**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**PROJECT I**

**Thuật toán phân cụm dữ liệu mờ bán giám sát**

**LÊ THANH TÙNG**

[tung.lt194205@sis.hust.edu.vn](mailto:tung.lt194205@sis.hust.edu.vn)

**LÊ TIẾN DŨNG**

dung.lt194027@sis.hust.edu.vn

**HOÀNG HỮU PHÚC**

phuc.hh194818@sis.hust.edu.vn

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | PGS. TS. Trần Đình Khang  Chữ ký của GVHD |
|  |  |
| **Viện:** | CNTT & TT |

**HÀ NỘI, 12/2021**

Mục lục

[LỜI CẢM ƠN 3](#_Toc93511724)

[TÓM TẮT 4](#_Toc93511725)

[NỘI DUNG 5](#_Toc93511726)

[I. Giới thiệu bài toán 5](#_Toc93511727)

[II. Cơ sở lý thuyết 5](#_Toc93511728)

[1. Phân cụm mờ bán giám sát thông thường (sSFCM) 5](#_Toc93511729)

[2. Phân cụm mờ bán giám sát với hệ số m thay đổi (sSMC-FCM) 8](#_Toc93511730)

[3. Đánh giá các thuật toán phân cụm mờ sử dụng độ đo ngoài 11](#_Toc93511731)

[III. Xây dựng chương trình 12](#_Toc93511732)

[1. Cài đặt chương trình 12](#_Toc93511733)

[2. Giao diện chương trình 13](#_Toc93511734)

[3. Các bước chạy chương trình 13](#_Toc93511735)

[IV. Kết quả 16](#_Toc93511736)

[1. Các thuật toán khi chạy 1 pha 16](#_Toc93511737)

[2. Các thuật toán khi chạy 2 pha 17](#_Toc93511738)

[TỔNG KẾT 21](#_Toc93511739)

[TÀI LIỆU THAM KHẢO 22](#_Toc93511740)

# LỜI CẢM ƠN

Trong suốt quá trình học và làm đồ án Project I, chúng em đã luôn được sự quan tâm, giúp đỡ tận tình của thầy và các bạn cùng viện CNTT&TT

Lời đầu tiên, em xin cảm ơn viện CNNT&TT đã tạo cơ hội cho em hoàn thành đồ án lần này

Đặc biệt, chúng xin cảm ơn **Pgs.TS Trần Đình Khang**, thầy đã tận tình hướng dẫn chúng em suốt kỳ học 20211 để giúp chúng em hoàn thành đủ và tốt nội dung đồ án

Cuối cùng thay mặt cả nhóm cảm ơn tất cả các thành viên trong nhóm đã đóng góp công sức để hoàn thành các nhiệm vụ được giao

# TÓM TẮT

Phân cụm là một phương pháp học máy không được giám sát với nhiều ứng dụng thực tế đã thu thập được nhiều quan điểm nghiên cứu rộng rãi. Chúng ta có 2 phương pháp phân cụm: phân cụm có giám sát và phân cụm không giám sát. Phân cụm có giám sát có thể đạt được kết quả rất thực tế nhưng lại không thể xử lý khi dữ liệu quá lớn. Mặt khác, phân cụm không giám sát có thể xử lý rất nhiều dữ liệu nhưng phương pháp này khá phức tạp và kết quả đôi khi trông hơi lạ.

Do đó gần đây, nghiên cứu phân cụm bán giám sát đã được nghiên cứu. Phương pháp phân cụm này cả lợi thế của cả hai phương pháp trên, đó chính là kết quả vừa thực tế, chi phí thấp và thời gian tính toán ngắn. Nội dung của báo cáo này tập trung vào phân tích các thuật toán phân cụm bán giám sát trên sự phân cụm mờ, tức là một số thành viên đã được coi như là được giám sát.

Để thực hiện đề tài này, ngôn ngữ lập trình được sử dụng là Python Ưu điểm của ngôn ngữ là làm việc với các tập dữ liệu và các phép biến đổi toán học hiệu quả

# NỘI DUNG

## Giới thiệu bài toán

Gần đây, các phương pháp phân cụm tự động đang dần trở nên quan trọng hơn bởi vì độ phức tạp cũng như kích thước của bài toán ngày càng tăng lên. Phân cụm các mẫu đang thu hút sự quan tâm của rất nhiều nhà nghiên cứu không chỉ xét về nền tảng toán học mà còn cả các ứng dụng kỹ thuật. Chúng ta có 2 phương pháp phân cụm: phân cụm có giám sát và phân cụm không giám sát.

Trong phân cụm có giám sát, một số dữ liệu được giám sát trước và việc phân cụm được thực hiện dựa trên dữ liệu được giám sát trước đó. Vì dữ liệu được giám sát được cung cấp nên kết quả của phương pháp này trở nên tự nhiên hơn phương pháp phân cụm không giám sát và thuật toán cũng được đơn giản hoá. Tuy nhiên, dữ liệu càng lớn thì việc đưa ra được nhiều dữ liệu đã được giám sát từ trước là ít khả thi hơn.

Trong phân cụm không giám sát, còn được gọi là phân cụm, dữ liệu được phân cụm vào trong các cụm trừ các tiêu chuẩn bên ngoài. Bởi vì việc phải loại trừ các tiêu chuẩn bên ngoài, các thuật toán trở nên phức tạp hơn. Tuy nhiên, chúng ta có thể nói rằng việc xây dựng các thuật toán phân cụm không giám sát có ý nghĩa hơn các thuật toán có giám sát trong lĩnh vực phân cụm mẫu từ việc xem xét rằng hiện nay chúng ta phải xử lý rất nhiều dữ liệu, khi đó việc cung cấp dữ liệu đã được giám sát rất là khó.

Do đó gần đây, nghiên cứu phân cụm bán giám sát đã được nghiên cứu. Phân cụm bán giám sát có thuộc tính của cả 2 phương pháp phân cụm trên, đó chính là cả những dữ liệu được đánh nhãn và những dữ liệu không được đánh nhãn đều được phân cụm trong phân cụm bán giám sát và do đó, nó trở nên khả thi hơn để đạt được kết quả thực tế hơn phân cụm giám sát và phân cụm không giám sát.

Trong đồ án này, chúng em sẽ trình bày một thuật toán phân cụm mờ bán giám sát mới. Thuật toán này được xây dựng bằng cách cho vào một số độ thuộc của dữ liệu vào phân cụm mờ. Thuật toán này có thể đạt được kết quả phân cụm thực tế và thuật toán cũng khá đơn giản.

## Cơ sở lý thuyết

### Phân cụm mờ bán giám sát thông thường (sSFCM)

Trong phân cụm mờ bán giám sát, một số độ thuộc của một phần dữ liệu vào một số cụm đã được cho trước. Độ thuộc này được gọi là độ thuộc có giám sát và được biểu thị là ***ik***.

**\*Bài toán:**

Bài toán cơ bản chính là tìm ra cách để phân cụm n dữ liệu:

vào c cụm:

với tâm cụm là:

trong không gian mẫu . Khoảng cách Euclid được định nghĩa trong không gian là:

chính là tỉ lệ phần tử thuộc cụm Mục đích của việc phân cụm chính là đạt được ma trận U:

***ik***  chính là độ thuộc được giám sát thể hiện khả năng thuộc về cụm và

được cho trước. Và điều kiện của ***ik*** là:

(1)

**\* Cơ sở lý thuyết**

Chúng ta sẽ tìm cách để làm giảm giá trị của hàm J(U,V), hàm được tạo ra bằng cách đưa vào hàm mục tiêu của sFCM.

(2)

với điều kiện là:

(3)

Do J là hàm lồi nên chúng ta có thể tìm ra kết quả từ phương trình

Từ đó ta có:

Suy ra:

(4)

Chúng ta phải xét đến 2 trường hợp: m > 1 và m=1. Trong đồ án này, chỉ xét đến m>1.

Thay vì xét (2), chúng ta sẽ tìm cách làm giảm giá trị của hàm J’(U,V)

(5)

khi điều kiện của và lần lượt là (3) và (1)

Chúng ta xét hàm Larange L(U):

L là hàm lồi với nên ta có thể tìm được kết quả từ phương trình

Suy ra:

Từ (3) chúng ta có :

Từ đó ta có chúng ta có kết quả:

(6)

với (7)

Chúng ta lại có:

Suy ra . Kết quả được lấy từ (5), từ nên nó cũng là phương án tối ưu cho (2).

**\*Thuật toán sFCM:**

Input:

+ Dữ liệu X

+ Độ thuộc được giám sát

+ Số cụm c

+ Giới hạn

+ Giới hạn maxStep

Output:

+ Tâm cụm V

+ Độ thuộc U

**\*Thuật toán sFCM:**

* Bước 1: Khởi tạo giá trị ban đầu của V, số lần chạy t=0.
* Bước 2: t=t+1
* Bước 3: Tính U trong khigiữ nguyên V và sử dụng công thức (6)
* Bước 4: Tính V trong khi giữ nguyên U và sử dụng công thức (4)
* Bước 5: Nếu tiêu chí dừng lại thoả mãn, thuật toán hoàn thành, U và V chính là kết quả cần tìm. Ngược lại, trở về Bước 2
  + **Tiêu chí dừng lại:**

**+** t > maxStep

+

### Phân cụm mờ bán giám sát với hệ số m thay đổi (sSMC-FCM)

**\*Cơ sở lý thuyết**

Trong phần này, chúng em sẽ tổng quan lại quá trình dẫn đến thuật toán sSMC-FCM. Mỗi phần tử có thể nhận giá trị tham số khác nhau. Phần tử giám sát thứ thuộc cụm được thể hiện bằng cách thay đổi . Quy ước giá trị có thể được cho như dưới đây, với điều kiện :

* với mọi , cho những phần tử không giám sát .
* , và với mọi , cho những phần tử giám sát thuộc cụm .

Phương trình của bài toán tối ưu (tìm min) được thể hiện như sau,

(8)

Trong đó là tập dữ liệu, số lượng dữ liệu, là số cụm, số điểm của phần tử đối với cụm với tâm cụm , giá trị tham số M của phần tử trong cụm , và bình phương khoảng cách giữa hai vector và .  
Gọi là tập các phần tử giám sát, ta có:

(9)

Để giải bài toán tối ưu được thể hiện ở phương trình (7), ta sử dụng phương pháp nhân tử Lagrange. Gọi

(10)

Hơn nữa, với

Ta có thể tính được V:

Từ đó, ta có

hay

(11)

Tiếp theo ta tính sử dụng phương trình:

(12)

Từ đó:

(13)

Kết hợp , ta chia thành 2 trường hợp:  
**Trường hợp 1:**  
Với những phần tử không được giám sát cho mọi , phương trình (13) trở thành:

(14)

Vì , nên , hay , hay:

(15)

Thay (14) vào (13), ta có:

(16)

**Trường hợp 2:**  
Với những phần tử giám sát thuộc cụm và với , Phương trình (12) trở thành:

(17)

Kết hợp với , để tính toán , ta cần giải hệ sau: với mọi   
Các bược để giải phương trình (17) được thể hiện bởi các phương trình (18)-(21) như dưới đây.

Trước hết, ta tính , sau đó:

(18)

Tính với mọi (19)   
Tính là nghiệm trong phương trình (20)

Cuối cùng : (21)

Từ đây, ta có thuật toán sSMC-FCM như dưới đây.

**\*Thuật toán**  
Input: Tập dữ liệu , tham số and , một tập các phần tử giám sát   
Output: Chia tập X và C cụm.

* Bước 1: Khởi tạo giá trị cho , đặt , .
* Bước 2: Với vòng lặp thứ , cập nhật theo phương trình (16) đối vói phần tử không giám sát, hoặc theo các phương trình (18)-(21) với những phần tử giám sát.
* Bước 3: Cập nhật cho bước tiếp theo , theo phương trình (10), với được tính ở phương trình (9).
* Bước 4: Nếu , đi đến Step Nếu không, gán , Trở lại Bước 2.
* Bước 5: Kết thúc.

### Đánh giá các thuật toán phân cụm mờ sử dụng độ đo ngoài

Sử dụng các thuật toán sSFCM và sFMC- FCM để phân cụm dữ liệu, ta dựa vào nhãn của dữ liệu và kết quả phân cụm dữ liệu để đánh giá các thuật toán sử dụng độ đo ngoài

**\*Rand Index**

**\*Adjusted Rand Index**

**\*Jaccard Coefficient**

Trong đó:

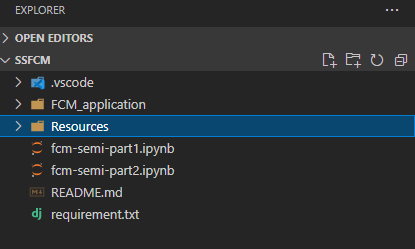
* a số cặp dữ liệu cùng nhãn và được phân vào cùng 1 cụm
* b số cặp dữ liệu cùng nhãn và được phân vào 2 cụm khác nhau
* c số cặp dữ liệu khác nhãn và được phân vào cùng 1 cụm
* d số cặp dữ liệu khác nhãn và được phân vào 2 cụm khác nhau
* M = a + b + c + d.

## Xây dựng chương trình

### Cài đặt chương trình

Yêu cầu: Python 3

* Bước 1: Pull code về máy
  + Link source code: [phuchoang92/sSFCM: Semi-Supervised Fuzzy C-means Clustering (github.com)](https://github.com/phuchoang92/sSFCM)

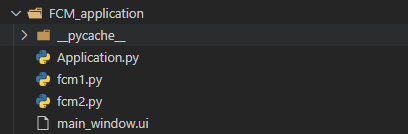


ảnh minh họa sau khi pull code

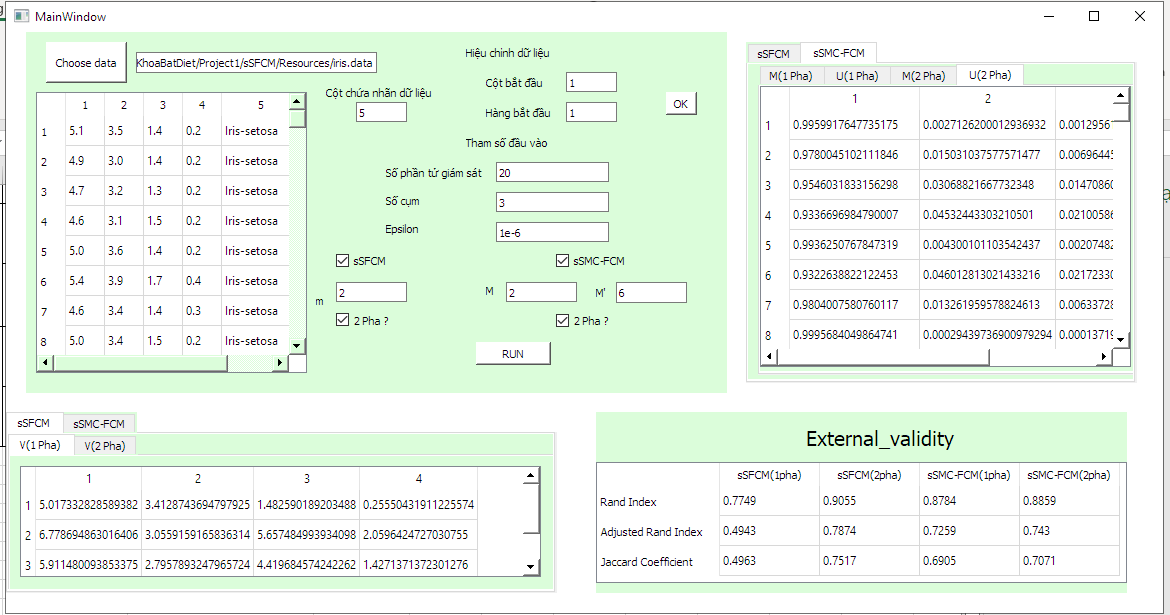
* Trong đó source code của chương trình nằm ở folder FCM\_application
* README.md file hướng dẫn chạy và sử dụng sources code
* Các file fcm-semi-part1.ipynb và fcm-semi-part2.ipynb là các file được viết trong quá trình phát triển (không phải mã nguồn cuối)
* requirement.txt là file chứa các thư viện của python cần thiết cho việc chạy chương trình
* Bước 2: Cài đặt các thư viện cần thiết
  + Sử dụng lệnh: pip install -r requirement.txt



* Bước 3: Chạy chương trình : Chạy file Application.py trong thư mục FCM\_application



### Giao diện chương trình

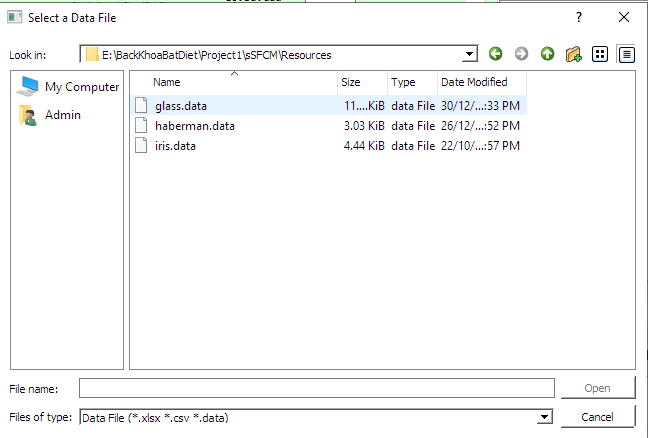


### Các bước chạy chương trình

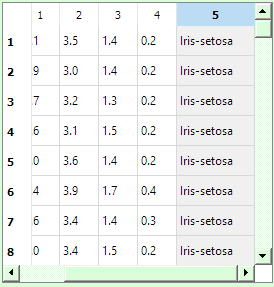
* Bước 1: Chọn bộ dữ liệu:
* Ấn nút chọn dữ liệu



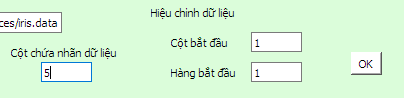
* Chọn bộ dữ liệu từ thư mục Resources hoặc bộ dữ liệu phân cụm tùy chọn (Dữ liệu phải ở dạng số, phải có nhãn dữ liệu)



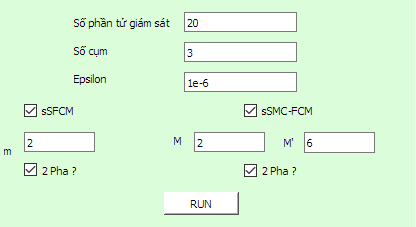
* Ví dụ Bảng dữ liệu hợp lệ



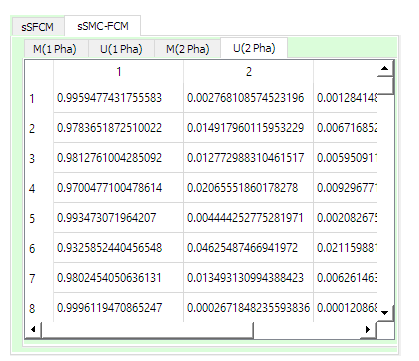
* Bước 2: Hiệu chỉnh dữ liệu
* Chọn cột chứa nhãn dữ liệu, điều chỉnh hàng và cột bắt đầu dữ liệu nếu sai lệch => Nhấn ok để hoàn tất xử lý dữ liệu



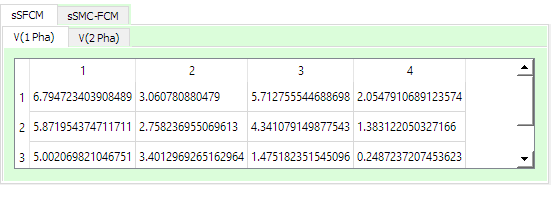
* Bước 3: Điền các tham số và chọn các tùy chọn để chạy thuật toán



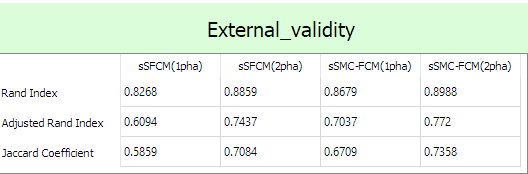
* Bước 4: Ấn run để chạy thuật toán. Ở bước này, quy trình chương trình chạy sẽ như sau:
  + Khởi tạo các phần tử giám sát dựa theo số lượng phần tử giám sát được nhập. – Cố định cho cả 2 thuật toán trong một lần chạy chương trình
  + Kiểm tra các dữ liệu đầu vào, tham số của từng thuật toán
  + Chạy các thuật toán được yêu cầu và trả về kết quả ra màn hình
* Bược 5: Xem kết quả
  + Các ma trận tham số



* Tâm cụm:



* Các độ đo:



## Kết quả

Sau đây là kết quả thực hiện phân cụm bằng thuật toán sSFCM và sSMC-FCM trên hai bộ dữ liệu Iris và WDBC

* Iris: 150 dữ liệu, 4 thuộc tính, chia làm 3 lớp
* WDBC: 569 dữ liệu, 30 thuộc tính, chia làm 2 lớp

Các tham số khi chạy thuật toán:

* sSFCM:
  + m=2
* sSMC-FCM:
  + M=2
  + M’=9 hoặc M’=20
* Epsilon =

### Các thuật toán khi chạy 1 pha

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Bộ dữ liệu** | **Thuật toán** | **Độ đo ngoài** | **Không giám sát** | **10%** | **20%** | **30%** |
| Iris | sSFCM | Rand Index | 0.8797 | 0.8242 | 0.774 | 0.6821 |
| Adjusted Rand Index | 0.7294 | 0.6035 | 0.4913 | 0.2807 |
| Jaccard Coefficient | 0.6943 | 0.5811 | 0.4933 | 0.3494 |
| sSMC-FCM  (M=2,M’=9) | Rand Index | 0.8797 | 0.868 | 0.829 | 0.8026 |
| Adjusted Rand Index | 0.7294 | 0.7022 | 0.6142 | 0.5554 |
| Jaccard Coefficient | 0.6943 | 0.6679 | 0.59 | 0.5423 |
| WDBC(Breast Cancet Wisconsin (Diagnostic) | sSFCM | Rand Index | 0.7504 | 0.6897 | 0.6686 | 0.6338 |
| Adjusted Rand Index | 0.4914 | 0.3693 | 0.3298 | 0.2623 |
| Jaccard Coefficient | 0.6499 | 0.5774 | 0.5431 | 0.496 |
| sSMC-FCM  (M=2, M’=6) | Rand Index | 0.7504 | 0.7332 | 0.7098 | 0.7284 |
| Adjusted Rand Index | 0.4914 | 0.4569 | 0.4102 | 0.4476 |
| Jaccard Coefficient | 0.6499 | 0.6289 | 0.5988 | 0.6215 |

*Kết quả chạy khi chạy các thuật toán với các phần tử giám sát được cho vào cụm ngẫu nhiên*

Rand Index khi chạy thuật toán với độ giám sát thay đổi

**Nhận xét:**

* Các thuật toán đều chạy tốt và hội tụ
* Khi các thuật toán chạy không giám sát, kết quả là như nhau (thuật toán chuyển về FCM)
* Khi tăng độ giám sát, do chọn các phần tử vào cụm ngẫu nhiên, dẫn đến độ đo ngoài bị giảm. Điều này sẽ được khắc phục khi chạy thuật toán 2 pha
* Buộc các phần tử vào các cụm không hợp lý có thể dẫn đến kết quả của bài toán phân cụm bị giảm đi
* Giá độ đo ngoài của thuật toán sSMC-FCM tốt hơn một chút so với thuật toán sSFCM

### Các thuật toán khi chạy 2 pha

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Bộ dữ liệu** | **Thuật toán** | **Độ đo** | **Không giám sát** | **10%** | **20%** | **30%** | **40%** | **50%** |
| Iris | sSFCM | Rand Index | 0.8797 | 0.9124 | 0.8988 | 0.9124 | 0.9341 | 0.9417 |
| Adjusted Rand Index | 0.7294 | 0.8019 | 0.7715 | 0.8022 | 0.8512 | 0.8681 |
| Jaccard Coefficient | 0.6943 | 0.7656 | 0.7348 | 0.7663 | 0.8188 | 0.8375 |
| sSMC-FCM | Rand Index | 0.8797 | 0.9124 | 0.8988 | 0.9124 | 0.9341 | 0.9417 |
| Adjusted Rand Index | 0.7294 | 0.8019 | 0.7715 | 0.8022 | 0.8512 | 0.8681 |
| Jaccard Coefficient | 0.6943 | 0.7656 | 0.7348 | 0.7663 | 0.8188 | 0.8375 |
| WDBC | sSFCM | Rand Index | 0.7504 | 0.7733 | 0.7946 | 0.8194 | 0.8423 | 0.863 |
| Adjusted Rand Index | 0.4914 | 0.5392 | 0.5831 | 0.6342 | 0.6811 | 0.7232 |
| Jaccard Coefficient | 0.6499 | 0.6731 | 0.696 | 0.7247 | 0.753 | 0.7806 |
| sSMC-FCM  M=2, M’=6 | Rand Index | 0.7504 | 0.7733 | 0.7579 | 0.8055 | 0.8222 | 0.8541 |
| Adjusted Rand Index | 0.4914 | 0.5392 | 0.5069 | 0.6056 | 0.64 | 0.705 |
| Jaccard Coefficient | 0.6499 | 0.6731 | 0.6583 | 0.7083 | 0.7281 | 0.7688 |
| sSMC-FCM  M=2, M’=20 | Rand Index | 0.7504 | 0.7733 | 0.7946 | 0.8194 | 0.8423 | 0.863 |
| Adjusted Rand Index | 0.4914 | 0.5392 | 0.5831 | 0.6342 | 0.6811 | 0.7232 |
| Jaccard Coefficient | 0.6499 | 0.6731 | 0.696 | 0.7247 | 0.753 | 0.7806 |

*Kết quả chạy khi chạy các thuật toán với các phần tử giám sát được cho vào cụm đã biết trước*

* Khi điều chỉnh giá trị tham số M’ của thuật toán sSMC-FCM, kết quả tăng lên
  + Khi M=2, M’=9:
* Khi M=2, M’=20:

**Nhận xét:**

* Các thuật toán đều chạy tốt và hội tụ
* Khi các thuật toán chạy không giám sát, kết quả là như nhau (thuật toán chuyển về FCM)
* Khi tăng dần độ giám sát, giá trị các độ đo ngoài đều tăng
* Kết quả của thuật toán sSMC-FCM phụ thuộc vào cách chọn cặp giá trị M, M’ khác nhau tùy vào bộ dữ liệu đang xét
* Khi giá trị M’ ở thuật toán sSMC-FCM đủ lớn, kết quả hai thuật toán gần như là như nhau

# TỔNG KẾT

Trong đồ án lần này, chúng em đã tìm hiểu về bài toán phân cụm có giám sát, cụ thể là thuật toán phân cụm sSFCM và sSMC-FCM. Hai thuật toán đều cho ra kết quả giống nhau và thỏa mãn lý thuyết về tính hội tụ của thuật toán. Đã xây dựng được chương trình minh họa cho cả hai thuật toán và chạy thử đều cho kết quả tốt, có thể xử lý được các bộ dữ liệu khác nhau. Kết quả của thuật toán sau khi chạy với một số bộ dữ liệu khác nhau khá chính xác so với mô tả của bộ dữ liệu và phù hợp với lý thuyết.

# TÀI LIỆU THAM KHẢO

1. ENDO Yasunori, HAMASUNA Yukihiro, YAMASHIRO Makito, MIYAMOTO Sadaaki. On Semi-Supervised Fuzzy c-Means Clustering;FUZZ-IEEE 2009, Korea, August 20-24, 2009
2. Khang, T.D.; Tran, M.-K.; Fowler, M. A Novel Semi-Supervised Fuzzy C-Means Clustering Algorithm Using Multiple Fuzzification Coefficients. Algorithms 2021, 14, 258. Academic Editor: Frank Werner, Published: 29 August 2021

https://doi.org/10.3390/a140902

1. Lucas Vendramin, Ricardo J. G. B. Campello and Eduardo R. Hruschka. Relative Clustering Validity Criteria: A Comparative Overview; Department of Computer Sciences of the University of São Paulo at São Carlos, C.P. 668, São Carlos, Brazil; Published online 30 June 2010 in Wiley InterScience (www.interscience.wiley.com)