data**	**Kmeans Acc**	**Fuzzy C-means Acc**
D1	1.0000	1.0000	Dry Bean	0.7865	0.8517
D2	1.0000	1.0000	-	-	-
D3	0.9614	0.9779	-	-	-
D4	0.9167	1.0000	-	-	-
D5	1.0000	0.8517	-	-	-
-	-	-	-	-	-
D21	1.0000	1.0000	-	-	-
D22	1.0000	0.8538	-	-	-
D23	1.0000	0.8333	-	-	-
D24	0.9880	0.9907	-	-	-
D25	1.0000	1.0000	-	-	-

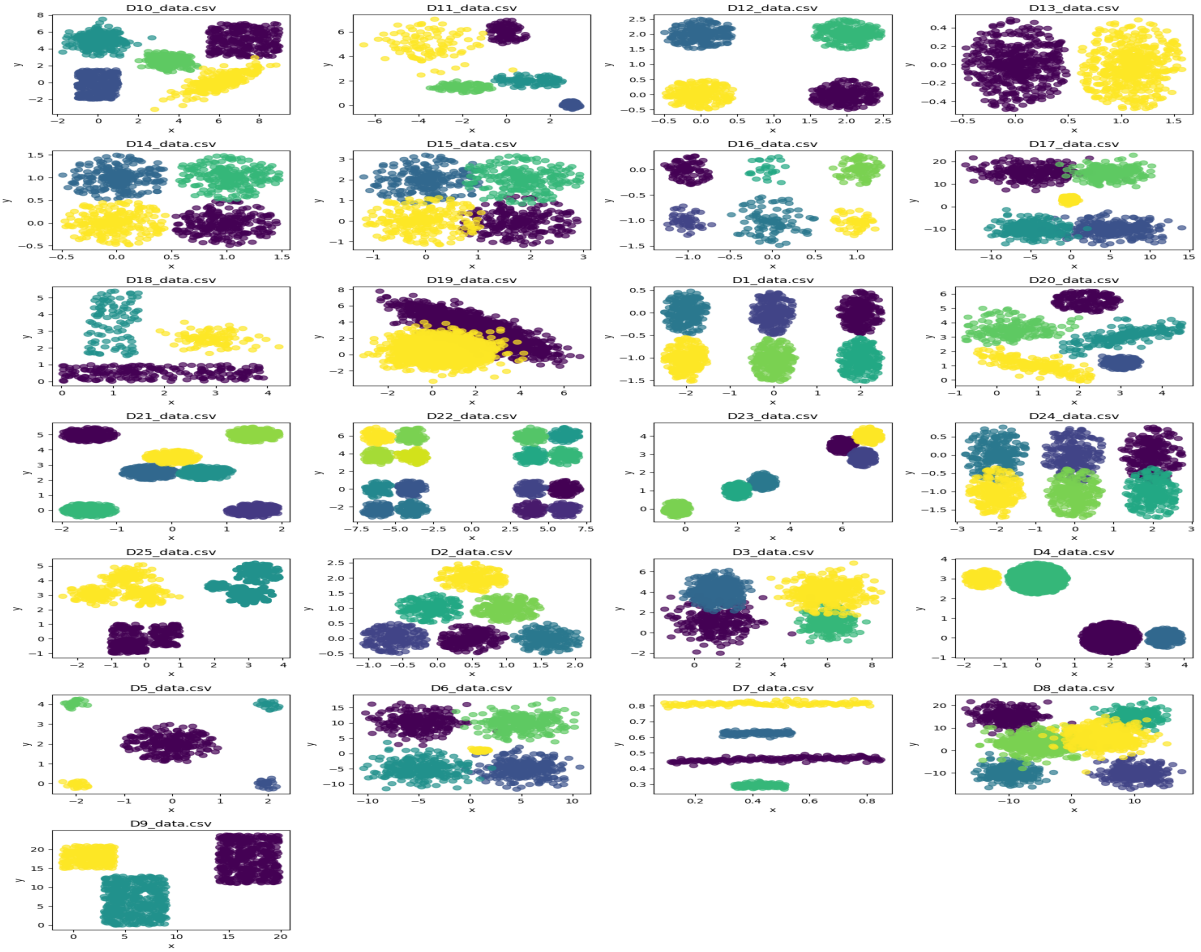
Table 1: Độ chính xác của thuật toán phân cụm dựa trên nhãn dữ liệu.

Độ chính xác của bộ dữ liệu này có phần tốt hơn so với dữ liệu gốc của bài báo tuy nhiên thì bộ dữ liệu thực tế có độ chính xác thấp nhưng vẫn ở mức chấp nhận được.

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	-	-

Table 2: Tham số của phân phối tiên nghiệm

Sử dụng lại bộ alpha theo kịch bản mà bài báo đã dùng dựa vì giá trị alpha này được dựa trên kinh nghiệm mà người dùng phân cụm để có và bộ dữ mới được dùng để đánh giá không có sự khác biệt nên sẽ không thay đổi quá nhiều. Chủ yếu vào các chỉ số phân cụm và số lượng cụm.



Hình 1: Bộ dữ liệu nhân tạo được gắn nhãn sẵn

Data	α	K1	K2	BCVI(CH)	CH	BCVI(SH)	SH	BCVI(DB)	DB
	a_1	-	-	4	8	4	6	4	2
D1	a_2	6	-	8	8	8	6	4	2
	a_3	-	-	7	10	6	6	4	2
	a_1	-	-	4	10	4	6	3	2
D2	a_2	6	-	10	10	8	6	3	2
	a_3	-	-	6	10	6	6	3	2
	a_1	-	-	4	10	4	4	4	2
D3	a_2	4	2	9	9	8	4	4	2
	a_3	-	-	7	9	5	4	4	2
	a_1	-	-	4	10	2	4	2	2
D4	a_2	4	2	10	10	10	10	2	2
	a_3	-	-	7	10	5	5	2	2
	a_1	-	-	3	10	4	5	2	2
D5	a_2	5	-	10	10	10	5	2	2
	a_3	-	-	7	10	5	5	2	2

Table 3:Bảng kết quả phân cụm theo dữ liệu nhân tạo.

Data	α	K1	K2	BCVI(CH)	CH	BCVI(SH)	SH	BCVI(DB)	DB
	a_1	-	-	3	3	4	3	3	2
Dry bean	a_2	7	-	10	6	8	3	6	2
	a_3	-	-	3	5	6	3	4	2

Table 4:Bảng kết quả phân cụm cứng với dữ liệu real world.

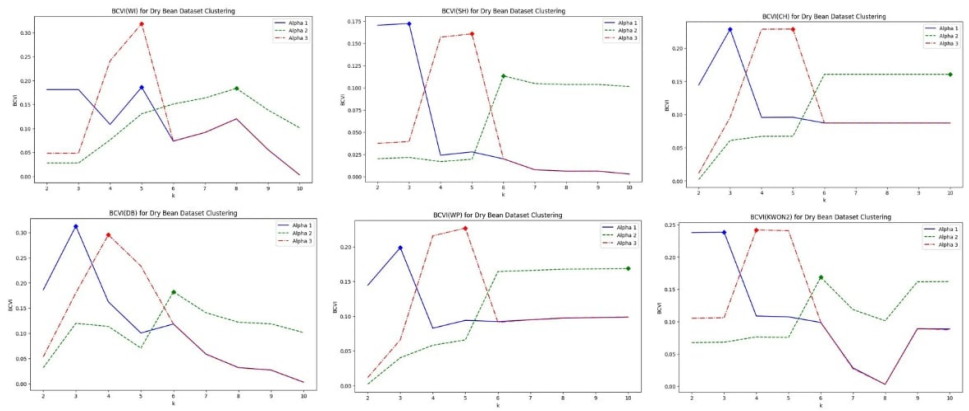


Figure 2: Các chỉ số CVI được áp dụng BCVI theo các α_1 α_2 α_3

Data	α	K1	K2	BCVI(KWON2)	KWON2	BCVI(WP)	WP
	a_1	-	-	4	2	4	4
D1	a_2	6	-	4	2	4	4
	a_3	-	-	4	2	4	4
	a_1	-	-	3	2	4	4
D2	a_2	6	-	3	2	4	4
	a_3	-	-	3	2	4	4
	a_1	-	-	4	2	4	4
D3	a_2	4	2	4	2	4	4
	a_3	-	-	4	2	4	4
	a_1	-	-	4	4	4	4
D4	a_2	4	2	4	4	4	4
	a_3	-	-	4	4	4	4
	a_1	-	-	4	4	3	2
D5	a_2	5	-	4	4	3	2
	a_3	-	-	4	4	3	2

Table 2: Bảng kết quả phân cụm mềm với dữ liệu nhân tạo.

5.1 Ứng dụng BCVI vào các bộ dữ liệu nhân tạo và dữ liệu thực

5.1.1 Trong dữ liệu nhân tạo

Trong phần này, dữ liệu nhân tạo giúp kiểm soát các yếu tố nhiễu và đánh giá hiệu suất của BCVI trong các điều kiện lý tưởng, từ đó giúp kiểm tra các đặc tính của BCVI. Điều chỉnh các kết quả sai: BCVI có thể điều chỉnh các kết quả phân cụm sai do các CVI cơ bản đưa ra đề xuất các lựa chọn phân cụm hợp lý hơn.

5.1.2 Trong dữ liệu thực tế

Với bộ dữ liệu *Dry bean dataset* với 7 loại đậu có đặc điểm tương đồng, được sử dụng để kiểm tra khả năng của BCVI trong môi trường thực tế. BCVI giúp xác định số lượng cụm chính xác (7 loại đậu) và cải thiện phân cụm khi CVI cơ bản không đạt yêu cầu.

5.1.3 Các giá trị của α

Ưu tiên các nhóm cụ thể, các giá trị α được chọn dựa trên mục tiêu người dùng mong muốn về số lượng nhóm. Ví dụ như: nhỏ, lớn, vừa phải. Các

giá trị α có thể điều chỉnh theo kích thước của dữ liệu và không nên lớn hơn nhiều kích thước dữ liệu.

5.2 Ứng dụng BCVI vào mô hình Detect Tumor Brain.

Trong phần này, chúng tôi trình bày ứng dụng của chỉ số BCVI trong việc phát hiện khối u não từ ảnh MRI, một ứng dụng quan trọng trong lĩnh vực y tế.

5.2.1 giới thiệu

Tập dữ liệu gồm 5712 bức ảnh về não (bao gồm : tumor và no tumor).Với định dạng ban đầu của ảnh là (512, 512, 3) pixels. Số 3 ở đây là dạng RGB (Red-Green-Blue).

5.2.2 Dữ liệu và tiền xử lý

Bộ dữ liệu MRI được sử dụng là tập hợp ảnh não chứa cả ảnh có khối u và ảnh bình thường. Các bước tiền xử lý bao gồm:

1. **Chuyển đổi định dạng:** Đưa tất cả các ảnh về định dạng chuẩn, đảm bảo kích thước ảnh đồng nhất (128x128 pixel).
2. **Chuẩn hóa dữ liệu:** Tất cả các pixel được chuẩn hóa về khoảng giá trị [0,1] để đảm bảo tính nhất quán khi huấn luyện mô hình.

5.2.3 Áp dụng thuật toán K-Means và các CVI

Phân chia ảnh thành K cụm (vùng) sao cho các pixel trong cùng cụm có đặc điểm tương đồng nhất. Xác định các vùng chứa cường độ sáng khác biệt (thường là khối u hoặc tổn thương não).

1. **Áp dụng K-means:**Toàn bộ ảnh MRI được phân cụm bằng cả thuật toán Kmeans.
2. **Tính toán CVI và áp dụng BCVI:** Sử dụng 1000 ảnh ngẫu nhiên để tính toán các CVI.Tiếp tục dùng BCVI lên các CVI đó để tính và đánh giá.
3. **Thiết lập kịch bản alpha:** theo kinh nghiệm của người dùng về bộ dữ liệu , thiết lập các kịch bản khác nhau. Khi áp dụng Kmeans với toàn bộ hình để minh họa việc phát hiện khối u, chúng tôi dùng k=5. Còn khi áp dụng Kmeans với 1000 hình ngẫu nhiên để tính CVI và áp dụng BCVI thì chúng tôi dùng k=8.

5.2.4 Kết quả và đánh giá

Hình ảnh minh họa sau khi áp dụng Kmeans vào toàn bộ data (với K=5)

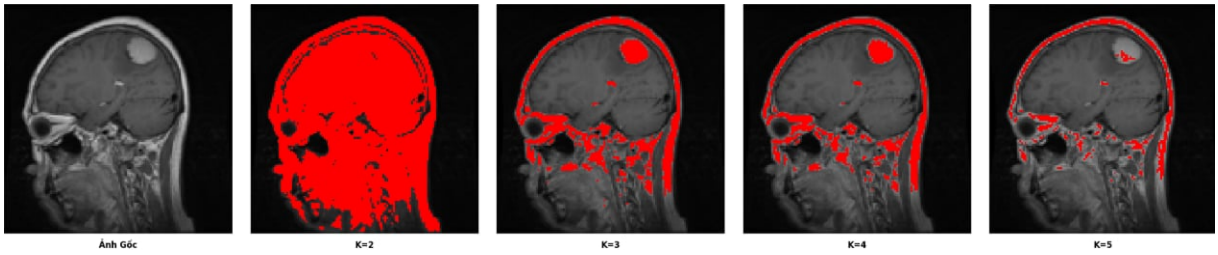


Figure 1: Ảnh phân đoạn khối u não với các giá trị K khác nhau trong thuật toán K-means.

Kết quả thực nghiệm sau khi áp dụng Kmeans và tính toán các CVI. Đây là các chỉ số thể hiện phân cụm và độ chính xác :

K	Inertia	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index	Starczewski Index	Wiroonsri Index
2	26296227.8378	0.7309	55827.0086	0.3958	125397.8447	162517.0332
3	9880553.4971	0.7505	87155.2644	0.4260	166192.1617	214486.9216
4	5680555.9299	0.7166	104001.0653	0.4668	166810.7235	226480.8955
5	3608324.5415	0.7126	124717.2935	0.4769	191747.4301	263212.0241
6	2519850.8685	0.6983	143284.2147	0.4880	210956.6441	295249.5027
7	1880161.8028	0.6905	160727.0104	0.4928	231775.4247	328013.4429
8	1450260.5362	0.6796	178374.6321	0.4969	251212.7654	361317.6248

Table 3: Kết quả trung bình các chỉ số theo K.

Kết quả sau khi áp dụng BCVI vào các CVI với alpha đã cho trước theo kịch bản :

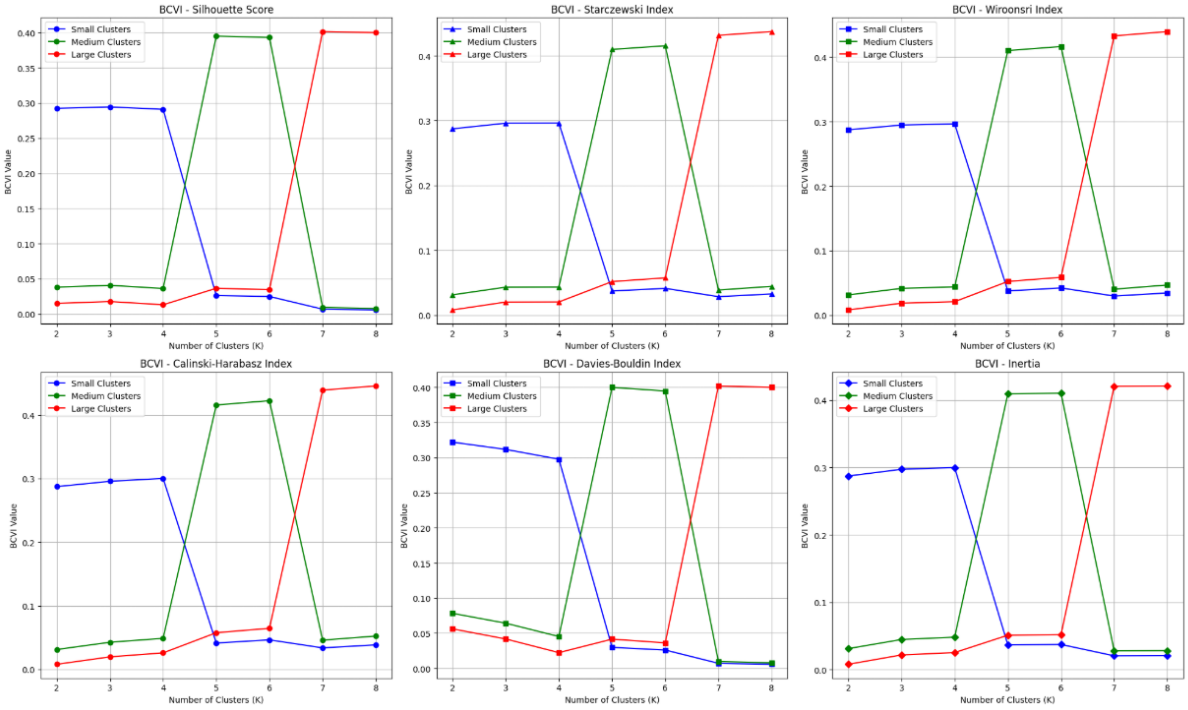


Figure 2: Các chỉ số CVI đã được áp dụng BCVI theo các alpha cho trước.

5.3 Hạn chế của nghiên cứu

Mặc dù BCVI đã cho thấy hiệu quả trong việc cải thiện độ chính xác khi xác định số cụm tối ưu, nhưng vẫn tồn tại một số hạn chế nhất định:

- **Phụ thuộc vào việc chọn giá trị α:** Giá trị của α có ảnh hưởng đáng kể đến kết quả phân cụm. Việc lựa chọn không phù hợp có thể dẫn đến kết quả phân cụm không chính xác. Ngoài ra, việc thiết lập giá trị α yêu cầu kiến thức chuyên môn từ người dùng, khiến việc áp dụng BCVI trở nên khó khăn đối với người không chuyên.
- **Độ nhạy với kích thước dữ liệu:** BCVI hoạt động tốt trên các tập dữ liệu vừa và nhỏ. Tuy nhiên, khi áp dụng trên các tập dữ liệu lớn với độ phân giải cao (như ảnh y tế MRI), việc tính toán CVI và BCVI có thể tốn nhiều tài nguyên tính toán và thời gian xử lý.
- **Ảnh hưởng từ chất lượng dữ liệu:** Các tập dữ liệu chứa nhiều nhiễu hoặc có sự phân tán không đồng nhất giữa các cụm có thể ảnh hưởng đến khả năng đánh giá chính xác số lượng cụm của BCVI. Đặc biệt, trong dữ liệu thực tế như MRI hoặc các bộ dữ liệu y tế, yếu tố này càng trở nên quan trọng.

- **Giới hạn với các chỉ số CVI cơ bản:** BCVI phụ thuộc vào các chỉ số CVI cơ bản như Calinski-Harabasz, Davies-Bouldin và Silhouette. Nếu các chỉ số này không đánh giá chính xác hoặc có độ nhiễu cao, kết quả của BCVI cũng có thể bị ảnh hưởng theo.
- **Không phù hợp cho mọi loại dữ liệu:** BCVI phù hợp hơn cho các bài toán phân cụm dữ liệu có cấu trúc rõ ràng. Với các loại dữ liệu phi cấu trúc hoặc dữ liệu có phân phối không đồng đều, BCVI có thể gặp khó khăn trong việc xác định số cụm hợp lý.

Kết luận

Trong nghiên cứu này, chúng tôi đã áp dụng thuật toán K-means và chỉ số xác thực cụm Bayesian Cluster Validity Index (BCVI) để phân cụm dữ liệu nhân tạo, dữ liệu số thực, ảnh y tế MRI não trong bài toán nhận dạng điểm khác thường của dữ liệu đối với dữ liệu số và phát hiện khối u não đối với dữ liệu MRI hình ảnh. Quá trình nghiên cứu bao gồm các bước tiền xử lý dữ liệu, chuẩn hóa ảnh và thử nghiệm trên các giá trị K khác nhau nhằm đánh giá độ chính xác của thuật toán phân cụm.

Kết quả cho thấy BCVI mang lại những lợi ích đáng kể khi được áp dụng kết hợp với K-means:

- BCVI cung cấp một phương pháp đáng tin cậy để xác định số cụm tối ưu K trong tập dữ liệu, giúp cải thiện khả năng phát hiện khối u và vùng bất thường trên ảnh MRI.
- So với các chỉ số truyền thống như Calinski-Harabasz, Davies-Bouldin và Silhouette Score, BCVI thể hiện tính linh hoạt hơn khi cho phép điều chỉnh dựa trên phân phối tiên nghiệm, kết hợp kiến thức chuyên môn từ người dùng.
- Trong thử nghiệm thực tế trên dữ liệu MRI não, BCVI đã chỉ ra rằng số lượng cụm tối ưu rơi vào khoảng $K = 4$, giúp phân tách rõ ràng giữa các vùng mô bình thường và vùng nghi ngờ có khối u.
- Chỉ số BCVI còn cho phép nhận diện và đánh giá độ tin cậy của các đỉnh cục bộ, từ đó cung cấp nhiều lựa chọn cho người dùng trong việc lựa chọn số cụm phù hợp.
- Các tập dữ liệu nhân tạo có thể được thiết kế để mô phỏng các trường hợp cụ thể, ví dụ như khi các CVI cơ bản đưa ra kết quả sai hoặc khi người dùng muốn có lựa chọn phân cụm khác với kết quả tối ưu mà CVI truyền thống đưa ra.
- BCVI có thể giúp xác định số lượng cụm tối ưu trong các tập dữ liệu thực tế phức tạp và linh hoạt hơn so với các CVI cơ bản, giúp cải thiện độ chính xác của phân cụm trong những tình huống không rõ ràng hoặc khi các CVI truyền thống không hoạt động hiệu quả.

Tuy nhiên, vẫn còn một số thách thức khi áp dụng BCVI vào các bài toán thực tế:

- Độ nhạy của BCVI phụ thuộc khá nhiều vào việc lựa chọn phân phối tiên nghiệm và kích thước tập dữ liệu.
- Đối với các tập dữ liệu có độ phức tạp cao như MRI não, việc điều chỉnh tham số α và lựa chọn chỉ số CVI gốc là cần thiết để đảm bảo kết quả chính xác.
- Dữ liệu thực tế có thể gây khó khăn cho BCVI khi đặc trưng giữa các nhóm dữ liệu rất tương đồng hoặc không đủ phân biệt. Điều này làm giảm hiệu quả của thuật toán phân cụm và có thể ảnh hưởng đến kết quả cuối cùng của BCVI.
- Mặc dù dữ liệu nhân tạo cho phép kiểm soát mọi yếu tố, nhưng chúng có thể không phản ánh chính xác các tình huống phức tạp mà dữ liệu thực tế mang lại. Điều này có thể khiến các kết quả từ dữ liệu nhân tạo không hoàn toàn phù hợp với các tình huống thực tế.

Hướng phát triển tiếp theo: Trong tương lai, chúng tôi dự định:

1. Thử nghiệm BCVI trên các tập dữ liệu y tế lớn hơn với độ phân giải cao hơn.
2. Tích hợp BCVI vào hệ thống hỗ trợ chẩn đoán y tế, giúp bác sĩ đưa ra quyết định chính xác hơn.

Kết quả từ nghiên cứu này cho thấy rằng BCVI là một công cụ hữu ích, mang lại giá trị trong việc xác định số cụm tối ưu và cải thiện độ chính xác của các mô hình phân cụm trong lĩnh vực y tế.

Tài liệu tham khảo

Preedasawakul, O., & Wiroonsri, N. (2025). A Bayesian cluster validity index. *Computational Statistics and Data Analysis*, 202, 108053. DOI: 10.1016/j.csda.2024.108053

Koklu, M., & Özkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174, 105507.
https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Brain Tumor MRI Dataset: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

https://github.com/O-PREEDASAWAKUL/FuzzyDatasets