**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

Trường công nghệ thông tin và truyền thông

A red and white logo

Description automatically generated

**BÁO CÁO NGHIÊN CỨU TỐT NGHIỆP 1**

The Fuzzy C-Means Clustering &

Semi-supervised Fuzzy C-Means Clustering

*Giảng viên hướng dẫn: Trần Đình Khang*

*Sinh viên thực hiện: Bùi Thế Hiếu - 20215047*

Contents

[**1.** **Tóm tắt** 2](#_Toc169780247)

[**2.** **Giới thiệu** 2](#_Toc169780248)

[**3.** **Cơ sở lí thuyết** 3](#_Toc169780249)

[**4.** **Phương pháp nghiên cứu** 5](#_Toc169780250)

[**5.** **Thực nghiệm và kết quả** 6](#_Toc169780251)

[**6.** **Thảo luận** 12](#_Toc169780252)

[**7.** **Bộ dữ liệu, mã nguồn** 14](#_Toc169780253)

1. **Tóm tắt**

* **Mục đích nghiên cứu:**

Nghiên cứu này nhằm mục đích tìm hiểu và đánh giá hiệu quả của hai kỹ thuật phân cụm mờ: Fuzzy C-Means (FCM) và Semi-Supervised Fuzzy C-Means (sSFCM) trong việc phân cụm dữ liệu.

* + FCM là một phương pháp phân cụm không giám sát
  + sSFCM là một biến thể bán giám sát, sử dụng thông tin nhãn để nâng cao độ chính xác của quá trình phân cụm.
* **Phương pháp nghiên cứu :**

Nghiên cứu này sử dụng cả hai phương pháp FCM và sSFCM để phân cụm một bộ dữ liệu mẫu. Quá trình nghiên cứu bao gồm các bước chuẩn bị, tiền xử lý dữ liệu và thực hiện phân cụm. Để đánh giá hiệu quả của các phương pháp, các tiêu chí như độ chính xác, tính nhất quán và hiệu suất tính toán được xem xét. Ngôn ngữ được sử dụng là Python và các thư viện hỗ trợ đã được sử dụng để triển khai và đánh giá.

* **Kết quả chính:**

Nghiên cứu cho thấy rằng khi phân cụm dữ liệu có nhãn, sSFCM đạt được độ chính xác cao hơn và tính nhất quán tốt hơn so với FCM. Tuy nhiên, để đạt được những kết quả này, SFCM cần sử dụng nhiều tài nguyên tính toán hơn và có thời gian thực thi lâu hơn so với FCM.

* **Kết luận chính:**

Nghiên cứu kết luận rằng sSFCM là phương pháp phân cụm hiệu quả hơn FCM khi có thông tin nhãn, đặc biệt phù hợp cho các ứng dụng đòi hỏi độ chính xác cao. Tuy nhiên, cần xem xét chi phí tính toán và thời gian thực thi khi lựa chọn phương pháp phân cụm cho từng ứng dụng cụ thể.

1. **Giới thiệu**

* **Bối cảnh:**

Fuzzy C-Means (FCM) là một phương pháp phân cụm mờ nổi tiếng, cho phép mỗi điểm dữ liệu thuộc về nhiều cụm với mức độ khác nhau. Tuy nhiên, FCM là phương pháp không giám sát và không sử dụng thông tin nhãn trong quá trình phân cụm. Semi-Supervised Fuzzy C-Means (sSFCM), một biến thể của FCM, tích hợp thông tin nhãn để nâng cao độ chính xác và hiệu quả của quá trình phân cụm. Việc đánh giá và so sánh hiệu quả giữa FCM và sSFCM có vai trò quan trọng trong việc chọn lựa phương pháp phù hợp cho các ứng dụng cụ thể.

* **Mục tiêu:**

Nghiên cứu này đặt mục tiêu so sánh hiệu quả của FCM và sSFCM trong việc phân cụm dữ liệu, nhằm xác định phương pháp nào phù hợp hơn cho các tập dữ liệu cụ thể.

* **Phạm vi**:

Nghiên cứu này tập trung vào việc áp dụng và so sánh FCM và sSFCM trên mẫu dữ liệu từ UCI Machine Learning Repository

1. **Cơ sở lí thuyết**

* **Giới thiệu về Phân cụm (Clustering):**

Phân cụm là một kỹ thuật quan trọng trong lĩnh vực khai thác dữ liệu và học máy, được sử dụng để tự động nhóm các đối tượng dữ liệu vào các nhóm có tính chất tương đồng nhau. Mục đích chính của phân cụm là tạo ra các nhóm (cụm) sao cho các đối tượng trong cùng một nhóm có tính chất gần nhau và đồng nhất, trong khi đối tượng ở các nhóm khác nhau có tính chất khác biệt.

Phân cụm giúp phân tích và hiểu sâu hơn về cấu trúc của dữ liệu mà không cần sự can thiệp của con người để định nghĩa các nhóm trước. Kỹ thuật này có nhiều ứng dụng rộng rãi trong thực tế như phân tích thị trường, nhận dạng mẫu, phân tích hình ảnh, xử lý ngôn ngữ tự nhiên, và nhiều lĩnh vực khoa học khác.

* **Fuzzy C-Means (FCM)**:

Fuzzy C-Means (FCM) là một trong những phương pháp phân cụm mờ quan trọng trong lĩnh vực khai thác dữ liệu và học máy. Đây là một phương pháp giúp nhóm các điểm dữ liệu vào các cụm dựa trên độ tương đồng của chúng với các trung tâm cụm. FCM cho phép mỗi điểm dữ liệu thuộc về một hoặc nhiều cụm với mức độ thuộc về mỗi cụm được biểu thị bằng giá trị mờ (fuzzy membership)

* Thuật toán FCM tối thiểu hóa hàm mục tiêu Jm:

A black and white text

Description automatically generated

Trong đó:

* + - xk​ là điểm dữ liệu k
    - vi​ là tâm cụm i
    - uik là mức độ thành viên của xk​ trong cụm i
    - m là số mờ
  + Các bước thực hiện thuật toán FCM:
    - Khởi tạo ma trận thành viên U ngẫu nhiên sao cho các ràng buộc được thỏa mãn
    - Tính toán các trung tâm cụm bằng công thức:

A black text with a white background

Description automatically generated with medium confidence

* + - Cập nhật ma trận thành viên U bằng công thức:

A math equation with numbers

Description automatically generated with medium confidence

* + - Lặp lại các bước 2 và 3 cho đến khi hội tụ
* **Semi-Supervised Fuzzy C-Means (sSFCM)**:

Semi-Supervised Fuzzy C-Means (sSFCM) là một biến thể của phương pháp Fuzzy C-Means (FCM), được sử dụng để phân cụm dữ liệu khi có sẵn thông tin nhãn từ trước. sSFCM tích hợp thông tin nhãn vào quá trình phân cụm để cải thiện độ chính xác và tính nhất quán của kết quả phân cụm.

* Thuật toán sSFCM tối thiểu hóa hàm mục tiêu Jm:

A black text on a white background

Description automatically generated

Trong đó:

* + - xk​ là điểm dữ liệu k
    - vi​ là tâm cụm i
    - uki là mức độ thành viên của xk​ trong cụm i
    - uki là mức độ thành viên giám sát của xk​ trong cụm i
    - m là số mờ
  + Các bước thực hiện thuật toán sSFCM:
    - Khởi tạo ma trận thành viên U ngẫu nhiên sao cho các ràng buộc được thỏa mãn
    - Khởi tạo ma trận giám sát ū dựa vào nhãn của dữ liệu
    - Tính toán các trung tâm cụm bằng công thức:

A black and white image of a mathematical equation

Description automatically generated

* + - Cập nhật ma trận thành viên U bằng công thức:

A math equations on a white background

Description automatically generated

* + - Lặp lại các bước 3 và 4 cho đến khi hội tụ
* **So sánh giữa FCM và sSFCM**:

FCM (Fuzzy C-Means) và sSFCM (Semi-Supervised Fuzzy C-Means) là hai phương pháp quan trọng trong lĩnh vực phân cụm dữ liệu, nhưng có những điểm khác biệt đáng chú ý:

Đặc điểm cơ bản:

FCM: Là phương pháp phân cụm mờ không giám sát. Nó dựa trên khoảng cách giữa các điểm dữ liệu và các trung tâm cụm để gán mỗi điểm vào các cụm với một mức độ (giá trị mờ) khác nhau.

sSFCM: Là biến thể bán giám sát của FCM. Ngoài việc sử dụng khoảng cách, sSFCM còn tích hợp thông tin nhãn từ các điểm dữ liệu đã biết trước để cải thiện độ chính xác và tính nhất quán của quá trình phân cụm.

* + Sử dụng thông tin nhãn:
    - FCM: Không sử dụng thông tin nhãn trong quá trình phân cụm, do đó thường áp dụng trong các trường hợp mà không có sẵn thông tin nhãn hoặc khi không cần độ chính xác cao nhất.
    - sSFCM: Sử dụng thông tin nhãn để hướng dẫn quá trình phân cụm. Điều này giúp cải thiện đáng kể độ chính xác của kết quả phân cụm, đặc biệt là trong các tập dữ liệu có nhãn rõ ràng và tin cậy.
* Hiệu suất và tính toán:
  + - FCM: Thường có tính toán đơn giản hơn so với sSFCM vì không phải xử lý thông tin nhãn. Điều này có thể dẫn đến thời gian thực thi nhanh hơn và yêu cầu ít tài nguyên tính toán hơn.
    - sSFCM: Yêu cầu nhiều tài nguyên tính toán hơn do phải tích hợp và xử lý thông tin nhãn. Việc này có thể làm gia tăng thời gian thực thi và yêu cầu bộ nhớ lớn hơn so với FCM.
* Tính chất của kết quả phân cụm:
  + - FCM: Cho phép mỗi điểm dữ liệu có mức độ thuộc về nhiều cụm khác nhau, biểu thị bằng giá trị mờ. Điều này phù hợp với dữ liệu có tính mờ mịn và không rõ ràng.
    - sSFCM: Tính chất này vẫn được giữ lại từ FCM, nhưng kết quả phân cụm thường có tính nhất quán cao hơn và chính xác hơn do sử dụng thông tin nhãn.
* **Tổng kết:**

FCM và sSFCM đều có vai trò quan trọng trong phân cụm dữ liệu, mỗi phương pháp có ưu điểm và hạn chế riêng. Lựa chọn phương pháp phù hợp sẽ phụ thuộc vào tính chất của dữ liệu và mục đích cụ thể của bài toán, có thể là cải thiện độ chính xác phân cụm hoặc giảm thiểu chi phí tính toán.

1. **Phương pháp nghiên cứu**

* **Thiết kế nghiên cứu:**

Nghiên cứu này nhằm so sánh hiệu quả của hai phương pháp phân cụm mờ: Fuzzy C-Means (FCM) và Semi-Supervised Fuzzy C-Means (sSFCM). Quá trình nghiên cứu bao gồm các giai đoạn chính: thu thập dữ liệu, tiền xử lý dữ liệu, thực hiện phân cụm bằng FCM và sSFCM, sau đó đánh giá kết quả dựa trên các tiêu chí như độ chính xác, tính nhất quán và hiệu suất tính toán.

* **Dữ liệu và nguồn dữ liệu**

Bộ dữ liệu Iris về các loài hoa ([Iris - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/53/iris))

Đây là một trong những bộ dữ liệu sớm nhất được sử dụng trong văn xuôi về các phương pháp phân loại và rất phổ biến trong thống kê và học máy. Bộ dữ liệu này bao gồm 3 lớp, mỗi lớp có 50 ví dụ, trong đó mỗi lớp tương ứng với một loại cây hoa Iris.

* **Quy trình phân cụm:**
  + Bước 1: Chuyển đổi dữ liệu từ định dạng CSV sang ma trận thuộc tính.
  + Bước 2: Thực hiện phân cụm bằng Fuzzy C-Means (FCM):
    - Khởi tạo các tâm cụm ban đầu.
    - Tính toán các giá trị thành viên cho mỗi điểm dữ liệu.
    - Cập nhật các tâm cụm dựa trên giá trị thành viên.
    - Lặp lại quá trình cho đến khi đạt điều kiện hội tụ.
  + Bước 3: Thực hiện phân cụm bằng Semi-Supervised Fuzzy C-Means (sSFCM):
    - Sử dụng thông tin nhãn để điều chỉnh các giá trị thành viên.
    - Khởi tạo các tâm cụm ban đầu theo dạng có giám sát.
    - Tính toán và cập nhật các giá trị thành viên có giám sát.
    - Lặp lại quá trình cho đến khi đạt điều kiện hội tụ.
  + Bước 4: Đánh giá kết quả:
    - Sử dụng các tiêu chí như độ chính xác, tính nhất quán và thời gian tính toán để so sánh kết quả của FCM và sSFCM.
    - Phân tích và trình bày kết quả thông qua các bảng biểu và hình vẽ.
* **Các chỉ số đánh giá:**
  + **Độ đo trong:**
    - **Partition Coefficient:** Chỉ số này đo lường mức độ mờ của phân cụm. Giá trị của Fc dao động từ 0 đến 1. Giá trị càng cao thì mức độ mờ càng thấp và phân cụm càng tốt.
    - **Entropy:** Chỉ số entropy đo lường mức độ lộn xộn hoặc không chắc chắn trong phân cụm. Giá trị thấp của H chỉ ra rằng các cụm ít bị chồng chéo hơn và rõ ràng hơn.
    - **Calinski-Harabasz:** Calinski-Harabasz (VRC) được thiết kế để đánh giá chất lượng của phân cụm dựa trên tỷ lệ giữa biến động giữa các cụm và biến động trong các cụm. Chỉ số này càng cao thì phân cụm càng tốt.
  + **Độ đo ngoài:**
    - **Rand Index:** Chỉ số Rand đo lường tỷ lệ các cặp đối tượng được phân loại giống nhau trong cả hai phân vùng, nằm trong khoảng từ 0 đến 1. Chỉ số Rand càng gần 1, phân cụm càng chính xác khi so với phân cụm tham chiếu.
    - **Adjusted Rand Index:** Chỉ số Adjusted Rand là một phiên bản cải tiến của chỉ số Rand, nhằm điều chỉnh tác động của các phân cụm ngẫu nhiên, nằm trong khoảng từ -1 đến 1, nhưng thường nằm trong khoảng từ 0 đến 1. ARI cao hơn (gần 1) cho thấy phân cụm tốt hơn khi so với các phân cụm ngẫu nhiên.
    - **Jaccard Coefficient:** Hệ số Jaccard đo lường sự tương đồng giữa hai tập hợp bằng cách tính tỷ lệ giữa kích thước của giao của chúng với kích thước của hợp của chúng. Hệ Số Jaccard nằm trong khoảng từ 0 đến 1, càng gần 1, phân cụm càng chính xác khi so với phân cụm tham chiếu.
* **Các công nghệ và phần mềm sử dụng:**
  + **Python:** Ngôn ngữ lập trình chính được sử dụng cho việc phân cụm và phân tích dữ liệu.
  + **NumPy và Pandas:** Hỗ trợ xử lý và phân tích dữ liệu.
  + **Matplotlib:** Tạo biểu đồ đánh giá kết quả.
  + **Streamlit:** Tạo giao diện.

1. **Thực nghiệm và kết quả**

* **Mô tả dữ liệu sử dụng:**

Bộ dữ liệu Iris đã được sử dụng trong bài báo kinh điển năm 1936 của R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", và cũng có thể được tìm thấy trên Kho dữ liệu Học máy của Đại học California Irvine (UCI).

Bộ dữ liệu bao gồm ba loài hoa Iris, mỗi loài có 50 mẫu cùng với một số thuộc tính về từng loài hoa. Một loài hoa có thể phân tách tuyến tính với hai loài hoa còn lại, nhưng hai loài hoa còn lại không thể phân tách tuyến tính với nhau.

Các cột trong bộ dữ liệu này bao gồm:

* Id
* SepalLengthCm (Chiều dài đài hoa, đơn vị: cm)
* SepalWidthCm (Chiều rộng đài hoa, đơn vị: cm)
* PetalLengthCm (Chiều dài cánh hoa, đơn vị: cm)
* PetalWidthCm (Chiều rộng cánh hoa, đơn vị: cm)
* Species (Loài hoa)
* **Các bước tiền xử lí dữ liệu**
* Đọc flie dữ liệu
* Loại bỏ các cột không cần thiết (id), khởi tạo ma trận ū dựa trên cột Species

A screen shot of a computer code

Description automatically generated

* **Tiến hành phân cụm**
* **FCM:**

+ Khởi tạo ma trận thành viên và tâm cụm:

A computer screen with text

Description automatically generated

+ Cập nhật tâm cụm:

A black screen with white text

Description automatically generated

+ Cập nhật ma trận thành viên:

A computer code on a black background

Description automatically generated

+ Tính khoảng cách từ ma trận thành viên mới đến tâm cụm mới

A screen shot of a computer code

Description automatically generated

+ Lặp lại cho đến khi hội tụ:

A screen shot of a computer program

Description automatically generated

* **sSFCM:**

+ Khởi tạo ma trận thành viên và tâm cụm:

A black screen with white text

Description automatically generated

+ Cập nhật ma trận thành viên:

A computer code with white text

Description automatically generated

+ Cập nhật tâm cụm:

A computer code on a black background

Description automatically generated

+ Tính khoảng cách từ ma trận thành viên mới đến tâm cụm mới:

A black background with white text

Description automatically generated

+ Lặp lại cho đến khi hội tụ:

A screen shot of a computer program

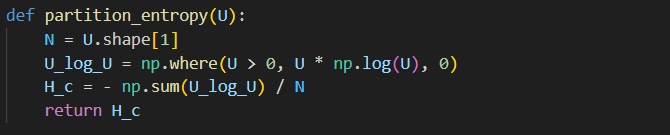
Description automatically generated

* **Các chỉ số đánh giá:**
* Độ đo trong:
  + Partition Coefficient

A black screen with white text

Description automatically generated

* + Entropy



* + Calinski-Harabasz

A computer screen shot of a program code

Description automatically generated

* Độ đo ngoài:

Tính toán các tham số

A computer screen shot of a program code

Description automatically generated

* + Rand Index

A black background with white text

Description automatically generated

* + Adjusted Rand Index

A computer screen with text

Description automatically generated

* + Jaccard Coefficient

**A black screen with white text

Description automatically generated**

* **Thiết kế giao diện:**
* File đầu vào
* Các tham số: số tâm cụm, hệ số mờ, epsilon có thể điều chỉnh
* Lựa chọn phương pháp phân cụm
* Hiển thị kết quả:
  + Tâm cụm
  + Các chỉ số đánh giá

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* **Đánh giá kết quả của 2 phương pháp:**
* **Độ chính xác:** Kết quả cho thấy sSFCM thường cho độ chính xác cao hơn so với FCM khi sử dụng các độ đo ngoài như Rand Index, Adjusted Rand Index, Jaccard Coefficient. Khi dữ liệu có sự can thiệp từ thông tin nhãn giám sát, sSFCM có xu hướng tạo ra các cụm rõ ràng hơn và tách biệt hơn. Tuy nhiên sSFCM tỏ ra yếu thế hơn khi sử dụng các độ đo trong như Partition Coefficient, Entropy hay Calinski-Harabasz.
* **Tính nhất quán:** Kết quả cho thấy FCM thường có tính nhất quán cao hơn so với sSFCM. Điều này có thể do có sự can thiệp từ nhãn giám sát, các hạn chế của nghiên cứu này và cần được xem xét thêm.
* **Thời gian tính toán (Computation Time):** thời gian tính toán của FCM và sSFCM có thể khác nhau và phụ thuộc vào nhiều yếu tố như kích thước của tập dữ liệu, độ phức tạp của thuật toán, và mức độ can thiệp từ thông tin nhãn giám sát. sSFCM thường yêu cầu nhiều thời gian tính toán hơn do tính phức tạp của quá trình tính toán ma trận membership và cập nhật các tâm cụm bán giám sát. Tuy nhiên, sự khác biệt này không đáng kể trên các tập dữ liệu nhỏ và đơn giản.

1. **Thảo luận**

* **Ý nghĩa của kết quả:** Nghiên cứu này nhấn mạnh vai trò của thông tin nhãn giám sát trong việc cải thiện độ chính xác và tính nhất quán của phương pháp phân cụm sSFCM so với FCM. Tuy nhiên, cần cân nhắc đến chi phí tính toán khi áp dụng sSFCM, đặc biệt là đối với các ứng dụng yêu cầu xử lý dữ liệu lớn và thời gian thực.
* **Hạn chế của nghiên cứu:**
  + Giới hạn của các chỉ số đánh giá: Nghiên cứu sử dụng các chỉ số được giới thiệu phía trên để đánh giá độ chính xác và tính nhất quán của phương pháp phân cụm. Tuy nhiên, các chỉ số này có thể không phản ánh đầy đủ các đặc trưng của dữ liệu thực tế, và việc áp dụng chúng cần phải cân nhắc kỹ lưỡng để đảm bảo tính đáng tin cậy của kết quả.
  + Giới hạn của thông tin nhãn giám sát: Sự hiệu quả của sSFCM phụ thuộc mạnh vào chất lượng và độ chính xác của thông tin nhãn giám sát được sử dụng. Nếu thông tin nhãn không đủ chính xác hoặc thiếu sót, kết quả phân cụm có thể bị sai lệch. Điều này đặt ra thách thức trong việc áp dụng sSFCM cho các tập dữ liệu thực tế có độ phức tạp và đa dạng cao.
  + Khả năng mở rộng và áp dụng: Nghiên cứu có thể hạn chế trong việc áp dụng và kiểm chứng trên nhiều loại dữ liệu khác nhau từ nhiều lĩnh vực khác nhau. Việc mở rộng nghiên cứu để bao gồm nhiều bộ dữ liệu đa dạng có thể cần thiết để đánh giá tính tổng quát và ứng dụng thực tế của các phương pháp phân cụm này.
  + Thời gian tính toán và chi phí tính toán: Sử dụng sSFCM có thể đòi hỏi nhiều tài nguyên tính toán hơn và thời gian thực thi lâu hơn so với FCM, đặc biệt là trên các tập dữ liệu lớn và phức tạp. Việc này có thể là một hạn chế đối với các ứng dụng yêu cầu phản hồi nhanh và xử lý dữ liệu lớn.
* **Các yếu tố ảnh hưởng đến kết quả:**
  + Kích thước tập dữ liệu: Kích thước của tập dữ liệu có thể ảnh hưởng đến hiệu suất của cả hai phương pháp. Các tập dữ liệu lớn thường đòi hỏi thời gian tính toán lâu hơn và tài nguyên tính toán cao hơn để xử lý.
  + Số lượng cụm: Số lượng cụm được chỉ định trước có thể ảnh hưởng đến độ chính xác của phân cụm. Việc lựa chọn số lượng cụm phù hợp là một yếu tố quan trọng để đạt được kết quả tốt nhất.
  + Hệ số mờ (fuzziness coefficient): Hệ số mờ trong FCM và sSFCM ảnh hưởng đến mức độ mờ của các cụm. Giá trị của hệ số này cần được điều chỉnh phù hợp để tối ưu hóa kết quả phân cụm.
  + Chất lượng của thông tin nhãn: Đối với sSFCM, chất lượng và độ tin cậy của thông tin nhãn là yếu tố quyết định đến hiệu quả phân cụm. Thông tin nhãn không chính xác có thể dẫn đến kết quả phân cụm kém chính xác.
  + Đặc điểm của dữ liệu: Tính phân tán, độ chi tiết, và độ phức tạp của dữ liệu có thể ảnh hưởng đến khả năng của thuật toán trong việc phân cụm chính xác và hiệu quả.
  + Các tham số thuật toán: Các tham số như số lần lặp, điều kiện dừng, và phương pháp khởi tạo các tâm cụm cũng có thể ảnh hưởng đến kết quả của phương pháp phân cụm.
* **Đề xuất hướng nghiên cứu tiếp theo:**
  + Tối ưu hóa sSFCM: Nghiên cứu và áp dụng các kỹ thuật tối ưu hóa để giảm chi phí tính toán và thời gian thực thi của sSFCM. Các phương pháp như sử dụng các thuật toán song song, phân tán, hoặc các kỹ thuật tối ưu hóa thuật toán có thể giúp cải thiện hiệu suất của sSFCM trên các tập dữ liệu lớn và phức tạp.
  + Nâng cao chất lượng nhãn: Phát triển các phương pháp tiền xử lý và xử lý dữ liệu để cải thiện chất lượng của thông tin nhãn trước khi áp dụng vào sSFCM. Các phương pháp này có thể bao gồm xử lý nhiễu, phân đoạn dữ liệu, và xây dựng mô hình dự đoán để tăng độ chính xác và tin cậy của thông tin nhãn.
  + Kiểm chứng trên nhiều bộ dữ liệu: Áp dụng và kiểm chứng FCM và sSFCM trên nhiều bộ dữ liệu khác nhau từ nhiều lĩnh vực khác nhau. Việc đánh giá tính tổng quát và hiệu quả của các phương pháp phân cụm trên các bộ dữ liệu đa dạng sẽ giúp khẳng định và mở rộng ứng dụng của chúng trong thực tế.
  + Phát triển các phương pháp đánh giá mới: Nghiên cứu và phát triển các phương pháp đánh giá hiệu quả hơn cho các phương pháp phân cụm mờ, đặc biệt là sSFCM. Các chỉ số mới có thể cân nhắc đến cả yếu tố can thiệp của thông tin nhãn giám sát để đánh giá hiệu quả của sSFCM một cách chính xác hơn.
  + Áp dụng trong các lĩnh vực ứng dụng cụ thể: Nghiên cứu ứng dụng của FCM và sSFCM trong các lĩnh vực như y tế, marketing, và khoa học dữ liệu để đánh giá khả năng áp dụng thực tiễn và tiềm năng của các phương pháp này trong các bối cảnh ứng dụng cụ thể.

1. **Bộ dữ liệu, mã nguồn**

Github: https://github.com/hiewbt/GR1