**TRƯỜNG ĐẠI HỌC BÁCH KHOA TP.HCM**

**KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH**

**BÁO CÁO ĐỀ TÀI SỐ 8:**

**NGHIÊN CỨU BÀI TOÁN GOM CỤM**

**TRONG KHAI PHÁ DỮ LIỆU**

CƠ SỞ LÝ THUYẾT HỌ GIẢI THUẬT

**FUZZY C-MEANS**

*Giảng viên hướng dẫn:*

*Nhóm thực hiện:*

***Châu Vĩnh Tuân 50802429***

***Phạm Nguyên Trình 50802353***

*Tháng 12 - 2011*

**MỤC LỤC**

[I. LÝ THUYẾT GOM CỤM: 2](#_Toc310924584)

[**1.** **Khái niệm gom cụm:** 2](#_Toc310924585)

[**2.** **Vai trò của­­­­­­­­­ gom cụm:** 3](#_Toc310924586)

[**3.** **Một số độ đo trong gom cụm:** 3](#_Toc310924587)

[**4.** **Mục đích của gom cụm:** 3](#_Toc310924588)

[**5.** **Một số phương pháp gom cụm điển hình:** 3](#_Toc310924589)

[**6.** **Một số mô hình cụm dữ liệu:** 4](#_Toc310924590)

[II. FUZZY C-MEANS (FCM): 5](#_Toc310924591)

[**1.** **Tìm hiểu Fuzzy C-Means:** 5](#_Toc310924592)

[**2.** **Giải thuật:** 6](#_Toc310924593)

[**3.** **Ưu và nhược điểm:** 6](#_Toc310924594)

[III. CHƯƠNG TRÌNH MẪU: 7](#_Toc310924595)

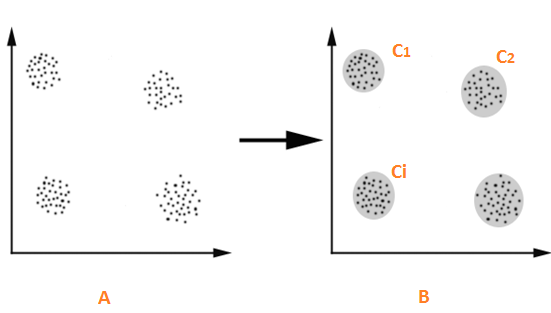
[**1.** **Hướng dẫn sử dụng phần mềm Fuzzy C-Mean Analyst:** 7](#_Toc310924596)

[IV. KẾT LUẬN VÀ KIẾN NGHỊ: 11](#_Toc310924597)

1. **LÝ THUYẾT GOM CỤM:**
2. **Khái niệm gom cụm:**

Gom cụm (hay phân cụm) dữ liệu là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn các điều kiện:

* Các đối tượng trong cùng một cụm “tương tự” nhau về một số tiêu chí nào đó.
* Các đối tượng khác cụm thì “không tương tự” nhau.

Giải quyết các vấn đề tìm kiếm, phát hiện các cụm, các mẫu dữ liệu trong một tập hợp ban đầu các dữ liệu không có nhãn.

A: là một tập các điểm dữ liệu trước khi gom cụm

B: là các tập điểm dữ liệu sau khi gom cụm

* + Ci : cụm thứ i

Chúng ta sẽ đưa ra một số thuật ngữ để thuận tiện trong việc mô tả:

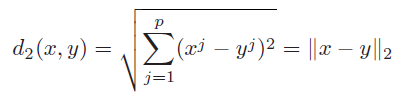
* Tập các đối tượng sẽ được phân vùng

Trong một số trường hợp có thể X là đối tượng vector trong không gian p chiều . Khi đó ta kí hiệu

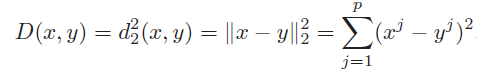
* Việc phân vùng dữ liệu dựa trên 2 khái niệm căn bản: sự không đồng dạng và tâm vùng. Việc so sánh sự không đồng dạng giữa 2 đối tượng X và X’ được biểu thị bởi quan hệ D(X,X’) trong đó đảm bảo tính đối xứng

Trong tính toán D(X,X’) sẽ cho chúng ta giá trị thực thể hiện mức độ “gần nhau” giữa 2 đối tượng. Một cách hiển nhiên:



* Việc đánh giá quan hệ không đồng dạng trong 1 không gian cho trước thường sử dụng nhiều đến khái niệm metric, metric giữa 2 đối tượng x,y m(x,y) cần thỏa mãn: 
* Khái niệm gần gũi chúng ta nhất là khoảng cách Euclid: 

Với những ứng dụng xây dựng trong không gian Euclid , hàm quan hệ đánh giá mức độ không đồng dạng D(X,Y) chúng ta dùng (được mô tả dưới đây) được xác định bằng bình phương khoảng cách Euclid:



1. **Vai trò của gom cụm:**

Gom cụm dữ liệu đóng vai trò quan trọng trong các ngành khoa học :

* Sinh học.
* Khôi phục dữ liệu
* Dự báo thời tiết
* Tâm lý học và Y học
* Kinh doanh

Gom cụm dữ liệu mang lại các tiện ích:

* Tổng kết
* Nén
* Tìm kiếm kết quả gần nhất

1. **Một số độ đo trong gom cụm:**

* Minkowski
* Euclidean – p = 2
* Độ đo tương tự: cosin hai vectơ

1. **Mục đích của gom cụm:**

Xác định được bản chất của việc gom thành nhóm các đối tượng trong một tập dữ liệu không có nhãn.

Không có tiêu chuẩn chung nào để gom cụm dữ liệu, gom cụm dựa vào tiêu chí người dùng cung cấp trong từng trường hợp.

1. **Một số phương pháp gom cụm điển hình:**

Đối với một tập dữ liệu cho trước, chúng ta có nhiều cách để thực hiện việc phân vùng dữ liệu. Việc nhìn nhận xem phương pháp gom vùng nào phù hợp phụ thuộc vào kết quả đầu ra mà chúng ta mong muốn.

Có các phương pháp:

* Hierarchical
* Partitional
* Exclusive
* Overlapping
* Fuzzy
* Complete
* Partial

1. **Một số mô hình cụm dữ liệu:**

* Phân tách
* Nguyên mẫu
* Đồ thị
* Dựa trên mật độ
* Chia sẻ

1. **FUZZY C-MEANS (FCM):**
2. **Tìm hiểu Fuzzy C-Means:**
   1. **Fuzzy logic:**
   * Fuzzy Logic là một hình thức logic có nhiều giá trị.
   * Biến Fuzzy Logic có thể có một giá trị chân lý dao động giữa [0,1]
   1. **Tập Fuzzy:**
   * Là tập hợp mà các phần tử có một mức độ thành viên nhất định
   * Tập Fuzzy được định nghĩa là cặp (A,m), trong đó A là tập hợp và m là ánh xạ *m: A🡪[0,1]*
     + Với mỗi phần tử ,  được gọi là hệ số thành viên của x trong (A,m). Cho tập hữu hạn A = {x1,...,xn}, tập Fuzzy (A,m) thường được mô tả nhă sau: {m(x1) / x1,...,m(xn) / xn}.
     + m(x) = 0 : x không thuộc (A, m)
     + m(x) = 1: x hoàn toàn thuộc (A, m)
   1. **Fuzzy C-Means:**
   * Fuzzy C-Means (FCM ) là một phương pháp của phân nhóm cho phép một phần dữ liệu thuộc về hai hoặc nhiều cụm
   * Thường xuyên được sử dụng trong nhận dạng mẫu
   * FCM được hiện thực dựa trên hàm mục tiêu:

Trong đó:

* + - m là bất kỳ số thực lớn hơn 1
    - uij là mức độ của các thành viên của xi trong cụm j
    - xi là chiều thứ i của dữ liệu đo d-chiều
    - cj là trung tâm của cụm kích thước d-chiều
    - ||\*|| Là bất kỳ chỉ tiêu nào thể hiện sự giống nhau giữa dữ liệu đo

1. **Giải thuật FCM:**

FCM được thực hiện lần lượt theo các bước:

* Bước 1: Khởi tạo ma trận U=[uij], U(0)
* Bước 2: sắp xếp các đối tượng sao cho gần tâm vùng nhất ,điều này có nghĩa là tại lần lặp thứ k: tính toán véc-tơ trung tâm C(k)=[cj] với U(k)
* Bước 3: Cập nhật U(k) và U(k+1)
* Bước 4: Kiểm tra

Nếu kết vẫn chưa thỏa, ta quay lại bước 2, nếu đã thỏa mãn, ta kết thúc tính toán.

1. **Ưu và nhược điểm:**
   1. **Ưu điểm:**
   * FCM khắc phục được các cụm dữ liệu chồng nhau trên các tập dữ liệu có kích thước lớn hơn, nhiều chiều và nhiều nhiễu. Cung cấp cho kết quả tốt nhất cho dữ liệu chồng chéo và tương đối tốt hơn thuật toán K-Means.
   * Không giống như K-Means, dữ liệu điểm duy nhất phải thuộc về một cụm duy nhất, ở mỗi điểm được phân vào cụm dựa vào kết quả tính toán hàm thành viên, vì vậy, một điểm có thể thuộc về nhiều hơn một cụm.
   1. **Nhược điểm:**
   * Cần tiên nghiệm số lượng các cụm .
   * ε càng thấp kết quả nhận được càng tốt nhưng chi phí tính toán càng nhiều
   * Nhạy cảm với nhiễu và các phần tử ngoại lai.
   * Khoảng cách Euclide các yếu tố cơ bản có thể không đồng đều.

1. **CHƯƠNG TRÌNH MẪU:**
2. **Hướng dẫn sử dụng phần mềm Fuzzy C-Mean Analyst:**
   1. **Yêu cầu hệ thống:**
   * Hệ điều hành: Windows 7/ Vista / XP (32bit)
   * Máy ảo Java (JVM) phiên bản 1.6 trở lên
   * Danh sách file ( yêu cầu không thay đổi hệ thống file này):

12/05/2011 04:05 PM 71,225 **FuzzyCMeanAnalyst.jar**

03/03/2011 06:57 AM 9,728 **gluegen-rt.dll**

03/03/2011 06:57 AM 416,768 **jogl\_desktop.dll**

03/03/2011 06:57 AM 73,216 **jogl\_es1.dll**

03/03/2011 06:57 AM 77,312 **jogl\_es2.dll**

12/05/2011 04:23 PM <DIR> **lib**

03/03/2011 06:57 AM 10,240 **nativewindow\_awt.dll**

03/03/2011 06:57 AM 36,864 **nativewindow\_win32.dll**

03/03/2011 06:57 AM 41,984 **newt.dll**

./lib

12/05/2011 04:05 PM 110,455 **gluegen-rt.jar**

12/05/2011 04:05 PM 2,419,760 **jogl.all.jar**

12/05/2011 04:05 PM 128,511 **nativewindow.all.jar**

12/05/2011 04:05 PM 176,393 **newt.all.jar**

* 1. **Hướng dẫn chạy phần mềm:**

Thực hiện theo các bước sau:

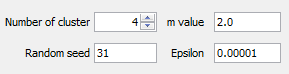
* + Khởi động phần mềm bằng cách chạy file

**FuzzyCMeanAnalyst.jar**

* + Chọn file input bằng cách click vào button **Browse**



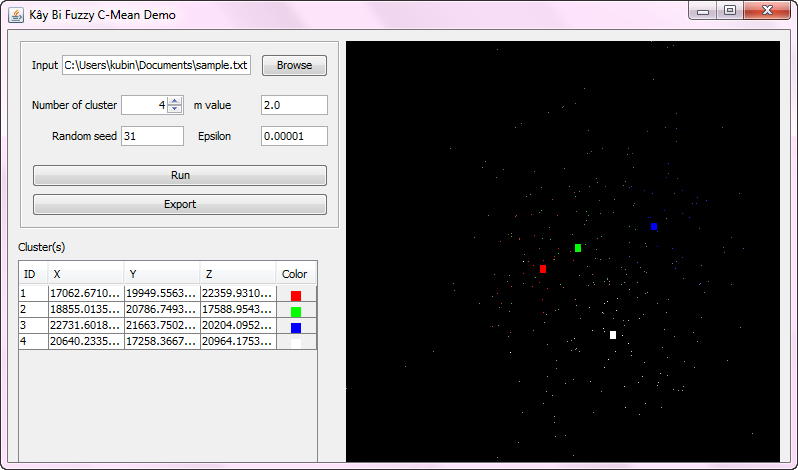
* + Thiết lập các thông số (**Number of clusters, m value, Random seed, Epsilon**) thích hợp



* + - * **Number** **of** **clusters**: số lượng cluster muốn phân tích
      * **m** **value**: giá trị m của công thức trong bài toán Fuzzy C-Mean
      * **Random** seed: giá trị để sinh ngẫu nhiên ma trận ban đầu U
      * **Epsilon**: độ chính xác của giải thuật
  + Click button **Run** để thực hiện việc tính toán và mô tả như hình minh họa



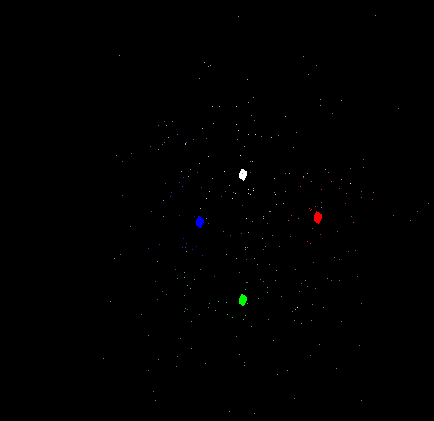
* + Kết quả:



* + Click button Export để xuất ra file định dạng plain text các thông số của giải thuật.

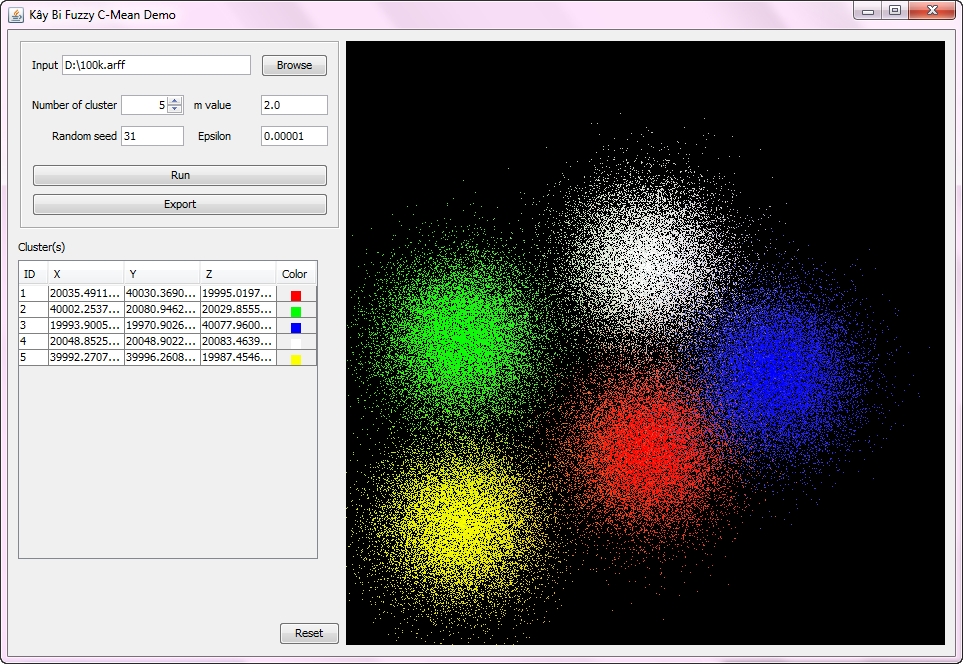


* + Để xem hình minh họa ở góc nhìn khác, có thể thực hiện click chuột và kéo trên màn hình mô phỏng hoặc nhấn các phím UP, DOWN, LEFT, RIGHT (nếu không có tác dụng thì click một lần lên màn hình mô phỏng và thự hiện lại thao tác).



* + Để về lại với góc nhìn ban đầu, nhấn Reset

1. **Kết quả chạy với dữ liệu mẫu:**



1. **TÀI LIỆU THAM KHẢO:**

* Data Mining: Concepts and Techniques (Second Edition) – Jiawei Han and Micheline Kamber
* Fuzzy Cluster Analysis – John Wiley and Sons
* Algorithms for Fuzzy Cluster, Methods in c-Means Clustering with Applications - Sadaaki Miyamoto,Hidetomo Ichihashi, KatsuhiroHonda