# LỜI MỞ ĐẦU

Trong những năm gần đây, do sự phát triển vượt bậc, không ngừng vươn lên của nền kinh tế đất nước, kéo theo các hệ thống dữ liệu phục vụ cho các lĩnh vực kinh tế - xã hội đã phát triển bùng nổ, lượng dữ liệu khổng lồ được tạo ra gay càng lớn. Sự phong phú về dữ liệu, thông tin cùng với khả năng kịp thời khai thác chúng đã mang đến những năng xuất và chất lượng  mới cho công tác quản lý, hoạt động kinh doanh,…Không chỉ dừng lai ở đó, các yêu cầu về thông tin, khám phá tri thức mới trong các lĩnh vực này, đặc biệt trong lĩnh vực ra quyết định, ngày càng đòi hỏi cao hơn. Trước nhu cầu  đó, hàng loạt các lĩnh vực nghiên cứu về tổ chức các kho dữ liệu và kho thông tin, các hệ trợ giúp quyết định, các thuật toán nhận dạng, …và đặc biệt là Data Mining ra đời.

` Data Mining là một lĩnh vực mới xuất hiện, nhằm tự động khai thác những thông tin, những tri thức có tính tiềm ẩn, hữu ích từ những CSDL lớn cho các đơn vị, tổ chức, doanh nghiệp, …từ đó làm thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh cho các đơn vị, tổ chức này. Từ những ứng dụng thành công trong khám phá tri thức, cho thấy Data Mining là một lĩnh vực phát triển bền vững  mang lại nhiều lợi ích và có nhiều triển vọng, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyên thống. Hiện nay, Data Mining đã ứng dụng ngày càng rộng rãi trong các lĩnh vực như: thương mại, tài chính, điều trị y học, viễn thông, tin-sinh, …

Một trong những hướng nghiên cứu chính của Data Mining là phân cụm dữ liệu(Data Clustering). Phân cụm dữ liệu là quá trình tìm kiếm và phát hiện ra các cụm dữ liệu tự nhiên tiềm ẩn, quan tâm trong cơ sở dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định. Có rất nhiều kĩ thuật trong phân cụm dữ liệu như:phân cụm dữ liệu phân hoạch, phân cụm dữ liệu phân cấp, phân cụm dựa trên mật độ, …Tuy nhiên các kĩ thuật này đều hướng tới hai mục tiêu chung đó là chất lượng các cụm khám phá được và tốc độ thực hiện của thuật toán. Trong đó, kĩ thuật phân cụm dữ liệu phân hoạch là một kĩ thuật có thể đáp

ứng được những mục tiêu đó của bài toán phân cụm với khả năng làm việc đối với CSDL lớn.

Yêu cầu về các phương pháp khai phá dữ liệu và việc thực hiện các thuật toán trong đó có hiệu quả trên thực tế là vấn đề đang thu hút được rất nhiều quan tâm. Đo đó, em đã chọn đề tài nghiên cứu " ***Tìm hiểu và cài đặt một số thuật toán phân cụm dữ liệu cơ bản"*** cho đồ án tốt nghiệp của mình.

Nội dung của đồ án gồm 3 chương:

**Chương 1: Giới thiệu về phân cụm dữ liệu**: Trong chương này em trình bày tổng quan về phân cụm dữ liệu, bao gồm các kiểu dữ liệu có thể phân cụm, các ứng dụng và các kỹ thuật phân cụm dữ liệu. Đây là một hướng tiếp cận chính trong Data Mining. Trong đó, đi sâu phân tích chi tiết các vấn đề cơ bản trong PCDL và ý nghĩa của PCDL, đặc điểm của các kiểu dữ liệu cơ bản thường sử dụng trong PCDL như: dữ liệu có thuộc tính hạng mục (Categorical), dữ liệu có thuộc tính số, … Các khái niệm về "tương tự" và "phi tương tự" cũng được trình bày trong chương này

**Chương 2: Trình bày về các phương pháp phân cụm dữ liệu phân hoạch**: trình bày vắn tắt về các thuật toán trong PCDL phân hoạch, trong đó đồ án đi sâu vào tìm hiểu về 2 thuật toán phân cụm dữ liệu phân hoạch điển hình: K-MEANS, PAM.

**Chương 3: Cài đặt thực nghiệm:** Để khẳng định cho khả năng và hiệu quả của thuật toán phân cụm dữ liệu phân hoạch, em đã lựa chọn và cài đặt các thuật toán K-MEANS, PAM, trên cơ sở dữ liệu là các điểm ảnh được biểu diễn bằng các toạ độ trong không gian. Kết quả của chương trình là một ảnh trên đó các điểm ảnh gần nhau đã được gom vào một nhóm.

Cuối cùng là phần kết luận trình bày tóm tắt các kết quả thu được và các đề xuất cho hướng phát triển của đề tài.

# CHƯƠNG 1:
# PHÂN CỤM DỮ LIỆU - Data Clustering

## 1. 1. Vấn đề phân cụm dữ liệu

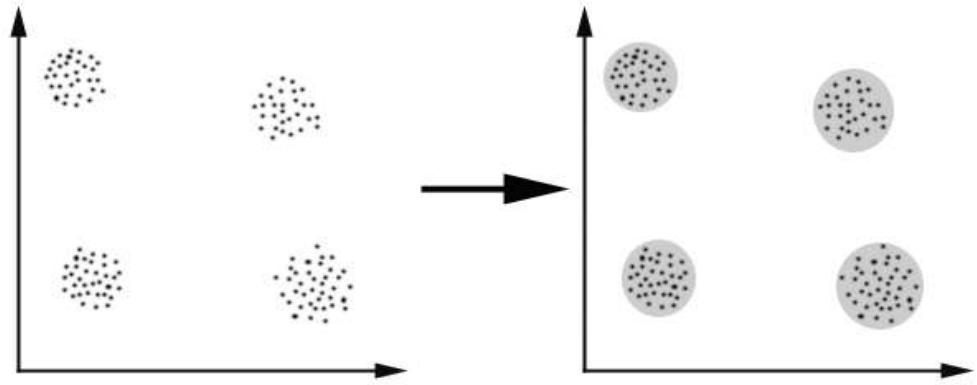
Phân cụm dữ liệu là một trong những hướng nghiên cứu trọng tâm của lĩnh vực khai phá dữ liệu (Data Mining) và lĩnh vực khám phá tri thức (KDD). Mục đích của phân cụm là nhóm các đối tượng vào các cụm sao cho các đối tượng trong cùng một cụm có tính tương đồng cao và độ bất tương đồng giữa các cụm lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định.

Ở một mức cơ bản nhất, người ta đã đưa ra định nghĩa PCDL như sau:

*"PCDL là một kỹ thuật trong DATA MINING, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định*"

Như vậy, PCDL là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" (Similar) với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định của phương pháp phân cụm.

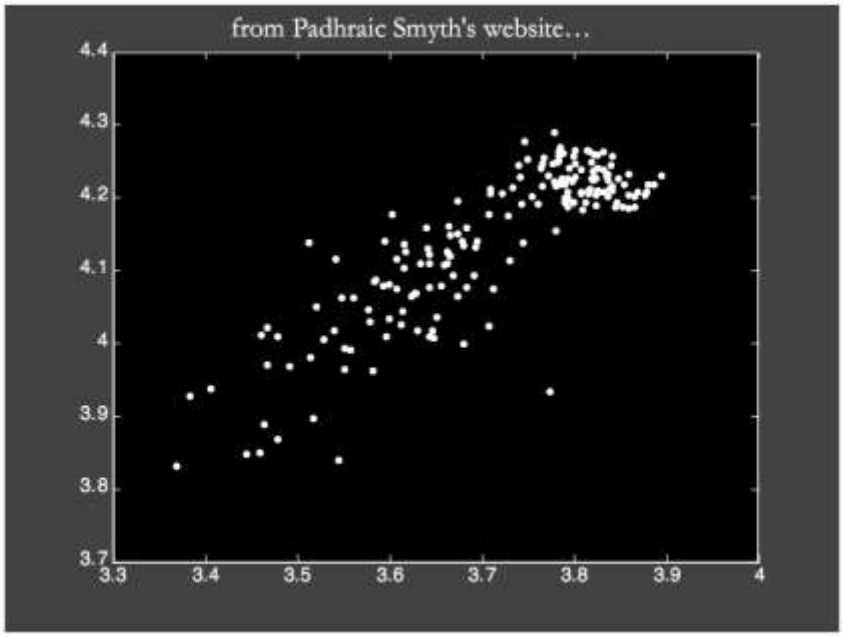
Chúng ta có thể minh hoạ vấn đề phân cụm như hình 1 sau đây:



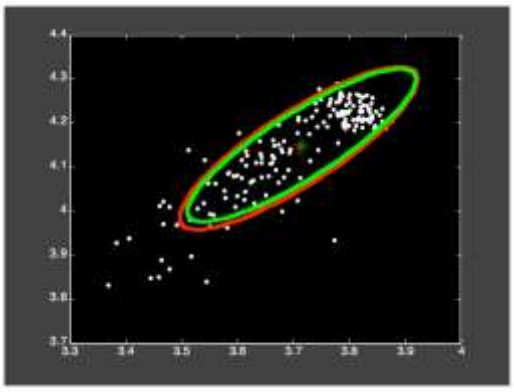
**Hình 1: Mô phỏng vấn đề PCDL**

Trong hình trên, sau khi phân cụm chúng ta thu được bốn cụm trong đó các phần tử "*gần nhau*" hay là "*tương tự*" thì được xếp vào một cụm, trong khi đó các phần tử "*xa nhau*" hay là "*phi tương tự*" thì chúng thuộc về các cụm khác nhau.

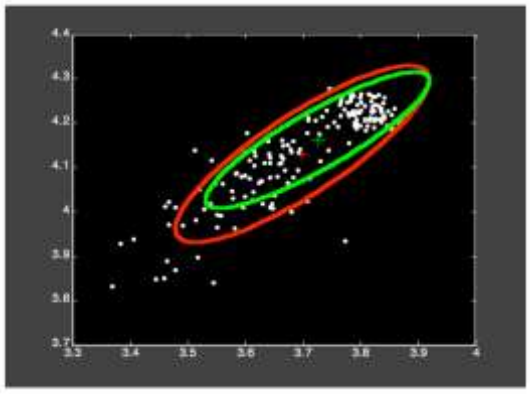
Để minh hoạ cụ thể hơn cho vấn đề này ta có thể quan sát các hình ảnh sau:



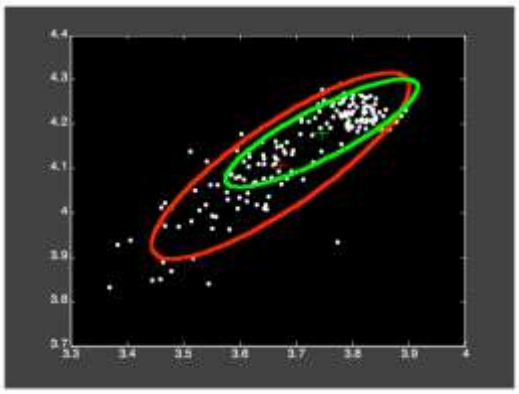

**Hình 2: Dữ liệu nguyên thuỷ**



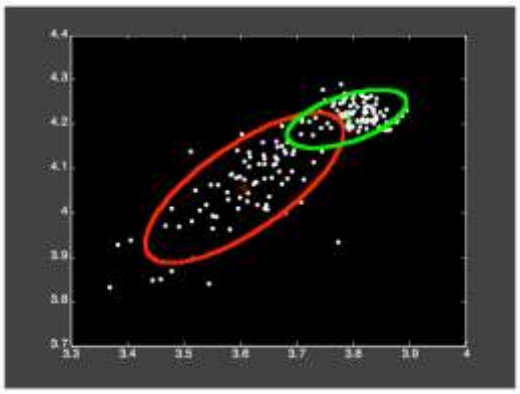
**Hình 3**



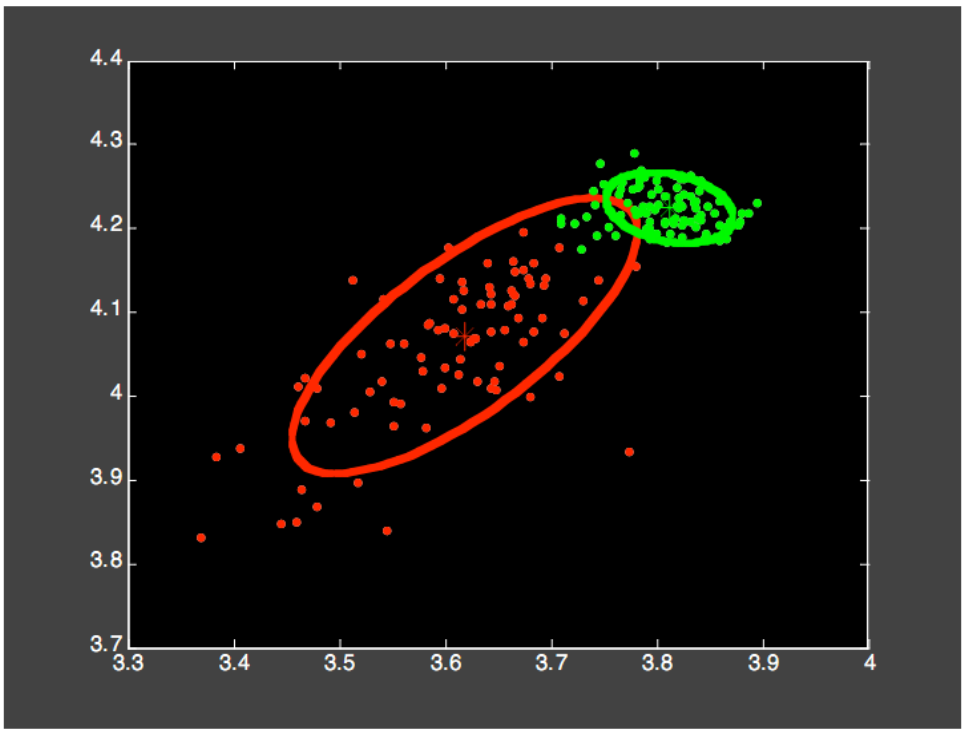
**Hình 4**



**Hình 5**



**Hình 6**




**Hình 7: Kết quả của quá trình phân cụm**

Các hình 2, 3, 4, 5, 6 ,7 là thể hiện quá trình phân cụm từ khi"bắt đầu" cho đến khi "kết thúc" .

Trong PCDL khái niệm (*Concept Clustering*) thì hai hoặc hoặc nhiều đối tượng cùng được xếp vào một cụm nếu chúng có chung một định nghĩa về khái niệm hoặc chúng xấp xỉ với các khái niệm mô tả cho trước, như vậy, ở đây PCDL không sử dụng khái niệm "*tương tự*" như đã trình bày ở trên.

Trong học máy, phân cụm dữ liệu được xem là vấn đề học không có giám sát, vì nó phải đi giải quyết vấn đề tìm một cấu trúc trong tập hợp các dữ liệu chưa biết biết trước các thông tin về lớp hay các thông tin về tập ví dụ huấn luyện. Trong nhiều trường hợp, khi phân lớp (Classification) được xem vấn đề học có giám sát thì phân cụm dữ liệu là một bước trong phân lớp dữ liệu, trong đó PCDL sẽ khởi tạo các lớp cho phân lớp bằng cách xác định các nhãn cho các nhóm dữ liệu.

Một vấn đề thường gặp trong PCDL đó là hầu hết các dữ liệu cần cho phân cụm đều có chứa dữ liệu "nhiễu" (noise) do quá trình thu thập thiếu chính xác hoặc thiếu đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lý dữ liệu nhằm khắc phục hoặc loại bỏ "nhiễu" trước khi bước vào giai đoạn phân tích phân cụm dữ liệu. "Nhiễu" ở đây có thể là các đối tượng dữ liệu không không chính xác, hoặc là các đối tượng dữ liệu khuyết thiếu thông tin về một số thuộc tính. Một trong các kỹ thuật xử lý nhiễu phổ biến là việc thay thế giá trị của các thuộc tính của đối tượng "nhiễu" bằng giá trị thuộc tính tương ứng của đối tượng dữ liệu gần nhất.

Tóm lại, phân cụm là một vấn đề khó, vì rằng người ta phải đi giải quyết các vấn đề con cơ bản như sau:

- ***Xây dựng hàm tính độ tương tự.***
- ***Xây dựng các tiêu chuẩn phân cụm.***
- ***Xây dựng mô hình cho cấu trúc cụm dữ liệu***
- ***Xây dựng thuật toán phân cụm và các xác lập các điều kiện khởi tạo.***
- ***Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm***

Phân cụm dữ liệu là bài toán thuộc vào lĩnh vực học máy không giám sát và đang được ứng dụng rộng rãi để khai thác thông tin từ dữ liệu

Phân cụm dữ liệu là quá trình phân chia một tập dữ liệu ban đầu thành các cụm sao cho các đối t ượng trong cùng một cụm “tương tự “. Việc tính “khoảng cách “ giữa các đối tượng, hay phép đo tương tự giữa các cặp đối tượng để phân

chia chúng vào các cum khác nhau. dựa vào hàm tính độ tương tự này cho phép xác định được hai đối tượng có tương tự hay không. theo quy ước, giá trị của hàm tính độ đo tương tự càng lớn thì sự tương đồng giữa các đối tượng càng lớn và ngược lại. hàm tính độ phi tương tự tỉ lệ nghịch với hàm tính độ t ương tự.

### 1.2. Bài toán phân cụm dữ liệu

Bài toán phân cụm dữ liệu thường được hiểu là một bài toán học không giám sát và được phát biểu như sau:

Cho tập N đối tượng dữ liệu X= {$x_1$, …, $x_n$ } (bài này ta hạn chế chỉ xét các đối tượng trong không gian số học n - chiều: $x_i \in R^n$), ta cần chia X thành các cụm đôi một không giao nhau: $X = \bigcup_{i=1}^{k} C_i$

sao cho các đối tượng trong cùng một cụm $C_i$ thì tương tự nhau và các đối tượng trong các cụm khác nhau thì khác nhau hơn theo một cách nhìn nào đó.

Số lượng k các cụm có thể cho trước hoặc xác định nhờ phương pháp phân cụm. Để thực hiện phân cụm ta cần xác định được mức độ tương tự giữa các đối tượng, tiêu chuẩn để phân cụm, trên cơ sở đó xây dựng mô hình và các thuật toán phân cụm theo nhiều cách tiếp cận. Mỗi cách tiếp cận cho ta kết quả phân cụm với ý nghĩa sử dụng khác nhau.

### 1.3. Kiểu dữ liệu và độ đo tương tự sử dụng trong bài toán phân cụm dữ liệu

Trong phần này chúng ta phân tích các kiểu dữ liệu thường được sử dụng trong PCDL. Trong PCDL, các đối tượng dữ liệu cần phân tích có thể là con người, cái nhà, tiền lương, các thực thể phần mềm, .... Các đối tượng này thường được diễn tả dưới dạng các đặc tính hay còn gọi là thuộc tính của nó. Các thuộc tính này là các tham số cho giải quyết vấn đề PCDL và sự lựa chọn chúng có tác động đáng kể đến các kết quả của phân cụm. Phân loại khái niệm các kiểu thuộc tính khác nhau là một vấn đề cần giải quyết đối với hầu hết các tập dữ liệu nhằm cung cấp các phương tiện thuận lợi để nhận dạng sự khác nhau của các phần tử dữ liệu. Dưới đây là cách phân lớp dựa trên hai đặc trưng là: kích thước miền (Domain Size) và hệ đo (Measurement Scale)

Cho một CSDL *D* chứa n đối tượng trong không gian k chiều trong đó x, y, z là các đối tượng thuộc *D:* x=($x_1$, $x_2$,. ., $x_k$); y=($y_1$, $y_2$,. ., $y_k$); z=($z_1$, $z_2$,. ., $z_k$),

trong đó $x_i$, $y_i$, $z_i$ với $i = \overline{1,k}$ là các đặc trưng hoặc thuộc tính tương ứng của các đối tượng x, y, z. Vì vậy, hai khái niệm "các kiểu dữ liệu" và "các kiểu thuộc tính dữ liệu" được xem là tương đương với nhau, như vậy, chúng ta sẽ có các kiểu dữ liệu sau :

Phân loại các kiểu dữ liệu dựa trên kích thước miền

- Thuộc tính liên tục (Continuous Attribute): nếu miền giá trị của nó là vô hạn không đếm được, nghĩa là giữa hai giá trị tồn tại vô số giá trị khác. Thí dụ như các thuộc tính về màu, nhiệt độ hoặc cường độ âm thanh.

- Thuộc tính rời rạc (DiscretteAttribute): Nếu miền giá trị của nó là tập hữu hạn, đếm được. Thí dụ như các thuộc tính về số serial của một cuốn sách, số thành viên trong một gia đình, …

Lớp các thuộc tính nhị phân là trường hợp đặc biệt của thuộc tính rời rạc mà miền giá trị của nó chỉ có 2 phần tử được diễn tả như: *Yes / No* hoặc *Nam/Nữ, False/true*, …

Phân loại các kiểu dữ liệu dựa trên hệ đo

Giả sử rằng chúng ta có hai đối tượng x, y và các thuộc tính $x_i$, $y_i$ tương ứng với thuộc tính thứ i của chúng. Chúng ta có các lớp kiểu dữ liệu như sau:

***** Thuộc tính định danh (nominal Scale): đây là dạng thuộc tính khái quát hoá của thuộc tính nhị phân, trong đó miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử - nghĩa là nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là $x \neq y$ hoặc x=y.

***** Thuộc tính có thứ tự (Ordinal Scale): là thuộc tính định danh có thêm tính *thứ tự,* nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì ta có thể xác định là $x \neq y$ hoặc x=y hoặc x>y hoặc x<y.

***** Thuộc tính khoảng (Interval Scale): Nhằm để đo các giá trị theo xấp xỉ tuyến tính. Với thuộc tính khoảng, chúng ta có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì ta

nói x cách y một khoảng $x_i - y_i$ tương ứng với thuộc tính thứ i. Một thí dụ về thuộc tính khoảng như thuộc tính *số Serial* của một đầu sách trong thư viện.

***** Thuộc tính tỉ lệ (Ratio Scale): là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc đầy ý nghĩa, thí dụ như thuộc tính chiều cao hoặc cân nặng lấy điểm 0 làm mốc.

Trong các thuộc tính dữ liệu trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục (Categorical), trong khi đó thì thuộc tính khoảng và thuộc tính tỉ lệ được gọi là thuộc tính số (Numeric).

Người ta còn đặc biệt quan tâm đến dữ liệu không gian (Spatial Data). Đây là loại dữ liệu có các thuộc tính số khái quát trong không gian nhiều chiều, dữ liệu không gian mô tả các thông tin liên quan đến không gian chứa đựng các đối tượng, thí dụ như thông tin về hình học, … Dữ liệu không gian có thể là dữ liệu liên tục hoặc rời rạc:

-Dữ liệu không gian rời rạc: có thể là một điểm trong không gian nhiều chiều và cho phép ta xác định được khoảng cách giữa các đối tượng dữ liệu trong không gian.

-Dữ liệu không gian liên tục: bao chứa một vùng trong không gian.

Thông thường, các thuộc tính số được đo bằng các đơn vị xác định như là *kilogams* hay là *centimeter*. Tuy nhiên, các đơn vị đo có ảnh hưởng đến các kết quả phân cụm. Thí dụ như thay đổi độ đo cho thuộc tính cân nặng từ kilogams sang *Pound* có thể mang lại các kết quả khác nhau trong phân cụm. Để khắc phục điều này người ta phải *chuẩn hoá dữ liệu*, tức là sử dụng các thuộc tính dữ liệu không phụ thuộc vào đơn vị đo. Thực hiện chuẩn hoá phụ thuộc vào ứng dụng và người dùng, thông thường chuẩn hoá dữ liệu được thực hiện bằng cách thay thế mỗi một thuộc tính bằng thuộc tính số hoặc thêm các trọng số cho các thuộc tính.

## 1.4. Khái niệm về tương tự và phi tương tự

Khi các đặc tính của dữ liệu được xác định, người ta đi tìm cách thích hợp để xác định "khoảng cách" giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính *độ tương tự (Similar)* hoặc là tính *độ phi tương tự (Dissimilar)* giữa các đối tượng dữ liệu. Giá trị của hàm tính độ đo tương tự càng lớn thì sự giống nhau giữa đối tượng càng lớn và ngược lại, còn hàm tính độ phi tương tự tỉ lệ nghịch với hàm tính độ tương tự. Độ tương tự hoặc độ phi tương tự có nhiều cách để xác định, chúng thường được đo bằng koảng cách giữa các đối tượng. Tất cả các cách đo độ tương tự đều phụ thuộc vào kiểu thuộc tính mà chúng ta phân tích. Thí dụ, đối với thuộc tính hạng mục (*Categorical*) người ta không sử dụng độ đo khoảng cách mà sử dụng một hướng hình học của dữ liệu.

Tất cả các độ đo dưới đây được xác định trong không đo gian metric. Bất kỳ một metric nào cũng là một độ đo, nhưng điều ngược lại không đúng. Để tránh sự nhầm lẫn, thuật ngữ độ đo ở đây đề cập đến hàm tính độ *tương tự* hoặc hàm tính *độ phi tương tự*. Một không gian metric là một tập trong đó có xác định các "*khoảng cách*" giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các phần tử của nó có thể là những đối tượng bất kỳ) các đối tượng dữ liệu trong CSDL D như đã đề cập ở trên được gọi là một không gian metric nếu

- ✓ Với mỗi cặp phần tử x, y thuộc X đều có xác định, theo một quy tắc nào đó, một số thực δ(x, y), được gọi là khoảng cách giữa x và y.

- ✓ Quy tắc nói trên thoả mãn hệ tính chất sau: *(i)*δ(x, y)>0 nếu x ≠y ; *(ii)*δ(x, y)=0 nếu =y; *(iii)* δ(x, y) = δ(y, x) với mọi x, y; *(iv)* δ(x, y) ≤ δ(x, z)+δ(z, y).

Hàm δ(x, y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.

**Sau đây là các phép đo độ tương tự áp dụng đối với các kiểu dữ liệu khác nhau:**

❖ ***Thuộc tính khoảng***: Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các metric khoảng cách như sau:

-Khoảng cách Minskowski: $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^{q})^{1/q}$, trong đó q là số tự nhiên dương.

-Khoảng cách Euclide: $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=2.

-Khoảng cách Manhattan: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=1.

-Khoảng cách cực đại: $d(x,y) = Max_{i=1}^{n} |x_i - y_i|$, đây là trường hợp của khoảng cách Minskowski trong trường hợp q-> ∞.

❖ ***Thuộc tính nhị phân:*** Trước hết chúng ta có xây dựng bảng tham số sau:

| | y:1 | y:0 | |
|---|---|---|---|
| x:1 | $\alpha$ | $\beta$ | $\alpha+\beta$ |
| x:0 | $\gamma$ | $\delta$ | $\gamma+\delta$ |
| | $\alpha+\gamma$ | $\beta+\delta$ | $\tau$ |

**Hình 8: Bảng tham số**

Trong đó: $\tau = \alpha + \gamma + \beta + \delta$, các đối tượng x, y mà tất cả các thuộc tính tính của nó đều là nhị phân biểu thị bằng 0 và 1. Bảng trên cho ta các thông tin sau:

- $\alpha$ là tổng số các thuộc tính có giá trị là 1 trong cả hai đối tượng x, y.
- $\beta$ là tổng số các giá trị thuộc tính có giá trị là 1 trong x và 0 trong y
- $\gamma$ là tổng số các giá trị thuộc tính có giá trị là 0 trong x và 1 trong y
- $\delta$ là tổng số các giá trị thuộc tính có giá trị là 0 trong x và y

Các phép đo độ tương tương đồng đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau:

-Hệ số đối sánh đơn giản: $d(x,y) = \frac{\alpha + \delta}{\tau}$, ở đây cả hai đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

-Hệ số Jacard: $d(x,y) = \frac{\alpha}{\alpha + \beta + \gamma}$, chú ý rằng tham số này bỏ qua số các đối sánh giữa 0-0. Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

- ***Thuộc tính định danh:*** Độ đo phi tương tự giữa hai đối tượng x và y được định nghĩa như sau: $d(x,y) = \frac{p-m}{p}$, trong đó m là số thuộc tính đối sánh tương ứng trùng nhau, và p là tổng số các thuộc tính.
- ***Thuộc tính có thứ tự:*** Phép đo độ phi tương tự giữa các đối tượng dữ liệu với thuộc tính thứ tự được thực hiện như sau, ở đây ta giả sử i là thuộc tính thứ tự có $M_i$ giá trị ($M_i$ kích thước miền giá trị):

- Các trạng thái $M_i$ được sắp thứ tự như sau: [1…$M_i$], chúng ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại $r_i$, với $r_i \in \{1 \ldots M_i\}$.

-Mỗi một thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy chúng ta chuyển đổi chúng về cùng miền giá trị [0, 1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính: $z_i^{(j)} = \frac{r_i^{(j)} - 1}{M_i - 1}$

-Sử dụng công thức tính độ phi tương tự của *thuộc tính khoảng* đối với các giá trị $z_i^{(j)}$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

❖ ***Thuộc tính tỉ lệ:*** Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính $x_i$, thí dụ $q_i = \log(x_i)$, lúc này $q_i$ đóng vai trò như thuộc tính khoảng (Interval -Scale). Phép biến đổi logarit này thích hợp trong trường hợp các giá trị của thuộc tính là số mũ.

Trong thực tế, khi tính độ đo tương tự dữ liệu, người ta chỉ xem xét một phần các thuộc tính đặc trưng đối với các kiểu dữ liệu hoặc là đánh trọng số cho cho tất cả các thuộc tính dữ liệu. Trong một số trường hợp, người ta loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Các trọng số này có thể sử dụng trong các độ đo khoảng cách trên, thí dụ với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng $w_i$ ($1 \le i \le k$), độ tương đồng dữ liệu được xác định như sau:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}.$$

Người ta có thể chuyển đổi giữa các mô hình cho các kiểu dữ liệu trên, thí dụ dữ liệu kiểu hạng mục có thể chuyển đổi thành dữ liệu nhị phân và ngược lại. Thế nhưng, giải pháp này rất tốt kém về chi phí tính toán, do vậy, cần phải cân nhắc khi áp dụng cách thức này.

Tóm lại, tuỳ từng trường hợp dữ liệu cụ thể mà người ta sử dụng các mô hình tính độ tương tự khác nhau. Việc xác định độ tương đồng dữ liệu thích hợp, chính xác, đảm bảo khách quan là rất quan trọng, góp phần xây dựng thuật toán PCDL có hiệu quả cao trong việc đảm bảo chất lượng cũng như chi phí tính toán của thuật toán.

## 1.5. Ứng dụng của phân cụm dữ liệu

Phân cụm DL là một trong những công cụ chính được ứng dụng trong nhiều lĩnh vực như thương mại và khoa học. Các kỹ thuật PCDL đã được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau**:**

- Thương mại: Trong thương mại, PCDL có thể giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu mua bán trong CSDL khách hàng.

- Sinh học: Trong sinh học, PCDL được sử dụng để xác định các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.

- Phân tích dữ liệu không gian: Do sự đồ sộ của dữ liệu không gian như dữ liệu thu được từ các hình ảnh chụp từ vệ tinh các thiết bị y học hoặc hệ thống thông tin địa lý (GIS), …làm cho người dùng rất khó để kiểm tra các dữ liệu không gian một cách chi tiết. PCDL có thể trợ giúp người dùng tự động phân tích và xử lý các dữ liêu không gian như nhận dạng và chiết xuất các đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong CSDL không gian.

- Lập quy hoạch đô thị*:* Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý, …nhằm cung cấp thông tin cho quy hoạch đô thị.

- Nghiên cứu trái đất: Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.

- Địa lý: Phân lớp các động vật và thực vật và đưa ra đặc trưng của chúng.

- Web Mining: PCDL có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa  trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu, …

# CHƯƠNG 2:
# PHÂN CỤM DỮ LIỆU PHÂN HOẠCH

## 2.1. Giới thiệu:

Có nhiều kỹ thuật phân cụm dữ liệu khác nhau. Việc lựa chọn phương pháp tuỳ thuộc vào yêu cầu cụ thể. Trong bài đồ án này trình bày về kỹ thuật phân cụm dữ liệu phân hoạch bởi cơ sở dữ liệu ta tiến hành nghiên cứu là cơ sở dữ liệu không gian tĩnh có chứa nhiễu.

Phương pháp phân cụm phân hoạch nhằm phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho: mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Các thuật toán phân hoạch dữ liệu có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề PCDL, do nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế người ta thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của các cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Với chiến lược này, thông thường người ta bắt đầu khởi tạo một phân hoạch ban đầu cho tập dữ liệu theo phép ngẫu nhiên hoặc theo heuristic, và liên tục tinh chỉnh nó cho đến khi thu được một phân hoạch mong muốn, thoả mãn ràng buộc cho trước. Các thuật toán phân cụm phân hoạch cố gắng cải tiến tiêu chuẩn phân cụm, bằng cách tính các giá trị đo độ tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược tham lam (Greedy) để tìm kiếm nghiệm.

Một số thuật toán phân cụm phân hoạch điển hình như : K-MEANS, PAM, CLARA, CLARANS:

- **Thuật toán K-MEANS**:có thể áp dụng đối với tập dữ liệu lớn.Tuy nhiên nó chỉ áp dụng với dữ liệu có thuộc tính số và khám phá ra các cụm có dạng hình cầu. Kmeans còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.Chất lượng phân cụm phụ thuộc nhiều vào các tham số đầu vào như số cụm k và k trọng tâm khởi tạo ban đầu.

- **Thuật toán PAM:** khắc phục nhược điểm của thuật toán KMEANS là có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc các phần tử ngoại lai. PAM áp dụng cho dữ liệu không gian, tuy nhiên PAM kém hiệu quả về thời gian tính toán khi giá trị của k và n là lớn.

- **Thuật toán CLARA:** thuật toán này nhằm khắc phục nhược điểm của thuật toán PAM trong trường hợp giá trị của k và n là lớn.CLARA tiến hành bằng cách trích mẫu cho tập dữ liệu có n phần tử và áp dụng thuật toán PAM cho mẫu này và tìm ra các đối tượng tâm medoid cho mẫu được trích từ dữ liệu này.

- **Thuật toán CLARANS:** nhằm để cải tiến cho chất lượng cũng như mở rộng áp dụng cho tập dữ liệu lớn. CLARANS cũng sử dụng các đối tượng trung tâm medoid làm đại diện cho các cụm dữ liệu. Ưu điểm của CLARANS là không gian tìm kiếm không bị giới hạn như đối với CLARA và trong cùng một lượng thời gian thì chất lượng của các cụm phân được là lớn hơn so với CLARA.

## 2.2. Thuật toán K-means:

Thuật toán phân hoạch K-means do MacQeen đề xuất trong lĩnh vực thống kê năm 1967, mục đích của thuật toán k-means là sinh ra k cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ $(i = \overline{1,n})$, sao cho hàm tiêu chuẩn: $E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x - m_i)$ đạt giá trị tối thiểu.

Trong đó: $m_i$ là trọng tâm của cụm $C_i$, D là khoảng cách giữa hai đối tượng.

Trọng tâm của một cụm là một véc tơ, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng vectơ dữ liệu trong cụm đang xét. Tham số đầu vào của thuật toán là số cụm k, và tham số đầu ra của thuật toán là các trọng tâm của các cụm dữ liệu. Độ đo khoảng cách D giữa các đối tượng dữ liệu thường được sử dụng dụng là khoảng cách Euclide, bởi vì đây là mô hình khoảng cách dễ để lấy đạo hàm và xác định các cực trị tối thiểu. Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tuỳ vào ứng dụng hoặc các quan điểm của người dùng.

**Thuật toán k-means bao gồm các bước cơ bản như sau:**

Bước 1. Chọn k phần tử ban đầu $z^j \; {}^{k}_{j=1}$ của D làm tâm các cụm con.

Bước 2. Với mỗi i = 1,. . . ., N ; xếp $x^i$ vào cụm $C_j$ nếu:

$$d(x^i, z^j) = \min \; d(x^i, z^q) / q \le k$$

Bước 3. Tính trung bình cộng của các phần của các cụm $C^j$ làm tâm mới:

$$z^j \leftarrow \frac{1}{\left|c^j\right|} \sum_{x \in C^j} x,$$

trong đó $\left|C^j\right|$ là số phần tử của cụm $C^j$.

Bước 4. Trở lại bước 2 để xếp lại các cụm con nhờ tâm mới cho tới khi các cụm không thay đổi.

Nếu mêtric được dùng là mêtric Euclide thì thuật toán hội tụ tới cực tiểu địa phương của hàm:

$$E = \sum_{j=1}^{k} \sum_{x \in c^j} d^2(x, z^j)$$

K-means biểu diễn các cụm bởi các trọng tâm của các đối tượng trong cụm đó.

Thuật toán k-means chi tiết được trình bày như sau:

BEGIN

1. Write (“Nhập số đối tượng dữ liệu”);readln(n);
2. Nhập n đối tượng dữ liệu;
3. Write (“Nhập số cụm dữ liệu”);readln(k);
4. MSE = + $\infty$;
5. For i = 1 to k do $m_i = x_{i+(i-1)*[n/k]}$; //Khởi tạo k trọng tâm
6. Do{
7. OldMSE = MSE;
8. MSE' = 0;
9. For j = 1 to k do
10. {$m'_j$ =0; $n'_j$ = 0;}
11. Endfor;
12. For i = 1 to n do
12. For j = 1 to k do

Tính toán khoảng cách Euclide

14. bình phương: $D^2(x_i, m_j)$;
15. Endfor
16. Tìm trọng tâm gần nhất $m_h$ tới $X_i$.
17. $m'_h = m'_h + X_i$; $n'_h = n'_h + 1$;
18. MSE' = MSE' + $D^2(x_i, m_h)$;
19. Endfor
20. $n_j$ = max ($n'_j$, 1); $m_j = m'_j / n_j$ ;
21. Endfor
22. MSE = MSE';
23} while (MSE < OldMSE)

END;

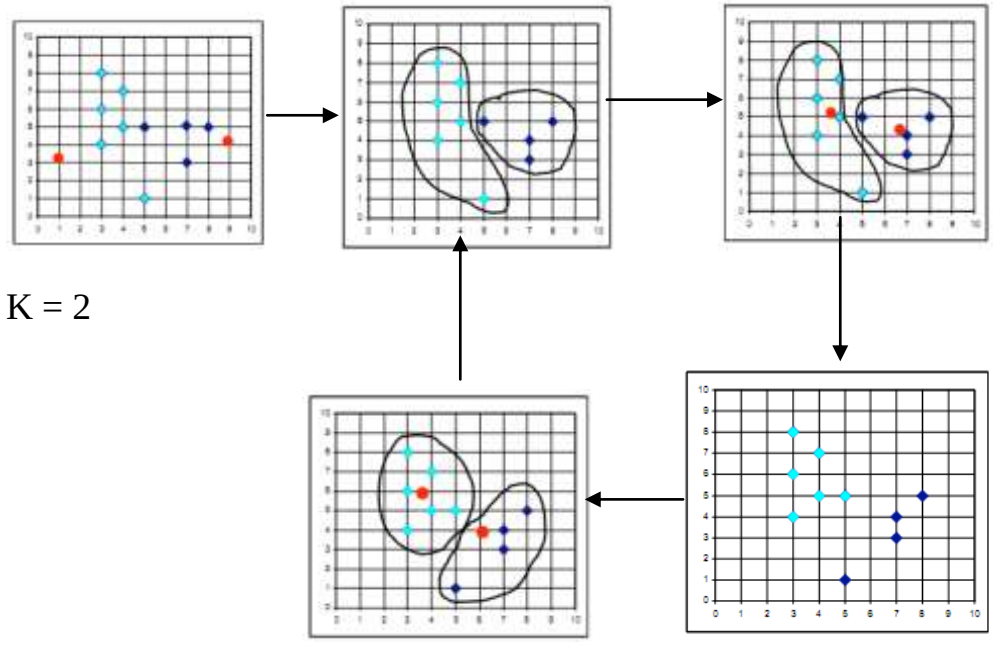
**Hình 9: Thuật toán k-means chi tiết**

Các khái niệm biến và hàm sử dụng trong thuật toán k-means trong hình 9 như sau:

- *MSE (Mean Squared Error)*: được gọi là sai số bình phương trung bình hay còn gọi là hàm tiêu chuẩn, MSE dùng để lưu giá trị của hàm tiêu chuẩn và được cập nhật qua mỗi lần lặp. Thuật toán dừng ngay khi giá MSE tăng lên so với giá trị MSE cũ của vòng lặp trước đó.
- $D^2(x_i, m_j)$*:* là khoảng cách Euclide từ đối tượng dữ liệu thứ i tới trọng tâm thứ j;
- $OldMSE, \overline{m'_j}, n'_j$: là các biến tạm lưu giá trị cho các trạng thái trung gian cho các biến tương ứng: giá trị hàm tiêu chuẩn, giá trị của vectơ tổng của các đối tượng trong cụm thứ j, số các đối tượng của cụm thứ j;

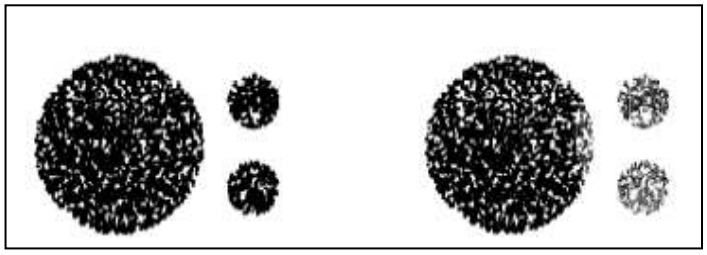
Thuật toán k-means tuần tự trên được chứng minh là hội tụ và có độ phức tạp tính toán là: $O((3nkd)\tau T^{flop})$. Trong đó: n là số đối tượng dữ liệu, k là số cụm dữ liệu, d là số chiều, $\tau$ là số vòng lặp, $T^{flop}$ là thời gian để thực hiện một phép tính cơ sở như phép tính nhân, chia, …Như vậy, do k-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.

Phương pháp k-mean có thể biểu diễn qua hình ảnh sau:




**Hình 10: Mô phỏngquá trình phân cụm của k-mean**

Tuy nhiên, nhược điểm của k-means là chỉ áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu, k-means còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu. Hình 11 diễn tả môi phỏng về một số hình dạng cụm dữ liệu khám phá được bởi k-means:



**Hình 11: Thí dụ về một số hình dạng cụm dữ liệu được khám phá bởi k-means**

Hơn nữa, chất lượng phân cụm dữ liệu của thuật toán k-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k trọng tâm khởi tạo ban đầu. Trong trường hợp, các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của k-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế người ta chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất.

Đến nay, đã có rất nhiều thuật toán kế thừa tư tưởng của thuật toán k-means áp dụng trong Data Mining để giải quyết với tập dữ liệu có kích thước rất lớn đang được áp dụng rất hiệu quả và phổ biến như thuật toán k-modes, PAM, CLARA, CLARANS, k- prototypes, …

### 2.3.Thuật toán PAM

PAM (Partitioning Around Medoids) là thuật toán mở rộng của thuật toán k-means, nhằm có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc các phần tử ngoại lai, đây là thụât toán PCDL được đề xuất bởi Kaufman và Rousseeuw. Thay vì sử dụng các trọng tâm như k-means, PAM sử dụng các đối tượng *medoid* để biểu diễn cho các cụm dữ liệu, một đối tượng *medoid* là đối tượng đặt tại vị trí trung tâm nhất bên trong của mỗi cụm. Vì vậy, các đối tượng *medoid* ít

bị ảnh hưởng của các đối tượng ở rất xa trung tâm, trong khi đó các trọng tâm của thuật toán k-means lại rất bị tác động bởi các điểm xa trung tâm này. Ban đầu, PAM khởi tạo k đối tượng *medoid* và phân phối các đối tượng còn lại vào các cụm với các đối tượng *medoid* đại diện tương ứng sao cho chúng tương tự với đối tượng *medoid* trong cụm nhất.

Thí dụ: Nếu $O_j$ là đối tượng không phải là **medoid** và $O_m$ là một đối tượng **medoid,** khi đó ta nói $O_j$ thuộc về cụm có đối tượng **medoid** là $O_m$ làm đại diện nếu: $d(O_j, O_m) = \min_{Oe} d(O_j, O_e)$. Trong đó: $d(O_j, O_e)$ là độ phi tương tự giữa $O_j$ và $O_e$, $\min_{Oe}$ là giá trị nhỏ nhất của độ phi tương tự giữa $O_j$ và tất cả các đối tượng **medoid** của các cụm dữ liệu. Chất lượng của mỗi cụm được khám phá được đánh giá thông qua độ phi tương tự trung bình giữa một đối tượng và đối tượng **medoid** tương ứng với cụm của nó, nghĩa là chất lượng phân cụm được đánh giá thông qua chất lượng của tất cả các đối tượng **medoid.** Độ phi tương tự ở đây thông thường được xác định bằng độ đo khoảng cách, thuật toán PAM ở đây được áp dụng cho dữ liệu không gian.

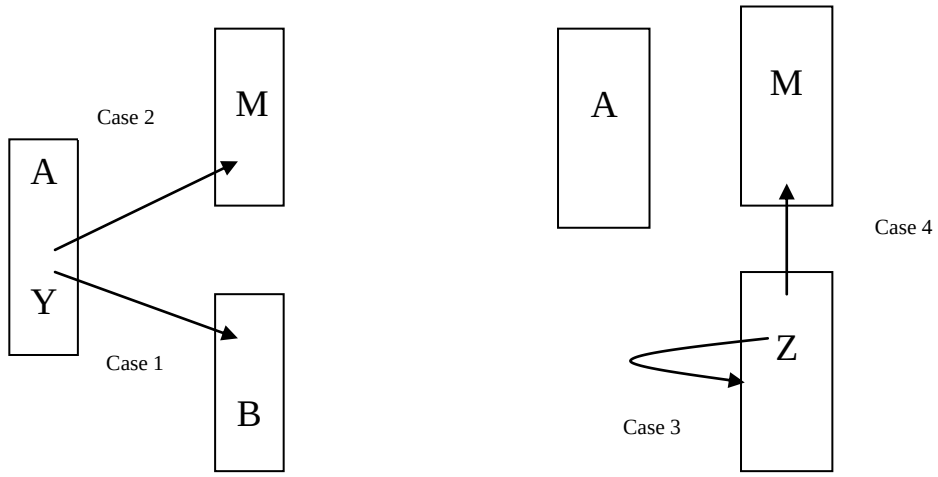
Để xác định các *medoid***,** PAM bắt đầu bằng cách lựa chọn k đối tượng *medoid* bất kỳ. Sau mỗi bước thực hiện, PAM cố gắng hoán chuyển giữa đối tượng *medoid* $O_m$ và một đối tượng $O_p$ không phải là *medoid***,** miễn là sự hoán chuyển này nhằm cải tiến chất lượng của phân cụm, quá trình này kết thúc khi chất lượng phân cụm không thay đổi. Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn, chất lượng phân cụm tốt nhất khi hàm tiêu chuẩn đạt giá trị tối thiểu.

**Cụ thể ta xét ví dụ sau:**

Cho hai đối tượng **medoid** A và B. Đối với tất cả các đối tượng Y thuộc cụm với đối tượng **medoid** đại diện A, chúng ta tìm **medoid** của cụm gần nhất để thay thể. Có hai trường hợp có thể xẩy ra, hoặc Y được chuyển tới cụm dữ liệu có đại diện là B hoặc được chuyển tới cụm dữ liệu có đại diện là M. Tiếp đến, chúng ta xét lần lượt cho tất cả các đối tượng trong cụm có đại diện là A. Tương tự như vậy, đối với tất các các đối tượng trong cụm có đối tượng đại diện

là B, chúng ta có thể di chuyển chúng tới cụm có đại diện là M hoặc là chúng ở lại B.

Thí dụ này có thể biểu diễn như hình 12 dưới đây:




**Hình 12: Thí dụ về các khả năng thay thế các đối tượng tâm *medoid***

Sau đây là một số khái niệm biến được sử dụng cho thuật toán PAM:

- $O_m$ : Là đối tượng **medoid** hiện thời cần được thay thế
- $O_p$: là đối tượng medoid mới thay thế cho $O_m$;
- $O_j$: Là đối tượng dữ liệu (không phải là **medoid**) có thể được di chuyển sang cụm khác.
- $O_{j,2}$: Là đối tượng **medoid** hiện thời gần đối tượng $O_j$ nhất mà không phải là các đối tượng A và M như trong ví dụ trên.

Bốn trường hợp như mô tả trong thí dụ trên, PAM tính giá trị $C_{jmp}$ cho tất cả các đối tượng $O_j$. $C_{jmp}$ ở đây nhằm để làm căn cứ cho việc hoán chuyển giữa $O_m$ và $O_p$. Các cách tính. Trong mỗi trường hợp $C_{jmp}$ được tính với 4 cách khác nhau như sau:

- **Trường hợp 1**: Giả sử $O_j$ hiện thời thuộc về cụm có đại diện là $O_m$ và $O_j$ tương tự với $O_{j,2}$ hơn $O_p$ ($d(O_j, O_p) \geq d(O_j, O_{j,2})$). Trong khi đó, $O_{j,2}$ là đối tượng **medoid** tương tự xếp thứ 2 tới $O_j$ trong số các **medoid.** Trong trường hợp

này, chúng ta thay thế $O_m$ bởi đối tượng medoid mới $O_p$ và $O_j$ sẽ thuộc về cụm có đối tượng đại diện là $O_{j,2}$. Vì vậy, giá trị hoán chuyển $C_{jmp}$ được xác định như sau:

$C_{jmp} = d(O_j, O_{j,2}) – d(O_j, O_m)$. (1)

Giá trị $C_{jmp}$ là không âm.

- **Trường hợp 2:** $O_j$ hiện thời thuộc về cụm có đại diện là $O_{m,}$ nhưng $O_j$ ít tương tự với $O_{j,2}$ so với $O_p$ (Nghĩa là, $d(O_j, O_p)<d(O_j, O_{j,2})$). Nếu $O_m$, được thay thế bởi $O_p$ thì $O_j$ sẽ thuộc về cụm có đại diện là $O_p$. Vì vậy, giá trị $C_{jmp}$ được xác định như sau: $C_{jmp}= (O_j, O_p) - d(O_j, O_m)$ (2). $C_{jmp}$ ở đây có thể là âm hoặc dương.

- **Trường hợp 3:** Giả sử $O_j$ hiện thời không thuộc về cụm có đối tượng đại diện là $O_m$ mà thuộc về cụm có đại diện là $O_{j,2}$. Mặt khác, giả sử $O_j$ tương tự với $O_{j,2}$ hơn so với $O_p$, khi đó, nếu $O_m$ được thay thế bởi $O_p$ thì $O_j$ vẫn sẽ ở lại trong cụm có đại diện là $O_{j,2}$. Do đó: $C_{jmp}= 0$ (3).

- **Trường hợp 4:** $O_j$ hiện thời thuộc về cụm có đại diện là $O_{j,2}$ nhưng $O_j$ ít tương tự tới $O_{j,2}$ hơn so với $O_{p.}$ Vì vậy, nếu chúng ta thay thế $O_m$ bởi $O_p$ thì $O_j$ sẽ chuyển từ cụm $O_{j,2}$ sang cụm $O_p$. Do đó, giá trị hoán chuyển $C_{jmp}$ được xác định là: $C_{jmp}= (O_j, O_p) - d(O_j, O_{j,2})$ (4). $C_{jmp}$ ở đây luôn âm.

- Kết hợp cả bốn trường hợp trên, tổng giá trị hoán chuyển $O_m$ bằng $O_p$ được xác định như sau: $TC_{mp} = \sum_j C_{jmp}$ (5). Sử dụng các khái niệm trên, thuật toán PAM có các bước thực hiện như hình 13 sau:

**Input**: Tập dữ liệu có n phần tử, số cụm k

**Out Put**: k cụm dữ liệu sao cho chất lượng phân hoạch là tốt nhất.

**BEGIN**

Bước 1: Chọn k đối tượng medoid bất kỳ;

2: Tính $TC_{mp}$ cho tất cả các cặp đối tượng $O_m$, $O_p$. Trong đó $O_m$ là đối tượng medoid và $O_p$ là đối tượng không phải là modoid.

3: Chọn cặp đối tượng $O_m$ và $O_p$. Tính $min_{Om,}$ $min_{Op}$, $TC_{mp}$.

$TC_{mp}$ là âm, thay thế $O_m$ bởi $O_p$ và quay lại bước 2. Nếu $TC_{mp}$ dương, chuyển sang bước 4.

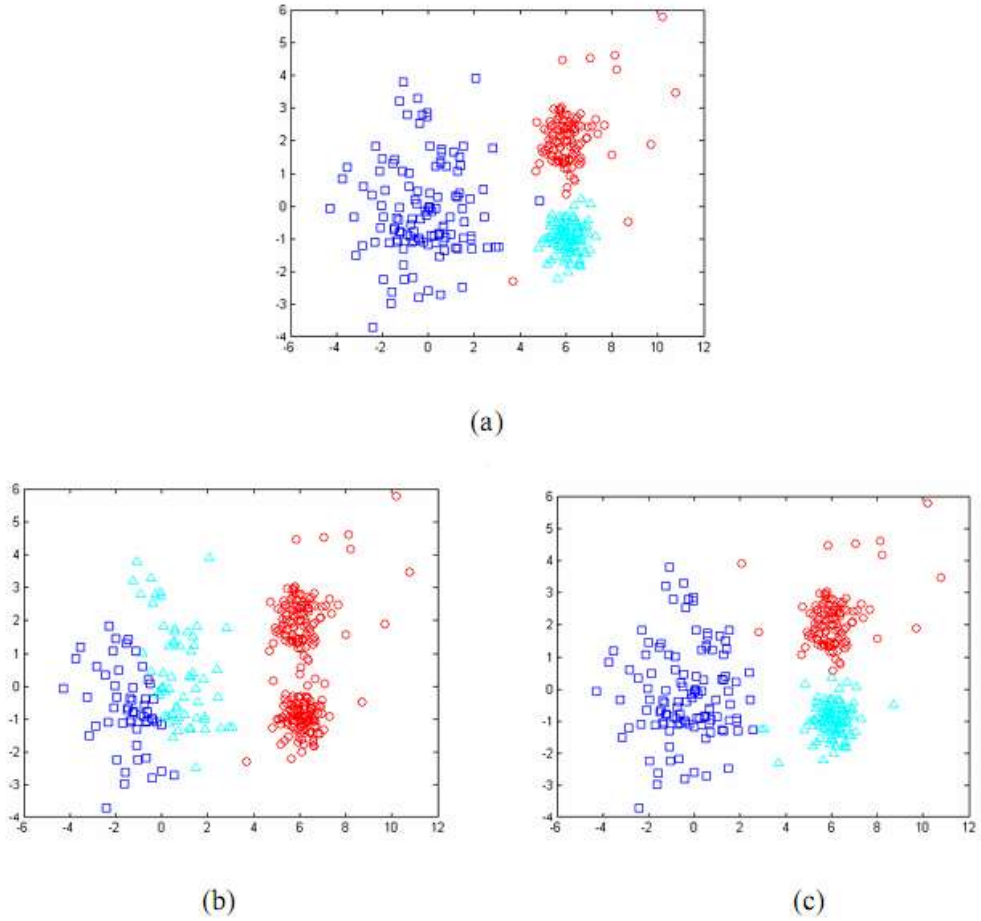
4: Với mỗi đối tượng không phải là medoid, xác định đối tượng medoid tương tự với nó nhất đồng thời gán nhãn cụm cho chúng.

**END**

**Hình 13: Các bước thực hiện của thuật toán PAM**

Trong bước 2 và 3, có PAM phải duyệt tất cả k(n-k) cặp $O_m$, $O_p$. Với mỗi vặp, việc tính toán $TC_{mp}$ yêu cầu kiểm tra n-k đối tượng. Vì vậy, độ phức tạp tính toán của PAM là $O(Ik\ (n-k)^2)$, trong đó I là số vòng lặp. Như vậy, thuật toán PAM kém hiệu quả về thời gian tính toán khi giá trị của k và n là lớn.

Sau đây là hình ảnh mô phỏng vấn đề phân cụm dữ liệu của thuật toán K-MEANS và PAM:




**Hình 14: Mô phỏng kết quả của 2 thuật toán k-means và pam**

- Hình (a) là mô phỏng dữ liệu ban đầu.

- Hình (b) là mô phỏng kết quả phân cụm dữ liệu sau khi áp dụng thuật toán k-means.

- Hình (c) là mô phỏng kết quả phân cụm dữ liệu của thuật toán pam.

## 2.4. Thuật toán CLARA

CLARA (Clustering LARge Application) được Kaufman đề xuất năm 1990, thuật toán này nhằm khắc phục nhược điểm của thuật toán PAM trong trường hợp giá trị của k và n là lớn. CLARA tiến hành trích mẫu cho tập dữ liệu có n phần tử, nó áp dụng thuật toán PAM cho mẫu này và tìm ra các các đối tượng tâm *medoid* cho mẫu được trích từ dữ liệu này. Người ta thấy rằng, nếu mẫu dữ liệu được trích theo cách ngẫu nhiên, thì các medoid của nó xấp xỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốt hơn, CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗi trường hợp

và tiến hành chọn kết quả phân cụm tốt nhất khi thực hiện phân cụm trên các mẫu này. Để cho chính xác, chất lượng của các cụm được đánh giá thông độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệu trong tập đối tượng ban đầu. Kết quả thực nghiệm chỉ ra rằng, 5 mẫu dữ liệu có kích thước 40+2k cho các kết quả tốt.

Các bước thực hiện của thuật toán CLARA như hình 15 sau.

BEGIN

1. For i = 1 to 5 do

2. Lấy một mẫu có 40 + 2k đối tượng dữ liệu ngẫu nhiên từ tập dữ liệu và áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các đối tượng medoid đại diện cho các cụm.

3. Đối với mỗi đối tượng $O_j$ trong tập dữ liệu ban đầu, xác định đối tượng medoid tương tự nhất trong số k đối tượng medoid.

4. Tính độ phi tương tự trung bình cho phân hoạch các đối tượng dành ở bước trước, nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.

5. Quay về bước 1.

END;

**Hình 15: Các bước thực hiện của thuật toán CLARA**

Độ phức tạp tính toán của nó là $O(k(40+k)^2 + k(n-k))$, và CLARA có thể thực hiện đối với tập dữ liệu lớn. Chú ý đối với kỹ thuật tạo mẫu trong PCDL: kết quả phân cụm có thể không phụ thuộc vào tập dữ liệu khởi tạo nhưng nó chỉ đạt tối ưu cục bộ. Thí dụ như: Nếu các đối tượng *medoid* của dữ liệu khởi tạo không nằm trong mẫu, khi đó kết quả thu được không đảm bảo là tốt nhất được.

## 2.5. Thuật toán CLARANS

Thuật toán CLARANS được Ng & Han đề xuất năm 1994 nhằm để cải tiến cho chất lượng cũng như mở rộng áp dụng cho tập dữ liệu lớn. CLARANS cũng sử dụng các đối tượng trung tâm medoids làm đại diện cho các cụm dữ liệu.

Như đã biết, PAM là thuật toán phân hoạch có kiểu k-medoids. Nó bắt đầu khởi tạo k tâm đại diện **medoid** và liên tục thay thế mỗi tâm bởi một đối tượng khác trong cụm cho đến khi là tổng khoảng cách của các đối tượng đến tâm cụm không giảm. CLARANS là thuật toán PCDL kết hợp thuật toán PAM với chiến lược tìm kiểm kinh nghiệm mới. Ý tưởng cơ bản của CLARANS là không xem xét tất cả các khả năng có thể thay thể các đối tượng tâm medoids bởi một đối tượng khác, nó ngay lập tức thay thế các đối tượng tâm này nếu việc thay thế này có tác động tốt đến chất lượng phân cụm chứ không cấn xác định cách thay thể tối ưu nhất. Một phân hoạch cụm phát hiện được sau khi thay thế đối tượng trung tâm được gọi là một láng giềng (Neighbor) của phân hoạch cụm trước đó. Số các láng giềng được hạn chế bởi tham số do người dùng đưa vào là **Maxneighbor**, quá trình lựa chọn các láng giềng này là hoàn toàn ngẫu nhiên. Tham số **Numlocal** cho phép người dùng xác định số vòng lặp tối ưu cục bộ được tìm kiếm. Không phải tất các các láng giềng được duyệt mà chỉ có **Maxneighbor** số láng giềng được duyệt.

Thuật toán chi tiết CLARANS như biểu diễn trong hình 16 sau:

```
Input: O, k, dist, numlocal, and maxneighbor;
Output: k cụm dữ liệu;
CLARANS (int k, function dist, int numlocal, int maxneighbor)
BEGIN
    for (i = 1; i <= numlocal; i++) {
        current. create_randomly(k);
        j = 1;
        while (j < maxneighbor) {
            current. select_randomly(old, new);
            diff = current. calculate_distance_difference(old, new);
            if (diff < 0) {
                current. exchange(old, new);
                j = 1;
            }
            else j++; // end if
        } // end while
        dist = current. calculate_total_distance();
        if (dist < smallest_dist) {
          best = current;
           smallest_dist = dist;
        } // end if
    } // end for
END;
```

**Hình 16: Thuật toán CLARANS**

**Trong đó:**

- **Create_Randomly(k)**: tạo ngẫu nhiên k cụm dữ liệu, nghĩa là thuật toán lựa chọn ngẫu nhiên k đối tượng medoid từ n đối tượng dữ liệu.

- **Select_randomly(old, new):** Thay thế một đối tượng tâm cụm medoid **old** bởi đối tượng khác **new.**

- **Calculate_distance_difference(old, new)**: Tính toán sự khác nhau về tổng khoảng cách giữa phân hoạch hiện thời và láng giềng của nó.

- **Exchange(old, new):** Hoán đổi giữa đối tượng tâm cụm medoid old với đối tượng không phải là medoid new, sau khi hoán đổi vai trò của chúng cũng được hoán đổi.

- **Calculate_total_distance()**: Tính tổng khoảng cách cho mỗi phân hoạch.

Như vậy, quá trình hoạt động của CLARANS tương tự với quá trình hoạt động của thuật toán CLARA. Tuy nhiên, ở giai đoạn lựa chọn các trung tâm **medoid** của cụm dữ liệu, CLARANS lựa chọn một giải pháp tốt hơn bằng cách lấy ngẫu nhiên một đối tượng của k đối tượng trung tâm **medoid** của cụm và cố gắng thay thế nó với một đối tượng được chọn ngẫu nhiên trong (n-k) đối tượng còn lại, nếu không có giải pháp nào tốt hơn sau một số cố gắng lựa chọn ngẫu nhiên xác định, thuật toán dùng và cho kết quả phân cụm tối ưu cục bộ.

Trong trường hợp tệ nhất, CLARANS so sánh một đối tượng với tất các đối tượng **Medoid**. Vì vậy, độ phức tạp tính toán của CLARANS là $O(kn^2)$, do vậy CLARANS không thích hợp với tập dữ liệu lớn (khi trường hợp xấu nhất xẩy ra). CLARANS có ưu điểm là không gian tìm kiếm không bị giới hạn như đối với CLARA, và trong cùng một lượng thời gian thì chất lượng của các cụm phân được là lớn hơn so với CLARA.

## 2.5. Nhận xét chung về họ các thuật toán phân hoạch

Thuật toán k-means chỉ thích hợp để tìm kiếm các cụm dữ liệu có dạng hình cầu, không thích hợp cho việc xác định các cụm với hình dạng bất kỳ, nhưng trong trường hợp các cụm khá gần nhau thì một số đối tượng của một cụm có thể là nằm cuối các trong các cụm khác. Thuật toán PAM là một cải tiến của k-means nhằm để khắc phục trong trường hợp dữ liệu chứa nhiễu hoặc các phần tử ngoại lai. CLARA và CLARANS là các thuật toán dựa trên hàm tiêu

chuẩn của thuật toán PAM, đây là các thuật toán có khả năng áp dụng với tập dữ liệu lớn, nhưng hiệu quả của chúng phụ thuộc vào kích thước của các mẫu được phân. Thuật toán CLARANS hiệu quả hơn so với thuật toán CLARA. Hạn chế chung của các thuật toán phân cụm phân hoạch là chỉ thích hợp đối với dữ liệu số và ít chiều, và chỉ khám phá ra các cụm dạng hình cầu, thế nhưng chúng lại áp dụng tốt với dữ liệu có các cụm phân bố độc lập và trong mỗi cụm có mật độ phân bố cao.

# CHƯƠNG 3:
# CÀI ĐẶT CHƯƠNG TRÌNH

## 3.1.Bài toán

**Input:** Có một tập rất lớn các điểm ảnh và phân ra làm k cụm.

**Output:** Các nhóm (cụm) điểm ảnh, trong đó các điểm ảnh có cùng màu sẽ được gom vào một nhóm.

Thuật toán phân cụm phân hoạch (kmean, pam) với các dữ liệu đầu vào là các điểm ảnh có 2 trường giá trị:

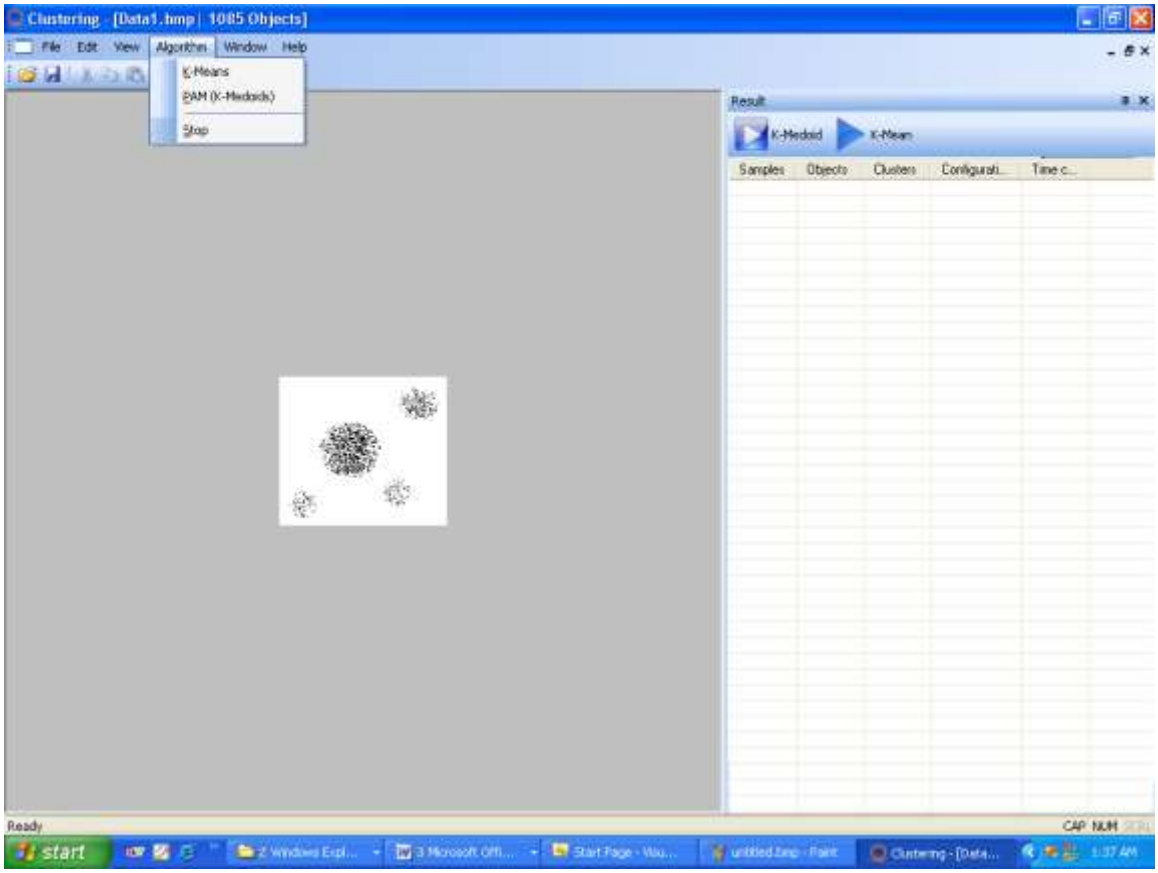
X:toạ độ x của điểm

Y:toạ độ y của điểm

## 3.2.Giới thiệu chương trình ứng dụng

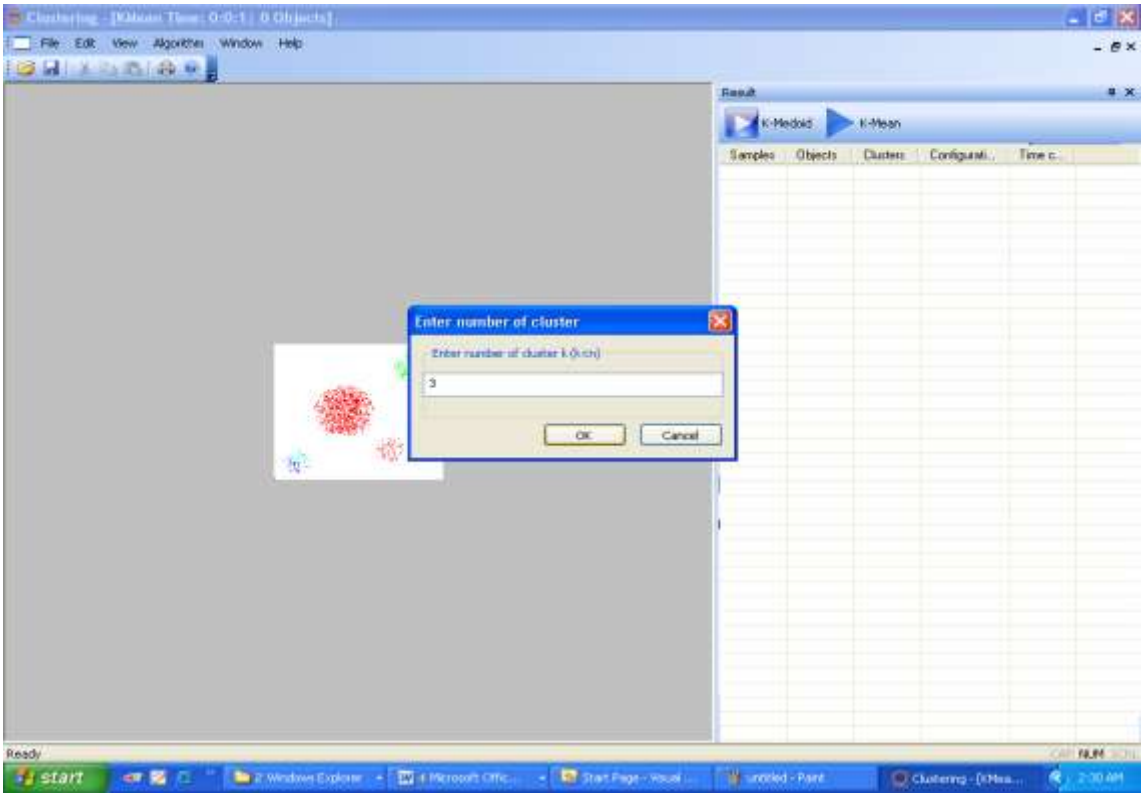
**Giao diện chương trình:**

Chương trình được viết bằng ngôn ngữ visual C++, chạy trên visual studio 2008.

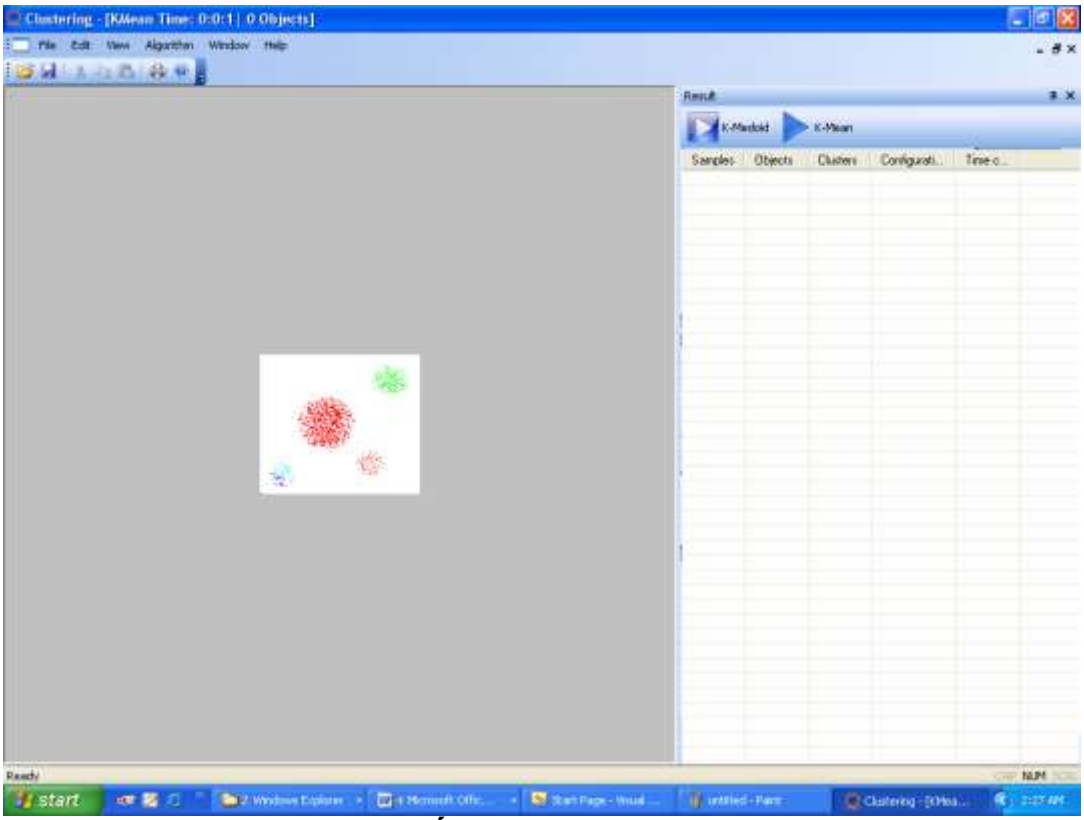



**Hình 9 : Giao diện chính của chương trình**

Khi chạy chương trình,các điểm ảnh được gọi ra. Ta sẽ chọn giá trị tham số là số cụm trong khung bên, chương trình sẽ chạy  thuật toán KMEANS và PAM và tiến hành phân cụm.Kết quả như sau:




**Hình 10: Yêu cầu nhập số cụm**



**Hình 11: Kết quả chạy chương trình**

**Nhận xét chương trình:**

Đây là chương trình thực hiện phân cụm trên một bài toán cụ thể, qua đó cho ta kiểm nghiệm được kết quả của thuật toán phân cụm dữ liệu phân hoạch KMEANS, PAM.

- Chương trình đã chạy và cho ra kết quả phân cụm các điểm ảnh theo tham số là số cụm nhập vào từ người dùng
- Ưu điểm của chương trình là thời gian chạy chương trình nhanh, có thể chạy trên cơ sở dữ liệu lớn.
- Tuy nhiên, thuật toán vẫn còn nhiều hạn chế nhất định được bộc lộ rõ nhất là tham số k được nhập vào từ phía người dùng làm giảm hiệu quả và tính tự động của thuật toán.

# KẾT LUẬN

## Kết quả đạt được của đồ án

Phân cụm dữ liệu trong lĩnh vực Data Mining là một hướng nghiên cứu rất quan trọng. Hiện nay, tuy có ít các kết quả khoa học mới trong PCDL, nhưng do các hệ thống CSDL ngày càng đa dạng, và tăng trưởng nhanh cả về chất lẫn về lượng. Hơn nữa, nhu cầu về khai thác các tri thức từ các CSDL này ngày càng lớn. Vì vậy, việc nghiên cứu các mô hình dữ liệu mới và hoàn thiện và áp dụng các phương pháp và kỹ thuật PCDL là việc làm rất cần thiết và có nhiều ý nghĩa trong khoa học cũng như trong thực tiễn.

Trong đồ án này, em đã trình bày tổng quan về phân cụm dữ liệu bao gồm các kiểu dữ liệu có thể phân cụm, các ứng dụng và các kỹ thuật phân cụm dữ liệu. từ đó, em tập trung đi sâu nghiên cứu về kỹ thuật phân cụm dữ liệu  phân hoạch  và các thuật toán điển hình của kỹ thuật này. Nhưng do thời gian có hạn em chỉ tập trung vào thuật toán KMEAN và PAM với cách thức tổ chức dữ liệu, thuật toán, đánh giá ưu nhược điểm của thuật toán.

Tuy nhiên, do năng lực và trình độ có hạn nên trong quá trình thực hiện đồ án em đã không tránh khỏi những thiếu sót. Kính mong các thầy cô và các bạn quan tâm giúp đỡ chỉ bảo để chương trình của em ngày m ột hoàn thiện hơn.

## Hướng nghiên cứu phát triển của đề tài trong thời gian tới

- Tìm hiểu và thử nghiệm thuật toán với một số ứng dụng khác trên thực tế.

# TÀI LIỆU THAM KHẢO


[1]. Nguyễn Thị Ngọc, *Phân cụm dữ liệu dựa trên mật độ*, Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDL Hải Phòng, 2008.

[2]. Trần Thị Quỳnh, *Thuật toán phân cụm dữ liệu nửa giám sát và giải thuật di truyền*, Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDL Hải Phòng, 2008.

[3]. Nguyễn. . Lâm, *Thuật toán phân cụm dữ liệu nửa giám sát*,- Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDL Hải Phòng, 2007.

[4]. Charles Elkan, *Department of Comput er Science and Engineering*, University of California, San Diego La jolla, CA 92093.

[5]. Andre w W. *Moore Associate Professoor School of Computer Science,* Carnegil Mellon University.

[6]. J.Han, M. Kamber and A.K.H. Tung, *Spatial Clustering Methods in Data Mining*, Sciences and Engineering Research Council of Canadia.

[7] Hae – sang Park, jong- Seok Lee and Chi- Hyuck jun

*Department of Industrial and Management Engineering* , POSTECH

San 31 Hyọa – dong , Pohang 790 – 784 , S.Korea.

[8*]* *Descriptive Modeling* – Based in Part on chapter 9 of Hand , Manilla , & Smyth and Section 14.3 of HTF.

# LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn chân thành tới Cô giáo -Thạc sỹ **Nguyễn Thị Xuân Hương** - người đã tận tình chỉ bảo và giành thời gian quý báu định hướng và hướng dẫn em hoàn thành tốt đề tài tốt nghiệp này.

Em xin chân thành cảm ơn các thầy cô giáo trong khoa Công nghệ Thông tin trường Đại Học Dân Lập Hải Phòng, những người đã giảng dạy, trang bị cho em những kiến thức quý báu trong suốt thời gian học tập tại trường.

Em xin trân trọng cảm ơn GS. TS. **Trần Hữu Nghị** - Hiệu trưởng trường Đại Học Dân Lập Hải Phòng - người đã ủng hộ, tạo điều kiện tốt nhất để chúng em có thể hoàn thành nhiệm vụ học tập trong 2 năm vừa qua.

Cuối cùng tôi xin bày tỏ lòng biết ơn tới những người thân trong gia đình và những người bạn đã luôn chia sẻ, động viên tôi trong suốt quá trình học tập cho đến nay.

*Hải Phòng, tháng 7 năm2010*

**Sinh viên thực hiện**

***Cao Mai Liên***

# MỤC LỤC