BỘ GIÁO DỤC VÀ ĐÀO TẠO
TR- ỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o-------


ISO 9001 : 2008

# ĐỒ ÁN TỐT NGHIỆP

## NGÀNH CÔNG NGHỆ THÔNG TIN

**HẢI PHÒNG 2013**

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TR- ỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o-------

# PHÂN CỤM DỮ LIỆU
# BÀI TOÁN VÀ MỘT SỐ GIẢI THUẬT THEO TIẾP CẬN PHÂN HOẠCH

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

HẢI PHÒNG - 2013

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TR- ỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o-------

# PHÂN CỤM DỮ LIỆU
# BÀI TOÁN VÀ MỘT SỐ GIẢI THUẬT THEO TIẾP CẬN PHÂN HOẠCH

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

Giáo viên h- ớng dẫn: PGS.TS Nguyễn Thanh Tùng
Sinh viên thực hiện: Phạm Văn Đức
Mã số sinh viên: 121323

HẢI PHÒNG - 2013

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o------

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
*Độc lập - Tự do - Hạnh phúc*
-------o0o-------

# NHIỆM VỤ ĐỀ TÀI TỐT NGHIỆP

Sinh viên: Phạm Văn Đức — Mã sinh viên: 121323

Lớp: CT1201 — Ngành: Công nghệ thông tin

Tên đề tài:

PHÂN CỤM DỮ LIỆU: Bài toán và các giải thuật theo tiếp cận phân hoạch

# NHIỆM VỤ ĐỀ TÀI

1. Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp
   a. Nội dung:
      - Thế nào là khai phá dữ liệu khám phá tri thức từ cơ sở dữ liệu
      - Kỹ thuật phân cụm dữ liệu trong khai phá dữ liệu, phân loại các thuật toán phân cụm và các lĩnh vực ứng dụng.
      - Một số thuật toán phân cụm theo tiếp cận phân hoạch: Thuật toán K-Means, thuật toán K-Medoids
      - Xây dựng chương trình demo một trong số các thuật toán phân cụm phân hoạch trình bày.
   b. Các yêu cầu cần giải quyết:
      - Về lý thuyết: Nắm được các khái niệm, kỹ thuật về giải thuật theo tiếp cận phân hoạch
      - Về thực hành: Xây dựng được chương trình demo một trong số các thuật toán phân cụm phân hoạch trình bày.

2. Các số liệu cần thiết để thiết kế, tính toán

3. Địa điểm thực tập tốt nghiệp.

.

## CÁN BỘ H□ ỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

**Người hướng dẫn thứ nhất:**

Họ và tên: **Nguyễn Thanh Tùng**

Học hàm, học vị: *Phó giáo sư, Tiến sĩ.*

Cơ quan công tác: *Nguyên cán bộ nghiên cứu Viện Khoa học và Công nghệ Việt Nam.*

Nội dung hướng dẫn:

.....

.....

.....

.....

.....

.....

.....

.....

Đề tài tốt nghiệp đ- ợc giao ngày 25 tháng 03. năm 2013

Yêu cầu phải hoàn thành tr- ớc ngày 25 tháng 06 năm 2013

| | |
|---|---|
| Đã nhận nhiệm vụ: Đ.T.T.N<br>*Sinh viên* | Đã nhận nhiệm vụ: Đ.T.T.N<br>*Cán bộ h□ớng dẫn Đ.T.T.N* |
| Phạm Văn Đức | PGS.TS Nguyễn Thanh Tùng |

*Hải Phòng, ngày .....tháng.....năm 20…*

HIỆU TR- ỞNG

**GS.TS.NGƯT Trần Hữu Nghị**

## PHẦN NHẬN XÉT TÓM TẮT CỦA CÁN BỘ H□ ỚNG DẪN

1. Tinh thần thái độ của sinh viên trong quá trình làm đề tài tốt nghiệp:

.....

.....

.....

.....

.....

.....

2. Đánh giá chất l- ợng của đề tài tốt nghiệp (so với nội dung yêu cầu đã đề ra trong nhiệm vụ đề tài tốt nghiệp)

.....

.....

.....

.....

.....

3. Cho điểm của cán bộ h- ớng dẫn:

.....

.....

.....

.....

.

Ngày.....tháng.....năm 20…

Cán bộ h- ớng dẫn chính

*(Ký, ghi rõ họ tên )*

**PHẦN NHẬN XÉT ĐÁNH GIÁ CỦA CÁN BỘ CHẤM PHẢN BIỆN**

**ĐỀ TÀI TỐT NGHIỆP**

1. Đánh giá chất lượng đề tài tốt nghiệp về các mặt thu thập và phân tích số liệu ban đầu, cơ sở lý luận chọn phương án tối ưu, cách tính toán chất lượng thuyết minh và bản vẽ, giá trị lý luận và thực tiễn của đề tài.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2. Cho điểm của cán bộ phản biện

*( Điểm ghi bằng số và chữ )*

.....

.....

Ngày.....tháng.....năm 20…

Cán bộ chấm phản biện

*( Ký, ghi rõ họ tên )*

# MỤC LỤC

**DANH MỤC HÌNH MINH HỌA**

**LỜI CẢM ƠN**

Trước hết em xin chân thành cảm ơn thầy giáo **PGS.TS Nguyễn Thanh Tùng** là giáo viên hướng dẫn em trong quá tình làm đồ án. Thầy đã giúp đỡ em rất nhiều và đã cung cấp cho em nhiều tài liệu quan trọng phục vụ cho quá trình tìm hiểu về đề tài "Bài toán và một số giải thuật theo tiếp cận phân hoạch".

Thứ hai, em xin chân thành cảm ơn các thầy cô trong bộ môn công nghệ thông tin đã chỉ bảo em trong quá trình học và rèn luyện trong 4 năm học vừa qua. Đồng thời em cảm ơn các bạn sinh viên lớp CT1201 đã gắn bó với em trong quá trình rèn luyện tại trường.

Cuối cùng em xin chân thành cảm ơn ban giám hiệu trường Đại Học Dân Lập Hải Phòng đã tạo điều kiện cho em có kiến thức. Đồng thời các thầy cô trong trường giảng dạy cho em nhiều kinh nghiệm trong cuộc sống.

Em xin chân thành cảm ơn!

*Hải Phòng, ngày tháng năm*

Sinh viên

Phạm Văn Đức

# LỜI MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển vượt bậc của công nghệ điện tử và truyền thông, khả năng thu thập và lưu trữ thông tin của các hệ thống thg tin không ngừng được nâng cao. Theo đó, lượng thông tin được lưu trữ trên các thiết bị nhớ không ngừng tăng lên. Khai phá dữ liệu là một lĩnh vực khoa học mới xuất hiện, nhằm tự động hóa việc khai thác những thông tin, những tri thức 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. Những ứng dụng thành công trong khám phá tri thức, cho thấy khai phá dữ liệu 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 truyền thống. Hiện nay, khai phá dữ liệu đã và đang được 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 khai phá dữ liệu 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 trong cơ sở dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc 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 này và có khả năng làm việc với các CSDL lớn.

Nghiên cứu và ứng dụng một cách hiệu quả các phương pháp khai phá dữ liệu là vấn đề hấp dẫn, đã và đang thu hút sự quan tâm chẳng những của các nhà nghiên cứu, ứng dụng mà của cả các tổ chức, doanh nghiệp. Do đó, em đã chọn đề tài nghiên cứu " *Phân cum dữ liệu: Bài toán và một số giải thuật theo tiếp cận phân hoạch"* cho đồ án tốt nghiệp của mình.

Nội dung của đồ án gồm 3 chương:

Chương 1: Khái quát về khai phá dữ liệu: Trong chương này em trình bày tổng quan về khai phá dữ liệu, quy trình khai phá, các kỹ thuật khai phá và các ứng dụng của khai phá dữ liệu, cuối cùng là các thách thức đặt ra.

Chương 2: Trình bày về các phương pháp phân cụm dữ liệu, trong đó đồ án đi sâu vào tìm hiểu về phương pháp phân cụm phân cấp với 2 thuật toán điển hình là: K-Means, K-Medoids.

Chương 3: Cài đặt thực nghiệm: Để khẳng định cho khả năng và hiệu quả của thuật toán phân cụm dữ liệu phân hoạch.

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: KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU

## 1.1. Khai phá dữ liệu là gì

*Khai phá dữ liệu* là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn (các kho dữ liệu). Về bản chất, khai phá dữ liệu liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy (regularities) trong tập dữ liệu.

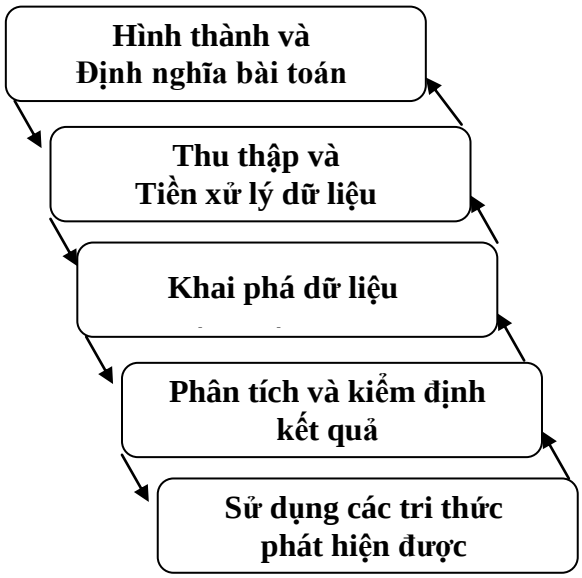
Năm 1989, Fayyad, Piatestsky-Shapiro và Smyth đã dùng khái niệm *Phát hiện tri thức trong cơ sở dữ liệu* (Kownledge Discovery in Database - KDD) để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, *khai phá dữ liệu* là một bước đặc biệt trong toàn bộ quá trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (pattern) (hay các mô hình) từ dữ liệu.

## 1.2. Quy trình khai phá dữ liệu

Quy trình phát hiện tri thức thường tuân theo các bước sau:

Bước thứ nhất: Hình thành và xác định bài toán. Bước này tìm hiểu lĩnh vực ứng dụng từ đó hình thành bài toán, xác định các nhiệm vụ cần phải hoàn thành. Điều này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu

*Bước thứ hai:* Thu thập và tiền xử lý dữ liệu: Tiến hành thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu (làm sạch dữ liệu), xử lý việc thiếu dữ liệu (làm giàu dữ liệu), biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức do dữ liệu được lấy từ nhiều nguồn khác nhau, không đồng nhất… có thể gây ra các nhầm lẫn. Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hoá.




***Hình 1.1.* Quy trình phát hiện tri thức**

*Bước thứ ba*: Khai phá dữ liệu, rút ra các tri thức: Trích ra các mẫu hoặc/ và các mô hình ẩn dưới các dữ liệu. Giai đoạn này rất quan trọng, bao gồm các công đoạn như: chức năng, nhiệm vụ và mục đích của khai phá dữ liệu, dùng phương pháp khai phá nào? Thông thường, các bài toán khai phá dữ liệu bao gồm: các bài toán mang tính mô tả - đưa ra tính chất chung nhất của dữ liệu, các bài toán dự báo bao

gồm cả việc phát hiện các suy diễn dựa trên dữ liệu hiện có. Tuỳ theo bài toán xác định được mà ta lựa chọn các phương pháp khai phá dữ liệu cho phù hợp.

*Bước thứ tư*: Sử dụng các tri thức phát hiện được, đặc biệt là làm sáng tỏ các mô tả và dự đoán.

Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện. Các kết quả của quá trình phát hiện tri thức có thể được đưa và ứng dụng trong các lĩnh vực khác nhau. Do các kết quả có thể là các dự đoán hoặc các mô tả nên chúng có thể được đưa vào các hệ thống hỗ trợ ra quyết định nhằm tự động hoá quá trình này.

***Tóm lại:*** KDD là một quá trình kết xuất ra tri thức từ kho dữ liệu mà trong đó khai phá dữ liệu là công đoạn quan trọng nhất.

## 1.3. Các kỹ thuật khai phá dữ liệu

Khai phá dữ liệu là lĩnh vực mà con người luôn tìm cách đạt được mục đích sử dụng thông tin của mình. Quá trình khai phá dữ liệu là quá trình phát hiện mẫu, trong đó phương pháp khai phá dữ liệu để tìm kiếm các mẫu đáng quan tâm theo dạng xác định. Có thể kể ra đây một vài phương pháp như: Sử dụng công cụ truy vấn, xây dựng cây quyết định, dựa theo khoảng cách (K-láng giềng gần), giá trị trung bình, phát hiện luật kết hợp,… Các phương pháp trên có thể được phỏng theo và được tích hợp vào các hệ thống lai để khai phá dữ liệu theo thống kê trong nhiều năm nghiên cứu. Tuy nhiên, với dữ liệu rất lớn trong kho dữ liệu thì các phương pháp này cũng đối diện với thách thức về mặt hiệu quả và quy mô.

### 1.3.1 Phương pháp suy diễn và quy nạp

Một cơ sở dữ liệu là một kho thông tin nhưng các thông tin quan trọng hơn cũng có thể được suy diễn từ kho thông tin đó. Có hai kỹ thuật chính để thực hiện việc này là suy diễn và quy nạp.

- ***Phương pháp suy diễn:***

Nhằm rút ra thông tin là kết quả logic của các thông tin trong cơ sở dữ liệu. Ví dụ như toán tử liên kết áp dụng cho bảng quan hệ, bảng đầu chứa thông tin về các nhân viên và phòng ban, bảng thứ hai chứa các thông tin về các phòng ban và các trưởng phòng. Như vậy sẽ suy ra được mối quan hệ giữa các nhân viên và các trưởng phòng. Phương pháp suy diễn dựa trên các sự kiện chính xác để suy ra các tri thức mới từ các thông tin cũ. Mẫu chiết xuất được bằng cách sử dụng phương pháp này thường là các luật suy diễn. Với tập dữ liệu khách hàng vay vốn ở trên, ta có mẫu chiết xuất được với ngưỡng thu nhập t là một luật như sau: "*Nếu thu nhập của khách hàng lớn hơn t đồng thì khách hàng có khả năng trả nợ*".

- ***Phương pháp quy nạp:***

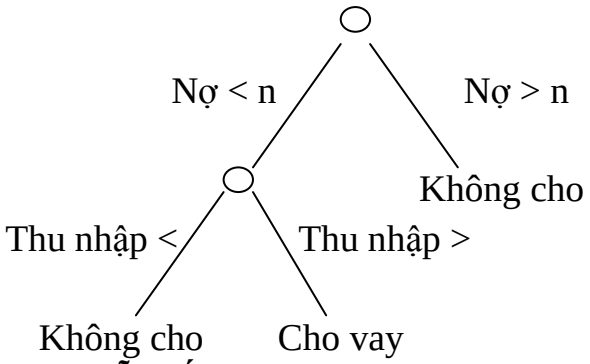
Phương pháp quy nạp suy ra các thông tin được sinh ra từ cơ sở dữ liệu. Có nghĩa là nó tự tìm kiếm, tạo mẫu và sinh ra tri thức chứ không phải bắt đầu với các tri thức đã biết trước. Các thông tin mà phương pháp này đem lại là các thông tin hay các tri thức cấp cao diễn tả về các đối tượng trong cơ sở dữ liệu. Phương pháp này liên quan đến việc tìm kiếm các mẫu trong CSDL.

Trong khai phá dữ liệu, quy nạp được sử dụng trong cây quyết định và tạo luật.

### *1.3.2. Cây quyết định và luật*

- ***Cây quyết định:***

Cây quyết định là một mô tả tri thức dạng đơn giản nhằm phân các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên các thuộc tính, các cạnh được gán các giá trị có thể của các thuộc tính, các lá mô tả các lớp khác nhau. Các đối tượng được phân lớp theo các đường đi trên cây, qua các cạnh tương ứng với các giá trị của thuộc tính của đối tượng tới lá. *Hình 1.3* mô tả một mẫu đầu ra có thể của quá trình khai phá dữ liệu dùng phương pháp cây quyết định với tập dữ liệu khách hàng xin vay vốn.




***Hình 1.2.* Mẫu kết quả với phương pháp cây quyết định**

- ***Tạo luật:***

Các luật được tạo ra nhằm suy diễn một số mẫu dữ liệu có ý nghĩa về mặt thống kê. Các luật có dạng *nếu P thì Q*, với P là mệnh đề đúng với một phần trong CSDL, Q là mệnh đề dự đoán. Ví dụ ta có một mẫu phát hiện được bằng phương pháp tạo luật: nếu giá 1 sản phẩm thấp hơn giá của một sản phẩm khác cùng loại 5000 đồng thì số lượng sản phẩm đó bán ra sẽ tăng 5% so với sản phẩm cùng loại. Những luật như thế này được sử dụng rất rộng rãi trong việc mô tả tri thức trong hệ chuyên gia. Chúng có thuận lợi là dễ hiểu đối với người sử dụng.

Cây quyết định và luật có ưu điểm là hình thức mô tả đơn giản, mô hình suy diễn khá dễ hiểu đối với người sử dụng. Tuy nhiên, giới hạn của nó là mô tả cây và luật chỉ có thể biểu diễn được một số dạng chức năng và vì vậy giới hạn về cả độ chính xác của mô hình

Đối với quy mô lớn, người ta dựa trên các phương pháp đánh giá mô hình theo xác suất với các mức độ mô hình phức tạp khác nhau. Các phương pháp tìm kiếm "tham lam", liên quan đến việc tăng và rút gọn các luật và các cấu trúc cây, chủ yếu được sử dụng để khai thác không gian siêu mũ (super-exponential space) của các mô hình. Cây và luật chủ yếu được sử dụng cho việc mô hình hóa dự đoán, phân lớp (Apte & Hong; Fayyad, Djorgovski, & Wei) và hồi quy. Chúng cũng có thể được áp dụng cho việc tóm tắt và mô hình hóa các mô tả (Agrawal et al.).

### *1.3.3. Phân nhóm và phân đoạn*

Kỹ thuật phân nhóm và phân đoạn là những kỹ thuật phân chia dữ liệu sao cho mỗi phần hoặc mỗi nhóm giống nhau theo một tiêu chuẩn nào đó. Mối quan hệ thành viên của các nhóm có thể dựa trên mức độ giống nhau của các thành viên và từ đó xây dựng nên các luật ràng buộc giữa các thành viên trong nhóm. Một kỹ thuật phân nhóm khác là xây dựng nên các hàm đánh giá các thuộc tính của các

thành phần như là hàm của các tham số của các thành phần. Phương pháp này được gọi là phương pháp phân hoạch tối ưu (optimal partitioning). Một ví dụ của phương pháp phân nhóm theo độ giống nhau là cơ sở dữ liệu khách hàng, ứng dụng của phương pháp tối ưu ví dụ như phân nhóm khách hàng theo số các tham số và các nhóm thuế tối ưu có được khi thiết lập biểu thuế bảo hiểm.

Mẫu đầu ra của quá trình khai phá dữ liệu sử dụng kỹ thuật này là các tập mẫu chứa các dữ liệu có chung những tính chất nào đó được phân tách từ cơ sở dữ liệu. Khi các mẫu được thiết lập, chúng có thể được sử dụng để tái tạo các tập dữ liệu ở dạng dễ hiểu hơn, đồng thời cũng cung cấp các nhóm dữ liệu cho các hoạt động cũng như công việc phân tích. Đối với cơ sở dữ liệu lớn, việc lấy ra các nhóm này là rất quan trọng.

### *1.3.4. Phương pháp ứng dụng K-láng giềng gần*

Sự miêu tả các bản ghi trong tập dữ liệu khi trỏ vào không gian nhiều chiều là rất có ích đối với việc phân tích dữ liệu. Việc dùng các miêu tả này, nội dung của vùng lân cận được xác định, trong đó các bản ghi gần nhau trong không gian được xem xét thuộc về lân cận (hàng xóm - láng giềng) của nhau. Khái niệm này được dùng trong khoa học kỹ thuật với tên gọi K-láng giềng gần, trong đó K là số láng giềng được sử dụng. Phương pháp này rất hiệu quả nhưng lại đơn giản. Ý tưởng thuật toán học K-láng giềng gần là "thực hiện như các láng giềng gần của bạn đã làm".

Kỹ thuật K-láng giềng gần là một phương pháp tìm kiếm đơn giản. Tuy nhiên, nó có một số mặt hạn chế giới hạn là phạm vi ứng dụng của nó. Đó là thuật toán này có độ phức tạp tính toán là luỹ thừa bậc 2 theo số bản ghi của tập dữ liệu.

Vấn đề chính liên quan đến thuộc tính của bản ghi. Một bản ghi gồm nhiều thuộc tính độc lập, nó bằng một điểm trong không gian tìm kiếm có số chiều lớn. Trong các không gian có số chiều lớn, giữa hai điểm bất kỳ hầu như có cùng khoảng cách. Vì thế mà kỹ thuật K-láng giếng không cho ta thêm một thông tin có ích nào, khi hầu hết các cặp điểm đều là các láng giềng. Cuối cùng, phương pháp K-láng giềng không đưa ra lý thuyết để hiểu cấu trúc dữ liệu. Hạn chế đó có thể được khắc phục bằng kỹ thuật cây quyết định.

### *1.3.5. Các phương pháp dựa trên mẫu*

Sử dụng các mẫu mô tả từ cơ sở dữ liệu để tạo nên một mô hình dự đoán các mẫu mới bằng cách rút ra những thuộc tính tương tự như các mẫu đã biết trong mô hình. Các kỹ thuật bao gồm phân lớp theo láng giềng gần nhất, các giải thuật hồi quy (Dasarathy 1991) và các hệ thống suy diễn dựa trên tình huống (case-based reasoning) (Kolodner 1993).

Khuyết điểm của các kỹ thuật này là cần phải xác định được khoảng cách, độ đo giống nhau giữa các mẫu. Mô hình thường được đánh giá bằng phương pháp đánh giá chéo trên các lỗi dự đoán (Weiss & Kulikowski, 1991). "Tham số" của mô hình được đánh giá có thể bao gồm một số láng giềng dùng để dự đoán và số đo khoảng cách. Giống như phương pháp hồi quy phi tuyến, các phương pháp này khá mạnh trong việc đánh giá xấp xỉ các thuộc tính, nhưng lại rất khó hiểu vì mô hình không được định dạng rõ ràng mà tiềm ẩn trong dữ liệu.

### *1.3.6. Phát hiện các luật kết hợp*

Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần dữ liệu trong cơ sở dữ liệu. Mẫu đầu ra của giải thuật khai phá dữ liệu là tập luật kết hợp tìm được. Ta có thể lấy một ví dụ đơn giản về luật kết hợp như sau: sự kết hợp giữa hai thành phần A và B có nghĩa là sự xuất hiện của A trong bản ghi kéo theo sự xuất hiện của B trong cùng bản ghi đó: $A \Rightarrow B$.

Cho một lược đồ $R=\{A_1,\ldots, A_p\}$ các thuộc tính với miền giá trị {0,1}, và một quan hệ r trên R. Một luật jết hợp trên r được mô tả dưới dạng $X \Rightarrow B$ với $X \subseteq R$ và $B \in R \backslash X$. Về mặt trực giác, ta có thể phát biểu ý nghĩa của luật như sau: nếu một bản ghi của bảng r có giá trị 1 tại mỗi thuộc tính thuộc X thì giá trị của thuộc tính B cũng là 1 trong cùng bản ghi đó. Ví dụ như ta có tập cơ sở dữ liệu về các mặt hàng bán trong siêu thị, các dòng tương ứng với các ngày bán hàng, các cột tương ứng với các mặt hàng thì giá trị 1 tại ô (20/10, bánh mì) xác định rằng bánh mì đã bán ngày hôm đó cũng kéo theo sự xuất hiện giá trị 1 tại ô (20/10, bơ).

Cho $W \subseteq R$, đặt s(W,r) là tần số xuất hiện của W trong r được tính bằng tỷ lệ của các hàng trong r có giá trị 1 tại mỗi cột thuộc W. Tần số xuất hiện của luật $X \Rightarrow B$ trong r được định nghĩa là $s(X \cup \{B\}, r)$ còn gọi là độ hỗ trợ của luật, độ tin cậy của luật là $s(X \cup \{B\}, r)/s(X, r)$. Ở đây X có thể gồm nhiều thuộc tính, B là giá trị không cố định. Nhờ vậy mà không xảy ra việc tạo ra các luật không mong muốn trước khi quá trìm tìm kiếm bắt đầu. Điều đó cũng cho thấy không gian tìm kiếm có kích thước tăng theo hàm mũ của số lượng các thuộc tính ở đầu vào. Do vậy cần phải chú ý khi thiết kế dữ liệu cho việc tìm kiếm các luật kết hợp.

Nhiệm vụ của việc phát hiện các luật kết hợp là phải tìm tất cả các luật $X \Rightarrow B$ sao cho tần số của luật không nhỏ hơn ngưỡng $\sigma$ cho trước và độ tin cậy của luật không nhỏ hơn ngưỡng $\theta$ cho trước. Từ một cơ sở dữ liệu ta có thể tìm được hàng nghìn và thậm chí hàng trăm nghìn các luật kết hợp.

Ta gọi một tập con $X \subseteq R$ là thường xuyên trong r nếu thỏa mãn điều kiện $s(X, r) \geq \sigma$. Nếu biết tất cả các tập thường xuyên trong r thì việc tìm kiếm các luật rất dễ dàng. Vì vậy, giải thuật tìm kiếm các luật kết hợp trước tiên đi tìm tất cả các tập thường xuyên này, sau đó tạo dựng dần các luật kết hợp bằng cách ghép dần các tập thuộc tính dựa trên mức độ thường xuyên.

Các luật kết hợp có thể là một cách hình thức hóa đơn giản. Chúng rất thích hợp cho việc tạo ra các kết quả có dữ liệu dạng nhị phân. Giới hạn cơ bản của phương pháp này là ở chỗ các quan hệ cần phải thưa theo nghĩa không có tập thường xuyên nào chứa nhiều hơn 15 thuộc tính. Giải thuật tìm kiếm các luật kết hợp tạo ra số luật ít nhất phải bằng với số các tập thường xuyên và nếu như một tập thường xuyên có kích thước K thì phải có ít nhất là $2^K$ tập thường xuyên. Thông tin về các tập thường xuyên được sử dụng để ước lượng độ tin cậy của các tập luật kết hợp.

## 1.4. Các ứng dụng của khai phá dữ liệu

Mặc dù còn rất nhiều vấn đề mà khai phá dữ liệu cần phải tiếp tục nghiên cứu để giải quyết nhưng tiềm năng của nó đã được khẳng định bằng sự ra đời của rất nhiều ứng dụng trong nhiều lĩnh vực của đời sống :

- *Trong một số lĩnh vực khoa học:* Quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ genen và một số bệnh di truyền…
- *Dự báo thời tiết*: Mô hình hóa những thay đổi thời tiết, phân tích những mẫu như mưa bão, lốc xoáy, sóng thần.. Để đưa ra nhưng dự đoán chính xác, kịp thời.
- *Bảo hiểm, tài chính và thị trường chứng khoán:* Phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thẻ tín dụng, phát hiện gian lận…
- *Điều trị y học và chăm sóc y tế:* Một số thông tin về chuẩn đoán bệnh lưu trong các hệ thống quản lý bệnh viện. Phân tích mối liên hệ giữa các triệu chứng bệnh, chuẩn đoán và phương pháp điều trị.

## 1.5. Một số thách thức đặt ra cho việc khai phá dữ liệu

Khai phá dữ liệu là kỹ thuật mới chứa nhiều tiềm năng mà người ta vẫn chưa khai phá hết. Và việc nghiên cứu và ứng dụng kỹ thuật khai phá dữ liệu luôn gặp nhiều khó khăn, nhưng đứng trước những khó khăn đó chúng ta cần tìm ra những hướng giải quyết để hoàn thiện hơn các kỹ thuật khai phá dữ liệu. Ta có thể liệt kê một số khó khăn như sau:

- *Dữ liệu lớn:*

Cho đến nay, các cơ sở dữ liệu với hàng trăm trường và bảng, hàng triệu bản ghi và với kích thước đến gigabytes đã là chuyện bình thường. Hiện nay đã bắt đầu xuất hiện các cơ sở dữ liệu có kích thước tới terabytes. Các phương pháp giải quyết hiện nay là đưa ra một ngưỡng cho cơ sở dữ liệu, lấu mẫu, các phương pháp xấp xỉ, xử lý song song (Agrawal et al, Holsheimer et al).

- *Kích thước lớn:*

Không chỉ có số lượng bản ghi lớn mà số các trường trong cơ sở dữ liệu cũng nhiều. Vì vậy mà kích thước của bài toán trở nên lớn hơn. Một tập dữ liệu có kích thước lớn sinh ra vấn đề làm tăng không gian tìm kiếm mô hình suy diễn. Hơn nữa, nó cũng làm tăng khả năng một giải thuật khai phá dữ liệu có thể tìm thấy các mẫu giả. Biện pháp khắc phục là làm giảm kích thước tác động của bài toán và sử dụng các tri thức biết trước để xác định các biến không phù hợp.

- *Dữ liệu động:*

Đặc điểm cơ bản của hầu hết các cơ sở dữ liệu là nội dung của chúng thay đổi liên tục. Dữ liệu có thể thay đổi theo thời gian và việc khai phá dữ liệu cũng bị ảnh hưởng bởi thời điểm quan sát dữ liệu. Ví dụ trong cơ sở dữ liệu về tình trạng bệnh nhân, một số giá trị dữ liệu là hằng số, một số khác lại thay đổi liên tục theo thời gian (ví dụ cân nặng và chiều cao), một số khác lại thay đổi tùy thuộc vào tình huống và chỉ có giá trị được quan sát mới nhất là đủ (ví dụ nhịp đập của mạch). Vậy thay đổi dữ liệu nhanh chóng có thể làm cho các mẫu khai thác được trước đó mất giá trị. Hơn nữa, các biến trong cơ sở dữ liệu của ứng dụng đã cho cũng có thể bị thay đổi, bị xóa hoặc là tăng lên theo thời gian. Vấn đề này được giải quyết bằng

các giải pháp tăng trưởng để nâng cấp các mẫu và coi những thay đổi như là cơ hội để khai thác bằng cách sử dụng nó để tìm kiếm các mẫu bị thay đổi.

- *Các trường không phù hợp:*

Một đặc điểm quan trọng khác là tính không thích hợp của dữ liệu, nghĩa là mục dữ liệu trở thành không thích hợp với trọng tâm hiện tại của việc khai thác. Một khía cạnh khác đôi khi cũng liên quan đến độ phù hợp là tính ứng dụng của một thuộc tính đối với một tập con của cơ sở dữ liệu.

- *Các giá trị bị thiếu:*

Sự có mặt hay vắng mặt của giá trị các thuộc tính dữ liệu phù hợp có thể ảnh hưởng đến việc khai phá dữ liệu. Trong hệ thống tương tác, sự thiếu vắng dữ liệu quan trọng có thể dẫn đến việc yêu cầu cho giá trị của nó hoặc kiểm tra để xác định giá trị của nó. Hoặc cũng có thể sự vắng mặt của dữ liệu được coi như một điều kiện, thuộc tính bị mất có thể được coi như một giá trị trung gian và là giá trị không biết.

- *Các trường bị thiếu:*

Một quan sát không đầy đủ cơ sở dữ liệu có thể làm cho các dữ liệu có giá trị bị xem như có lỗi. Việc quan sát cơ sở dữ liệu phải phát hiện được toàn bộ các thuộc tính có thể dùng để giải thuật khai phá dữ liệu có thể áp dụng nhằm giải quyết bài toán. Giả sử ta có các thuộc tính để phân biệt các tình huống đáng quan tâm. Nếu chúng không làm được điều đó thì có nghĩa là đã có lỗi trong dữ liệu. Đối với một hệ thống học để chuẩn đoán bệnh sốt rét từ một cơ sở dữ liệu bệnh nhân thì trường hợp các bản ghi của bệnh nhân có triệu chứng giống nhau nhưng lại có các chuẩn đoán khác nhau là do trong dữ liệu đã bị lỗi. Đây cũng là vấn đề thường xảy ra trong cơ sở dữ liệu kinh doanh. Các thuộc tính quan trọng có thể sẽ bị thiếu nếu dữ liệu không được chuẩn bị cho việc khai phá dữ liệu.

- *Độ nhiễu và không chắc chắn:*

Đối với các thuộc tính đã thích hợp, độ nghiêm trọng của lỗi phụ thuộc vào kiểu dữ liệu của các giá trị cho phép. Các giá trị của các thuộc tính khác nhau có thể là các số thực, số nguyên, chuỗi và có thể thuộc vào tập các giá trị định danh. Các giá trị định danh này có thể sắp xếp theo thứ tự từng phần hoặc đầy đủ, thậm chí có thể có cấu trúc ngữ nghĩa.

Một yếu tố khác của độ không chắc chắn chính là tính kế thừa hoặc độ chính xác mà dữ liệu cần có, nói cách khác là độ nhiễu của dữ liệu. Dựa trên việc tính toán trên các phép đo và phân tích có ưu tiên, mô hình thống kê mô tả tính ngẫu nhiên được tạo ra và được sử dụng để định nghĩa độ mong muốn và độ dung sai của dữ liệu. Thường thì các mô hình thống kê được áp dụng theo cách đặc biệt để xác định một cách chủ quan các thuộc tính để đạt được các thống kê và đánh giá khả năng chấp nhận của các (hay tổ hợp các) giá trị thuộc tính. Đặc biệt là với dữ liệu kiểu số, sự đúng đắn của dữ liệu có thể là một yếu tố trong việc khai phá.

- *Mối quan hệ phức tạp giữa các trường:*

Các thuộc tính hoặc các giá trị có cấu trúc phân cấp, các mối quan hệ giữa các thuộc tính và các phương tiện phức tạp để diễn tả tri thức về nội dung của cơ sở dữ liệu yêu cầu các giải thuật phải có khả năng sử dụng một cách hiệu quả các thông tin này. Ban đầu, kỹ thuật khai phá dữ liệu chỉ được phát triển cho các bản ghi có

giá trị thuộc tính đơn giản. Tuy nhiên, ngày nay người ta đang tìm cách phát triển các kỹ thuật nhằm rút ra mối quan hệ giữa các biến này.

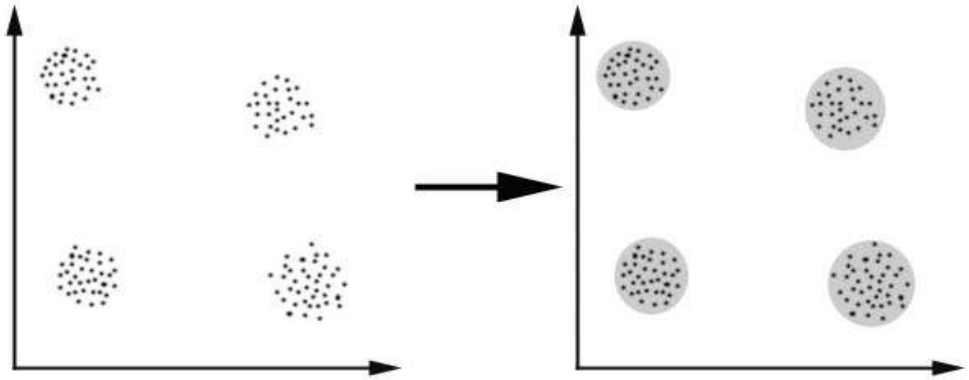
## 1.6. Kết luận chương 1

Khai phá dữ liệu là lĩnh vực đã và đang trở thành một trong những hướng nghiên cứu thu hút được sự quan tâm của nhiều chuyên gia về CNTT trên thế giới và được ứng dụng trong nhiều lĩnh vực khác nhau. Tại Việt Nam kỹ thuật này còn tương đối mới mẻ tuy nhiên cũng đang được nghiên cứu và dần đưa vào ứng dụng. Trong những năm gần đây, rất nhiều các phương pháp và thuật toán mới liên tục được công bố. Điều này chứng tỏ những ưu thế, lợi ích và khả năng ứng dụng thực tế to lớn của khai phá dữ liệu. Trong chương nàyđã trình bày mộtcách tổng quan về khai phá tri thức và khai phá dữ liệu.

# Chương 2. PHÂN CỤM DỮ LIỆU VÀ CÁC GIẢI THUẬT THEO TIẾP CẬN PHÂN HOẠCH

## 2.1. Phân cụm dữ liệu là gì?

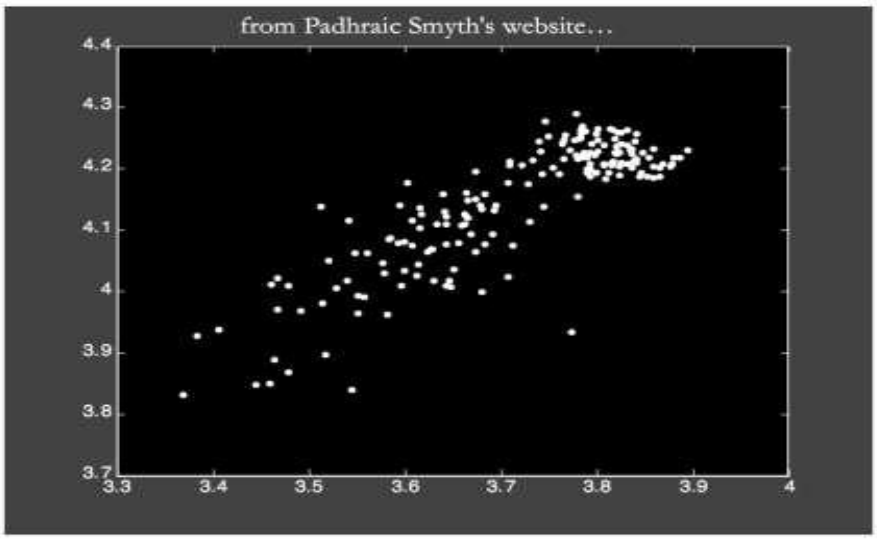
Phân cụm dữ liệu - PCDL (*Data Clustering*) là hình thức học không giám sát (unsupervised learning)trong đó các mẫu học chưa được gán nhãn.Mục đích của PCDL là tìm những mẫu đại diện hoặc nhóm dữ liệu tương tự nhau thành những cụm. Các điểm dữ liệu trong các cụm khác nhau có độ tương tự thấp hơn các điểm nằm trong cùng một cụm.



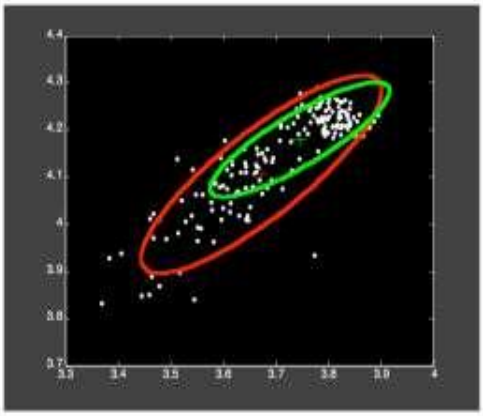
**Hình 2.1: Mô phỏng vấn đề PCDL**

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

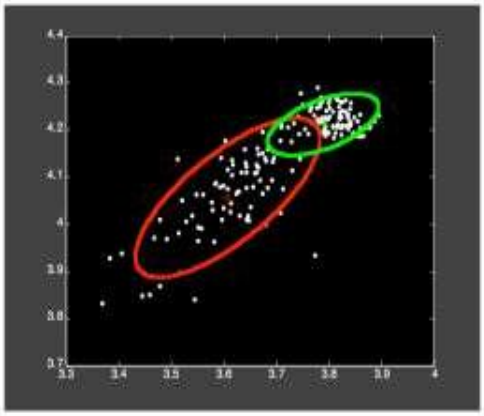
Để minh hoạ cụ thể hơn cho vấn đề này ta có thể quan sát các hình ảnh sau:



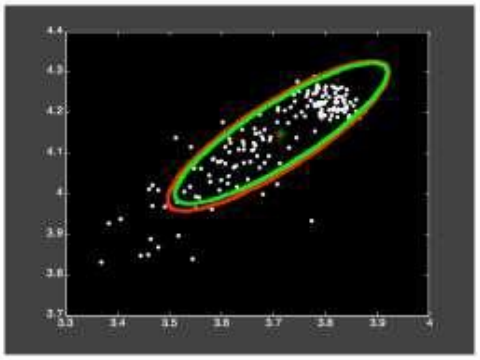

**Hình 2.2: Dữ liệu nguyên thuỷ**



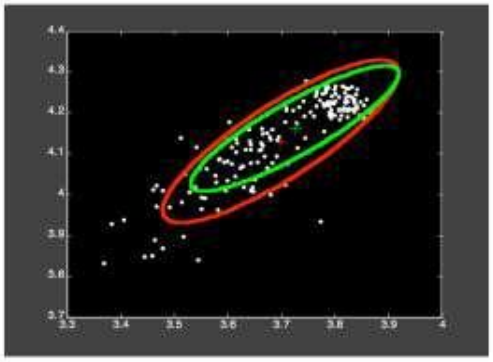
**Hình 2.3**



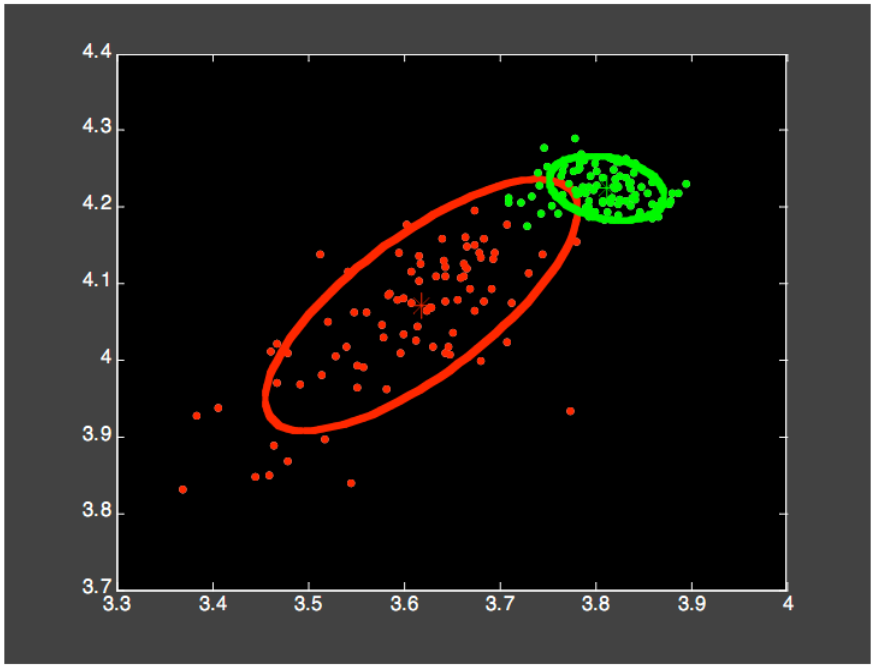
**Hình 2.4**



**Hình 2.5**



**Hình 2.6**




**Hình 2.7: Kết quả của quá trình phân cụm**

Các hình 2.2, 2.3, 2.4, 2.5, 2.6 ,2.7 là thể hiện quá trình phân cụm từ khi"bắt

đầu" cho đến khi "kết thúc" .

## 2.2. Các ứng dụng của phân cụm

Phân cụm dữ liệu có rất nhiều ứng dụng trong các lĩnh vực khác nhau:

- *Thương mại*: Giúp các doanh nhân khám phá ra các nhóm khách hàng quan trọng để đưa ra các mục tiêu tiếp thị.
- *Sinh học*: Xác định các loài sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.
- *Lập quy hoạch đô thị*: Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý, nhằm cung cấp thông tin cho quy hoặch đô thị.
- *Thư viện***:** Phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả
- *Bảo hiểm*: Nhận dạng nhóm tham gia bảo hiểm có chi phí bồi thường cao, nhận dạng gian lận thương mại.
- *Nghiên cứu trái đất*: Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhân dạng các vùng nguy hiểm.
- *World Wide Web*: 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 khai phá dữ liệu từ dữ liệu.

## 2.3. Các yêu cầu đối với thuật toán phân cụm dữ liệu

Theo các nghiên cứu cho thấy hiện này chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc CSDL. Hơn nữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của các CSDL, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng thuật toán phân cụm phù hợp. Vì vậy, phân cụm dữ liệu vẫn đang làmột vấn đề khó và mở vì phải giải quyết nhiều vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt là với kho dữ liệu hỗn hợp đang ngày càng tăng và đây cũng là một trong những thách thức lớn trong lĩnh vực KPDL.Vậy phân cụm dữ liệu là một thách thức trong lĩnh vực nghiên cứu vì những ứng dụng tiềm năng của chúng được đưa ra ngay chính trong những yêu cầu đặc biệt của chúng. Do đặc thù của của cơ sở dữ liệu là lớn, phức tạp, và có dữ liệu nhiễu nên những thuật toán phân cụm được áp dụng phải thoả mãn những yêu cầu sau:

- Thuật toán phải hiệu quả và thời gian chạy phải là tăng tuyến tính theo kích thước của dữ liệu.
- Thuật toán phải xử lý và áp dụng được với cơ sở dữ liệu nhiều nhiễu, phức tạp gồm cả dữ liệu không gian, phi không gian, dữ liệu số, phi số, kiểu nhị phân, dữ liệu định danh, hạng mục, thích nghi với kiểu dữ liệu hỗn hợp.
- Thuật toán phải có khả năng xác định được những cụm với hình dáng bất kỳ bao gồm cả những cụm có hình dạng lồng nhau, cụm có hình dạng lõm, hình cầu, hình que.
- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào.Do các giá trị đầu vào thường ảnh hưởng rất lớn đến thuật toán phân cụm và rất phức tạp để xác định các giá trị vào thích hợp đối với các CSDL lớn.
- Thuật toán phải thực hiện với mọi thứ tự đầu vào dữ liệu. Nói cách khác kết quả của thuật toán nên độc lập với dữ liệu đầu vào (Cùng một tập dữ liệu, khi đưa

vào xử lý cho thuật toán PCDL với các thứ tự vào của các đối tượng dữ liệu ở các lần thực hiện khác nhau thì không ảnh hưởng lớn đến kết quả phân cụm)

- Thuật toán không đòi hỏi những tri thức về cơ sở dữ liệu từ người dùng.
- Thuật toán phải làm việc được với cơ sở dữ liệu chứa nhiều lớp đối tượng dữ liệu phức tạp và có tính chất khác nhau.
- Thuật toán phải thích nghi với dữ liệu đa chiều: Thuật toán có khả năng áp dụng hiệu quả cho dữ liệu có số khác chiều nhau.
- Thuật toán phải dễhiểu,dễ cài đặt và khả thi: Người sử dụng có thể chờ đợi những kết quả phân cụm dễ hiểu, dễ lý giải và dễ sử dụng. Nghĩa là, sự phân cụm có thể cần được giải thích ý nghĩa và ứng dụng rõ ràng. Việc nghiên cứu cách để một ứng dụng đạt được mục tiêu rất quan trọng có thể gây ảnh hưởng tới sự lựa trọn các phương pháp phân cụm.

## 2.4. Các kiểu dữ liệu trong phân cụm

Trong phân cụm, các đối tượng dữ liệu thường được diễn tả dưới dạng các đặc tính hay còn gọi là thuộc tính ( Khái niệm "các kiểu dữ liệu" và "các kiểu thuộc tính dữ liệu" được xem là tương đương với nhau). Các thuộc tính này là các tham số để giải quyết vấn đề phân cụm và sự lựa chọn chúng có tác động đáng kể đến kết quả phân cụm. Phân loại các kiểu thuộc tính khác nhau là 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. Các thuật toán phân cụm thường sử dụng một trong hai cấu trúc dữ liệu sau:

***Ma trận dữ liệu (Data matrix, object-by-variable structure)***: là mảng n hàng, p cột, trong đó p là số thuộc tính của mỗi đối tượng. Mỗi hàng biểu diễn một đối tượng, các phần tử trong mỗi hàng chỉ giá trị thuộc tính tương ứng của đối tượng đó. Mảng được cho như sau:

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{21} & \dots & x_{2f} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

***Ma trận phi tương tự (Dissimilarity matrix, object-by-object structure):*** là mảng n hàng, n cột. Phần tử d(i,j) chứa khoảng cách hay độ khác biệt giữa các đối tượng i và đối tượng j, d(i,j) là một số không âm, trong đó nếu d(i,j) xấp xỉ 0 thì hai đối tượng i và j là khá "gần" nhau, nếu d(i,j) càng lớn thì hai đối tượng i, j khá khác

nhau. Do d(i,j) = d(j,i) = 0 nên ta có thể biểu diễn ma trận phi tương tự như sau:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & ... & ... & ... \\ d(3.1) & d(3,2) & 0 & ... & ... \\ ... & ... & ... & ... & ... \\ d(n,1) & d(n,2) & ... & ... & 0 \end{bmatrix}$$

Với d(i,j) là khoảng cách giữa đối tượng i và đối tượng j.

Phần lớn các thuật toán phân cụm sử dụng cấu trúc ma trận phi tương tự. Do vậy, nếu dữ liệu cần phân cụm được tổ chức dưới dạng ma trận dữ liệu thì cần biến đổi về dạng ma trận phi tương tự trước khi tiến hành phân cụm.

Có hai đặc trưng để phân loại: kích thước miền và hệ đo.

Cho một CSDL D chứa n đối tượng trong không gian k chiều; 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 $i = 1,.., k$ là các đặc trưng hoặc thuộc tính tương ứng của các đối tượng *x, y, z;* như vậy sẽ có các kiểu dữ liệu sau:

### 2.4.1. *Kiểu dữ liệu dựa trên kích thước miền*

- *Thuộc tính liên tục*: 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 (ví dụ, các thuộc tính mầu, nhiệt độ hoặc cường độ âm thanh,…)
- *Thuộc tính rời rạc*: Nếu miền giá trị của nó là tập hữu hạn, đếm được (ví dụ: các thuộc tính số,…) trường hợp đặc biệt của thuộc tính rời rạc là thuộc tính nhị phân mà miền giá trị chỉ có hai phân tử (ví dụ: Yes/No, True/False, On/Off..)

### *2.4.2. Kiểu dữ liệu dựa trên hệ đo*

- *Thuộc tính định danh*: Là dạng thuộc tính khái quát hoá của thuộc tính nhị phân, trong đó có miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử. 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: Là thuộc tính định danh nhưng có thêm tính thứ tự nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì có thể xác định là $x \neq y$ hoặc x = y hoặc x > y hoặc x< y.
- *Thuộc tính khoảng*: để đo các giá trị theo xấp xỉ tuyến tính, với thuộc tính khoảng có thể xác định một thuộc tính là đứng trược hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì có thể nói x cách y một khoảng $x_i - y_i$ tương ứng với thuộc tính thứ i.

Việc lựa chọn đơn vị đo cho các thuộc tính cũng ảnh hưởng đến chất lượng phân cụm. Nếu đơn vị độ đo của một thuộc tính càng được chia nhỏ, thì khoảng cách xác định của thuộc tính đó càng lớn và ảnh hưởng nhiều hơn đến kết quả phân cụm. Để tránh phụ thuộc vào việc lựa chọn đơn vị đo, dữ liệu cần được chuẩn hóa. Việc chuẩn hóa sẽ gán cho tất cả các thuộc tính một trọng số bằng nhau. Tuy nhiên, trong nhiều trường hợp người sử dụng có thể thay đổi trọng số cho các thuộc tính ưu tiên.

Để chuẩn hóa các độ đo, một cách làm phổ biến là biến đổi các thuộc tính về dạng không có đơn vị đo. Giả sử đối với các thuộc tính f, ta thực hiện như sau:

- Tính độ lệch trung bình:

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

Trong đó $x_{1f},\ldots,x_{nf}$ là giá trị thuộc tính f của n phần tử dữ liệu, và $m_f$ là giá trị trung bình của f, được cho như sau:

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$$

- Độ đo được chuẩn hóa:

- *Thuộc tính nhị phân* là thuộc tính có hai giá trị là 0 và 1.
- *Thuộc tính tính tỷ lệ*: Là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc.

Trong các thuộc tính trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục, còn thuộc tính khoảng cách và thuộc tính tỷ lệ được gọi là thuộc tính số.

Đặc biệt, còn có dữ liệu không gian là loại dữ liệu có 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 (ví dụ: thông tin về hình học, Quan hệ metric, Quan hệ hướng, …) Dữ liệu không gian có thể là dữ liệu liên tục hoặc rời rạc.

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

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

## 2.5. Phép đo độ tương tự và khoảng cách đối với các kiểu dữ liệu:

### *2.5.1. Khái niệm tương tự, phi tương tự*

Khi các đặc tính của dữ liệu được xác định, phải 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 hoặc là để tính độ phi tương tự giữa các đối tượng dữ liệu. Giá trị của hàm tính độ đo tương tự càng lớn thì sự giống nhau giữa các 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 phi tương tự có nhiều cách để xác định, chúng thường được đo bằng khoảng cách giữa các đối tượng. Tất cả các cách đo độ tương tự đều phụ thuộc vào kiểu thuộc tính mà con người phân tích. Ví dụ, thuộc tính hạng mục thì 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 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 "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 đề cập ở trên được gọi là một không gian metric nếu:

- Với mỗi cặp phần tử x, y thuộc X đều xác định theo một quy tắc nào đó,

một số thực d(x,y) được gọi là khoảng cách giữa x và y.

- Quy tắc nói trên thỏa mãn hệ tính chất sau:

(i) $d(x,y) > 0$ nếu $x \neq y$;
(ii) $d(x,y) = 0$ nếu $x = y$ ;
(iii) $d(x,y) = d(y,x)$ với mọi x,y ;
(iv) $d(x,y) \leq d(x,z) + d(z,y)$ ;

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.

### *2.5.2. Thuộc tính khoảng*

Một thành phần quan trọng trong thuật toán phân cụm là phép đo khoảng cách giữa hai điểm dữ liệu. Nếu thành phần của vectơ thể hiện dữ liệu thuộc trong cùng một đơn vị giống nhau thì nó tồn tại khoảng cách Euclidean có thể xác định được nhóm dữ liệu tương tự. Tuy nhiên, không phải lúc nào khoảng cách Euclidean cũng cho kết quả chính xác.

Tuy nhiên chú ý rằng đây không phải vấn đề đồ thị: vấn đề phát sinh từ công thức toán học được sử dụng để kết hợp khoảng cách giữa các thành phần đơn đặc tính dữ liệu vectơ vào trong một độ đo khoảng duy nhất mà có thể được sử dụng cho mục đích phân cụm: các công thức khác nhau dẫn tới những cụm khác nhau.

Các thuật toán cần có các phép đo khoảng cách hoặc độ tương tự giữa hai đối tượng để thực hiện phân cụm. Kiến thức miền phải được sử dụng để để trình bày rõ ràng phép đo khoảng thích hợp cho mỗi ứng dụng. Hiện nay, phép đo có nhiều mức độ khách nhau tùy theo từng trường hợp.

- *Khoảng cách Minkowski:*

$$d(i,j)= \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q} \quad (q>0)$$

- *Khoảng cách Euclidean*: là khoảng cách Minkowski khi q=2. Khoảng cách Euclidean chính là khoảng cách hình học trong không gian n chiều

$$d(i,j)=\sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

- *Khoảng cách Manhattan:* là khoảng cách *Minkowski* khi q=1.

$$d(i,j)=\sqrt{|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|}$$

- *Khoảng cách có trọng:*

$$d(i,j)= \sqrt[q]{w_1|x_{i1} - x_{j1}|^q + w_2|x_{i2} - x_{j2}|^q + \cdots + w_p|x_{ip} - x_{jp}|^q} \quad (q>0)$$

Khoảng cách có trọng là sự cải tiến của khoảng cách *Minkowski*, trong đó có tính ảnh hưởng của từng thuộc tính đến khoảng cách giữa hai đối tượng. Thuộc tính có trọng số w càng lớn thì ảnh hưởng càng nhiều đến khoảng cách d. Việc chọn trọng số tùy thuộc vào ứng dụng và mục tiêu cụ thể.

### 2.5.3. *Thuộc tính nhị phân*

Tất cả các phép đo được định nghĩa ở trên là đa số thích hợp cho các biến liên tục. Cho các biến danh nghĩa, "phép đo khoảng cách" là 0 nếu các trường hợp có cùng giá trị danh nghĩa, và 1 nếu các trường hợp có các giá trị danh nghĩa khác nhau, hoặc với độ đo tương tự 1 (nếu các trường hợp có cùng giá trị danh nghĩa) và 0 (nếu không giống nhau).

Do đó nếu xem xét p biến định danh, có thể đánh giá độ tương tự của các trường hợp bằng số các biến mà có giá trị giống nhau. Nói chung định nghĩa với một biến nhị phân mới từ mỗi biến danh nghĩa, bằng việc nhóm các nhãn danh nghĩa thành

hai lớp, một nhãn là 1, nhãn khác là 0. Xây dựng và xem xét bảng ngẫu nhiên các sự kiện có thể xảy ra và định nghĩa các thuộc tính của đối tượng x, y bằng các biến số nhị phân 0 và 1.

| | | Y | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| X | 1 | a | b | a+b |
| | 0 | c | d | c+d |
| | | a+c | b+d | p=a+b+c+d |

**Hình 2.8: Bảng tham số**

Trong đó:
a là tổng số các thuộc tính có giá trị 1 trong hai đối tượng x, y
b là tổng số các thuộc tính có giá trị 1 trong x và giá trị 0 trong y
c là tổng số các thuộc tính có giá trị 0 trong x và giá trị 1 trong y
d là tổng số các thuộc tính có giá trị 0 trong hai đối tượng x, y
p là tổng tất cả các thuộc tính của hai đối tượng x, y
Ta có tổng số các thuộc tính về đối tượng p = a + b + c + d.

Các phép đo độ tương tự giữa hai đối tượng trong trường hợp dữ liệu thuộc tính nhị phân được định nghĩa như sau:

- *Hệ số đối sánh đơn giản*:

$$d(x,y)=\frac{a+d}{p}$$

cả hai đối tượng có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

- *Hệ số Jaccard:*

$$d(x,y)=\frac{a}{a+b+c}$$

tham số này bỏ qua số các đối sánh 0-0

Công thức này sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu cao hơn nhiều so với các thuộc tính có giá trị 0. Như vậy thuộc tính nhị phân ở đây là không đối xứng.

### *2.5.4. 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, *p* là tổng số các thuộc tính.

### *2.5.5. 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 ($M_i$ là 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 ri, với $r_i \in \{1 \dots M_i\}$. $M_i \in$
- 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 :

$$Z_i^{(j)} = \frac{r_i^{(j)} - 1}{Mi - 1}$$

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

### *2.5.6. Thuộc tính tỉ lệ (Ratio Scale)*

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. 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:

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

Có thể chuyển đổi giữa các mô hình cho các kiểu dữ liệu trên, ví dụ như dữ liệu kiểu hạng mục có thể chuyển đổi thành dữ liệu nhị phân hoặc ngược lại. Giải pháp này rất tốn kém về chi phí tính toán, do vậy, cần phải cân nhắc khi áp dụng cách thức này.

Tóm lại, tùy từng trường hợp dữ liệu cụ thể mà có thể 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.

## 2.6. Các hướng tiếp cận bài toán phân cụm dữ liệu

Có rất nhiều các phương pháp phân cụm dữ liệu khác nhau. Việc lựa chọn phương pháp nào tuỳ thuộc vào kiểu dữ liệu, mục tiêu và ứng dụng cụ thể. Nhìn chung, có thể chia thành các phương pháp sau:

### 2.6.1. Các phương pháp phân hoạch

Đây là các phương pháp tạo phân hoạch cơ sở dữ liệu D có n đối tượng thành k cụm sao cho

- Mỗi cụm chứa ít nhất một đối tượng.
- Mỗi đối tượng thuộc về một cụm duy nhất.
- k là số cụm đã được cho trước.

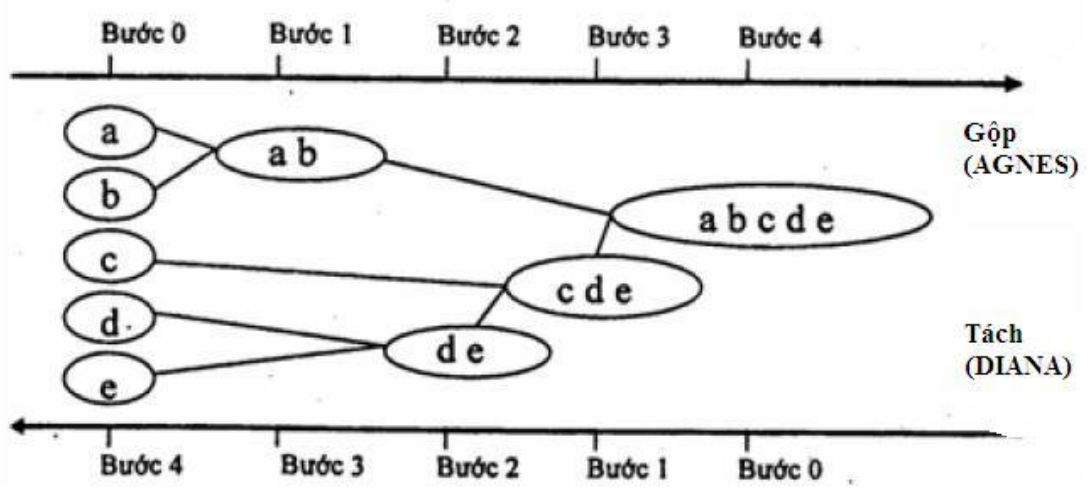
**Các phương pháp tiếp cận phân hoạch**

- Tối ưu toàn cục bằng vét cạn: với k cho trước có thể có ($k^n$ - (k-1) -…- 1) khả năng phân hoạch khác nhau. Đây là con số quá lớn nếu n là khá lớn do đó hầu như không thể thực hiện được.
- Các phương pháp heuristic:
  - K-means (MacQueen'67): Mỗi cụm được đại diện bằng trọng tâm của cụm. Phương pháp này sẽ được trình bày kỹ hơn ở phần sau.
  - K-medoids ( kaufman & Rouseau'87) còn được gọi là PAM( partition around medoids): Mỗi cụm được đại diện bởi một đối tượng của cụm.

### 2.6.2. Phương pháp phân cấp( Hierachical methods)

Đây là các phương pháp tạo phân cấp cụm (hierarchical clustering) chứ không tạo phân hoạch các đối tượng. Phương pháp này không cần phải xác định số cụm từ đầu. Số cụm sẽ do khoảng cách giữa các cụm hoặc điều kiện dừng quyết định. Tiêu chuẩn phân cụm thường được xác định bởi ma trận khoảng cách. Phân cấp cụm thường được biểu diễn dưới dạng đồ thị dạng cây các cụm (dendogram). Lá của cây biểu diễn đối tượng riêng lẻ, nút trong biểu diễn các cụm.

**Các phương pháp tiếp cận để phân cụm phân cấp gồm**




**Hình 2.9: Hai phương pháp tiếp cận phân cấp**

- **Gộp:**

B1. Xuất phát mỗi đối tượng và tạo một cụm chứa nó

B2. Nếu hai cụm đủ gần nhau (dưới một ngưỡng nào đấy) sẽ được gộp lại thành một cụm duy nhất.

B3. Lặp lại B2 dến khi chỉ còn một cụm duy nhất là toàn bộ không gian.

- **Tách:**

B1. Xuất phát từ một cụm duy nhất là toàn bộ không gian.

B2. Chọn cụm có độ phân biệt cao nhất (ma trận phân biệt có phần tử lớn nhất hoặc trị trung bình lớn nhất) để tách đôi. Bước này áp dụng các phương pháp phân hoạch đối với cụm đã chọn.

B3. Lặp lại B2 đến khi mỗi đối tượng thuộc một cụm hoặc đạt điều kiện dừng (đủ số cụm cần thiết hoặc khoảng cách giữa các cụm đạt ngưỡng đủ nhỏ).

**Các khoảng cách giữa các cụm thường được dùng là:**

- Khoảng cách nhỏ nhất. hay còn gọi là khoảng cách lien kết đơn (single link) hay khoảng cách người láng giềng gần nhất. Đây là loại khoảng cách phù hợp để phát hiện các cụm có dạng chuỗi hơn là dạng khối.

$$d(C_i,C_j) = \min_{x\in Ci,\ y\in Cj}\{d(x,y)\}$$

- Khoảng cách lớn nhất: hay gọi là khoảng cách liên kết hoàn toàn (complete link) hoặc khoảng cách người láng giềng xa nhất. Đây là loại khoảng cách phù hợp để phát hiện các cụm có dạng khối hơn là dạng chuỗi.

$$d(C_i,C_j) = \max_{x\in Ci,\ y\in Cj}\{d(x,y)\}$$

- Khoảng cách trung bình:

$$d(C_i,C_j) = \text{avg}_{x\in Ci,\ y\in Cj}\{d(x,y)\}$$

- Khoảng cách trọng tâm. Khoảng cách giữa hai trọng tâm của hai cụm được chọn làm khoảng cách của hai cụm đó. Khoảng cách phù hợp để phát hiện các cụm có dạng khối và tốc độ tính toán nhanh do chỉ quan tâm đến trọng tâm nên giảm khối lượng tính toán.

### 2.6.3. Các phương pháp dựa trên mật độ (Density based Methods)

Các ký hiệu và khái niệm:

- p, q, o là các điểm dữ liệu bất kỳ (các đối tượng)
- Với Eps dương cho trước,tập hợp NEps(p) ={q | d(q,p) ≤Eps } được gọi là lân cận bán kính Eps của p.
- p được gọi là điểm hạt nhân nếu thỏa mãn

$$|\text{NEps}(p)| \geq \min \text{Pts}$$

Trong đó min Pts: số nguyên dương cho trước, min Pts là ngưỡng tối thiểu để coi một điểm là trù mật. Từ đây khi nói một điểm là hạt nhân thì ta hiểu là nó gắn với một bán kính và một ngưỡng trù mật nhất định.

- p được gọi là điểm biên nếu nó không phải là điểm nhân.
- q được gọi là đi tới được trực tiếp theo mật độ từ p nếu p là một điểm nhân và q thuộc lân cận của p.
- $p_n$ được gọi là đi tới được theo mật độ từ $p_1$ nếu tồn tại một dãy các điểm $p_i$ ($i=2,\ldots,n$) sao cho $p_i$ liên thông mật độ trực tiếp từ $p_{i+1}$.
- p và q được gọi là có kết nối theo mật độ nếu tồn tại điểm o sao cho cả p và q đều liên thông mật độ từ o.

### 2.6.4. Phân cụm dữ liệu dựa trên lưới

Ý tưởng: dùng các cấu trúc dữ liệu dạng lưới với nhiều cấp độ phân giải. Những ô lưới có mật độ cao sẽ tạo thành những cụm. Phương pháp này rất phù hợp với các phân tích phân cụm ừng dụng trong không gian (phân loại sao, thiên hà, …). Ngoài ra còn có các thuật toán khác như thuật toán STING, WaveCluster, CLIQUE.

### 2.6.5. Phương pháp dựa trên mô hình (Gom cụm khái niệm, mạng neural)

Đây là các phương pháp dựa trên sự phù hợp giữa dữ liệu và các mô hình toán học. Ý tưởng của các phương pháp này là: Dữ liệu phát sinh từ một sự kết hợp nào đó của các phân phối xác xuất ẩn. Có hai phương pháp tiếp cận chính:

- Tiếp cận thống kê (phương pháp COBWEB, CLASSIT, AUTOCLASS).
- Tiếp cận mạng noron (học cạnh tranh, bản đồ tự cấu trúc SOM).

## 2.7. Các vấn đề có thể gặp phải:

- Các kỹ thuật phân cụm hiện tại chỉ giải quyết được một phần các yêu cầu của bài toán.
- Một vấn đề thường gặp trong phân cụm là hầu hết các dữ liệu cần cho phân cụm đều có chứa dữ liệu nhiễu do quá trình thu thập thiếu chính xác hoặc thiếu đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lí dữ liệu nhằm khắc phục hoặc loại bỏ nhiễu trước khi chuyển sang giai đoạn phân tích cụm dữ liệu.
- Việc phân cụm một dữ liệu với kích thước và số lượng lớn là vấn đề khó khăn bởi vì độ phức tạp thời gian tăng cao.
- Khả năng hiệu quả của các phương pháp phân cụm phụ thuộc vào định nghĩa "khoảng cách" (khi phân cụm dựa trên khoảng cách);
- Nếu một khoảng cách không tồn tại, thì chúng ta phải "định nghĩa" nó, quá trình thực hiện việc này không hề dễ dàng, đặc biệt là trong không gian đa chiều.

## 2.8. Phương pháp phân hoạch (Partion Methods)

### 2.8.1. Thuật toán K-Means

Cho k là số cụm sau khi phân hoạch. ($1 \leq k \leq n$, với n là số điểm( đối tượng) trong không gian giữ liệu)

Thuật toán k-means gồm 4 bước:

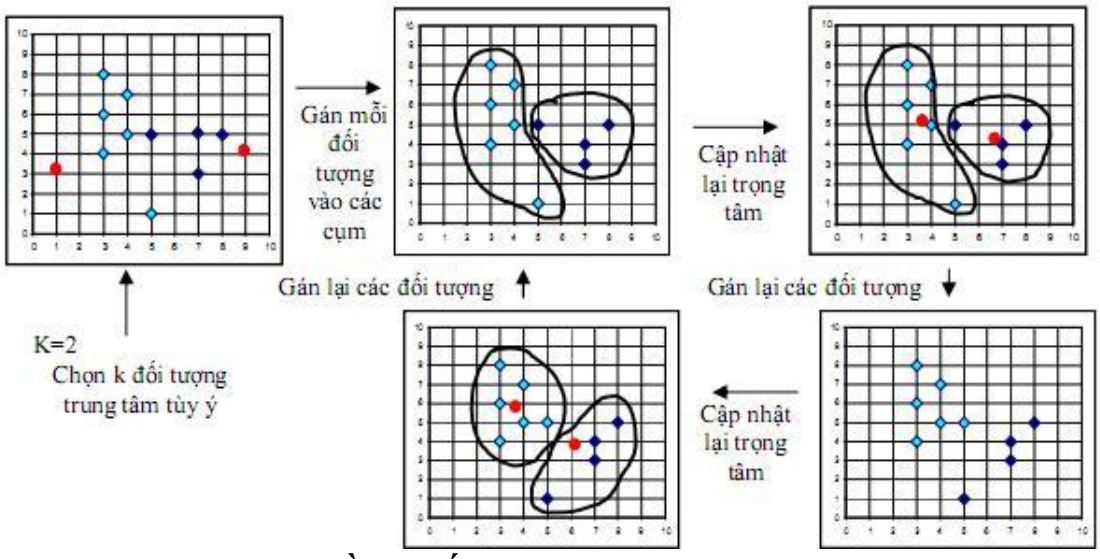
B1. Chọn ngẫu nhiên k điểm làm trọng tâm ban đầu của k cụm.

B2. Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần điểm đang xét nhất. Nếu không có phép gán nào thì dừng. Vì không có phép gán nào có nghĩa là các cụm đã ổn định và thuật toán không thể cải thiện làm giảm độ phân biệt hơn được nữa.

B3. Tính lại trọng tâm cho từng cụm.

B4. Quay lại bước 2.

Minh họa thuật toán với k=2




***Hình 2.10: Ví dụ về một số hình dạng cụm dữ liệu được khám phá bởi K-means***

**Ưu điểm của phương pháp gom cụm k-means**

- Tương đối nhanh .Độ phức tạp của thuật toán là O(tkn) với t là số lần lặp ( t khá nhỏ so với n), k là số cụm cần phân hoạch, n là số điểm trong không gian dữ liệu.
- K-means phù hợp với các cụm có dạng hình cầu.

**Nhược điểm của phương pháp k-mean**

- Không đảm bảo đạt được tối ưu toàn cục và kết quả đầu ra phụ thuộc nhiều vào việc chọn k điểm khởi đầu. Do đó có thể phải chạy lại thuật toán với nhiều bộ khởi đầu khác nhau để có được kết quả đủ tốt. Trong thực tế có thể áp dụng thuật giải di truyền để phát sinh các bộ khởi đầu.
- Cần phải xác định trước số cụm.
- Khó xác định số cụm thực sự mà không gian dữ liệu có. Do đó có thể phải thử với các giá trị k khác nhau.
- Khó phát hiện các loại cụm có hình dạng phức tạp và nhất là các dạng cụm không lồi.
- Không thể xử lý nhiễu và mẫu cá biệt.
- Chỉ có thể áp dụng khi tính được trọng tâm.

### 2.8.2. Thuật toán K-Medoids

Thuật toán K-Medoids là cải tiến của thuật toán k-means, k-medoids khác k-means ở:

- Chiến lược chọ k trọng tâm đầu tiên.
- Phương pháp tính độ phân biệt
- Phương pháp tính trọng tam trong cụm

Thuật toán K-Medoids được thực hiện qua các bước sau:

B1: Chọn ngẫu nhiên k điểm $O_i$ ( i=1,…,k) làm trung tâm (medoids) ban đầu của k cụm.

B2: Gán ( hoặc gán lại) từng điểm vào cụm có trung tâm gần điểm đang xét nhất

B3: Với mỗi điểm trung tâm $O_i$ ( i=1,…,k):

B3.1. Lần lượt xét các điểm không là trung tâm (non-medoids) x.

B3.2. Tính S là độ lợi khi hoán đổi $O_i$ bởi x. S được xác định như sau:

$$S=E_x - E_{Oi}$$

với $E_{Oi}$ và $E_x$ lần lượt là giá trị hàm mục tiêu trước và sau khi thay $O_i$ bởi x.

$$E = \sum_{i=1}^{k} \sum d(p, O_i)^2$$

B3.3. Nếu S là âm thì thay thế $O_i$ trong bộ k trung tâm bởi x ( chọn trung tâm mới tốt hơn).

B4. Nếu có ít nhất 1 sự thay đổi trong B3 thì tiếp tục quay lại B2. Ngược lại thì kết thúc thuật toán.

**Ưu điểm thuật toán K-medoids**

K-medoids làm việc được với nhiễu và biệt lệ.

**Nhược điểm thuật toán K-medoids**

K-medoids chỉ hiệu quả khi tập dữ liệu không quá lớn vì có độ phức tạp là $O(k(n-k)^2t)$. Trong đó:

n là số điểm trong không gian dữ liệu, k là số cụm cần phân hoạch, t là số lần lặp ( t khá nhỏ so với n).

## 2.9. Kết luận chương 2

Trong chương 2 chúng ta có 2 vấn đề quan tâm đó là phân cụm dữ liệu và các giải thuật theo tiếp cận phân hoạch.

Mục đích của phân cụm dữ liệu là gom các dữ liệu tương tự nhau thành những cụm, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định.Phân cụm dữ liệu là một trong những hướng nghiên cứu trọng tâm của lĩnh vực khai phá dữ liệu khám phá tri thức.

Ưu điểm các giải thuật theo tiếp cận phân hoạch là đơn giản, dễ áp dụng và hiệu quả đối với cơ sở dữ liệu nhỏ với các cụm đưa ra có hình dạng lồi. Tuy nhiên, do các cụm trong phương pháp phân hoạch được biểu diễn bởi các tâm của cụm và mỗi một điểm dữ liệu được chia vào một cụm dựa vào khoảng cách từ điểm đó tới tâm của cụm. Chính vì thế phương pháp phân hoạch chỉ có thể đưa ra được các cụm có hình dạng là đa giác lồi mà không thể đưa ra được các cụm có dạng lõm phủ lên nhau hoặc lồng nhau. Ngoài ra, nếu cơ sở dữ liệu có nhiễu hoặc có đối tượng dữ liệu quá xa tâm (outline) thì phương pháp phân cụm phân hoạch cùng không áp dụng được vì trong các trường hợp đó, các đối tượng dữ liệu nhiễu hoặc các đối tượng dữ liệu xa tâm (outline) sẽ làm tâm của cụm bị lệch đi. Do đó, không đưa ra được các cụm chính xác.

# Chương 3: CÀI ĐẶT VÀ THỬ NGHIỆM

## 3.1. Môi trường cài đặt

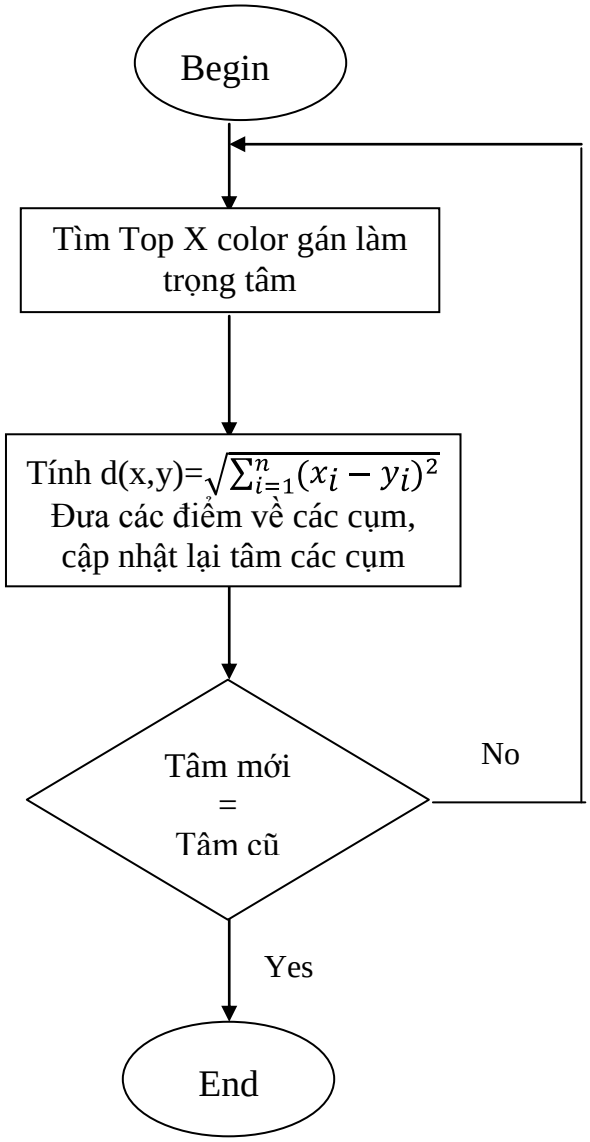
Chương trình được lập trình với ngôn ngữ C# của Visual Studio 2008. Được cài đặt và chạy trên windown XP SP3.

**Input:** Đưa vào một bức ảnh định dạng JPEG

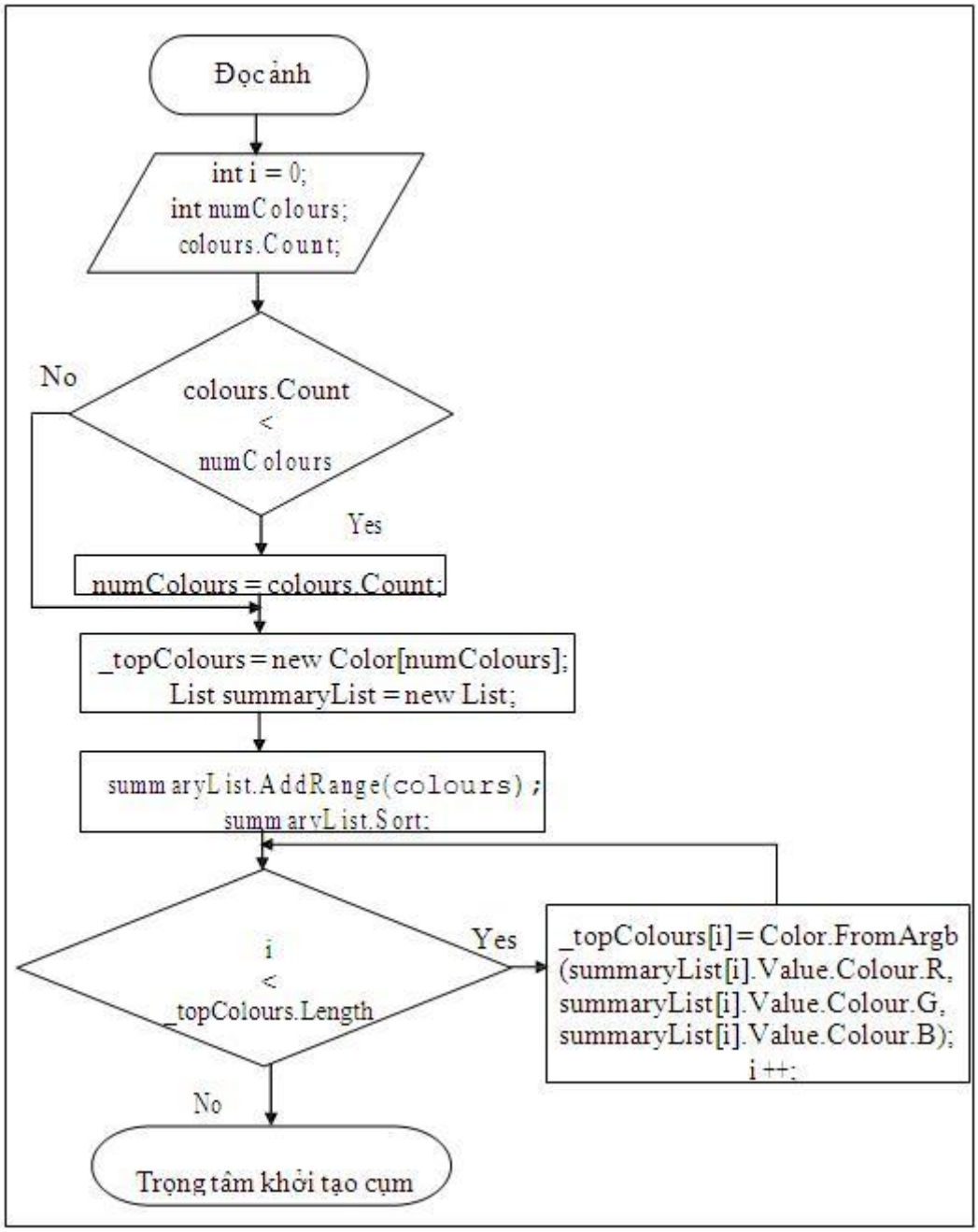
**Output:** Các nhóm (cụm) điểm ảnh, trong đó các điểm ảnh có cùng màu sẽ được gom vào một nhóm.

## 3.2. Giới thiệu chương trình ứng dụng

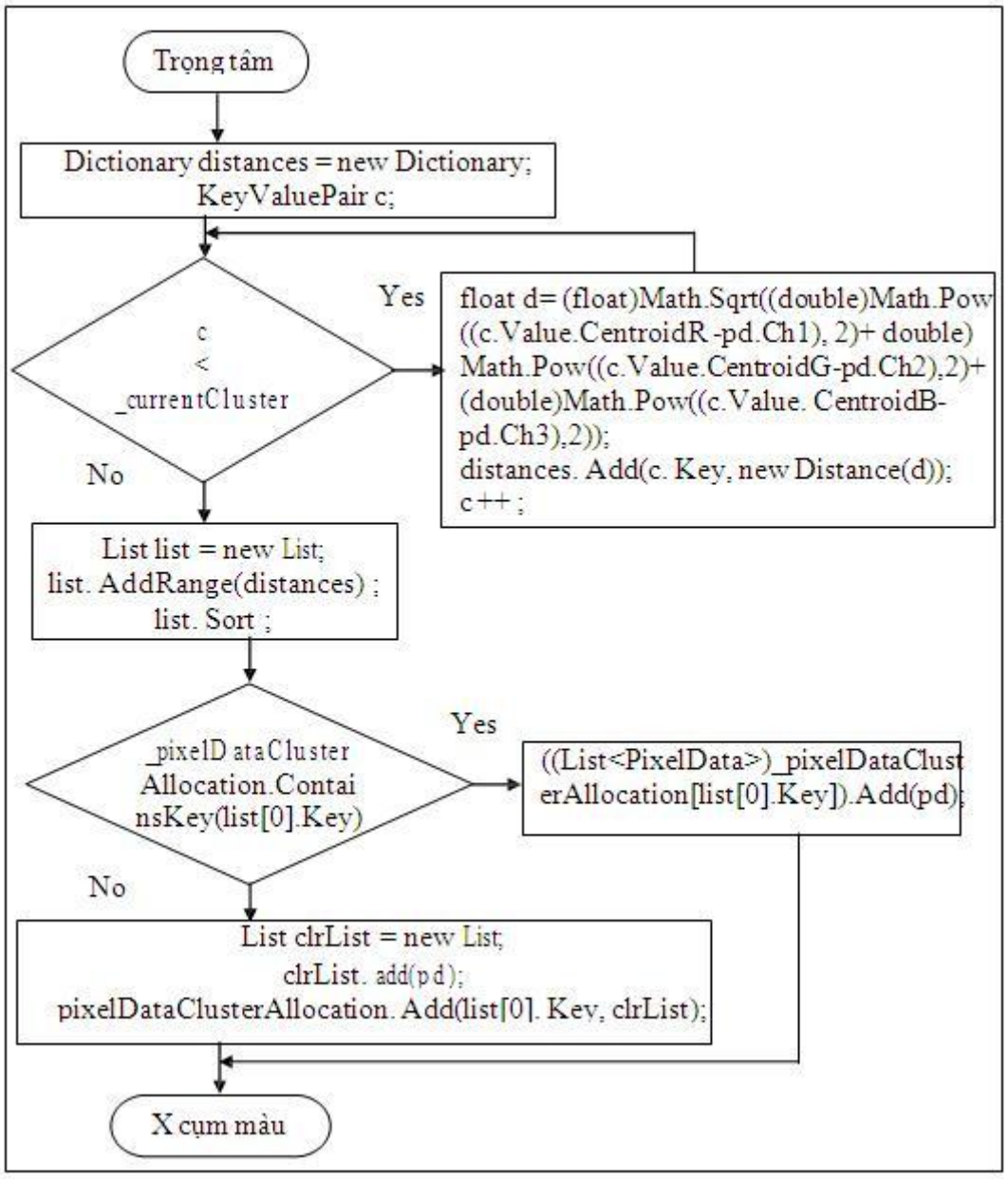
### 3.2.1. Lưu đồ thuật toán sử dụng trong chương trình



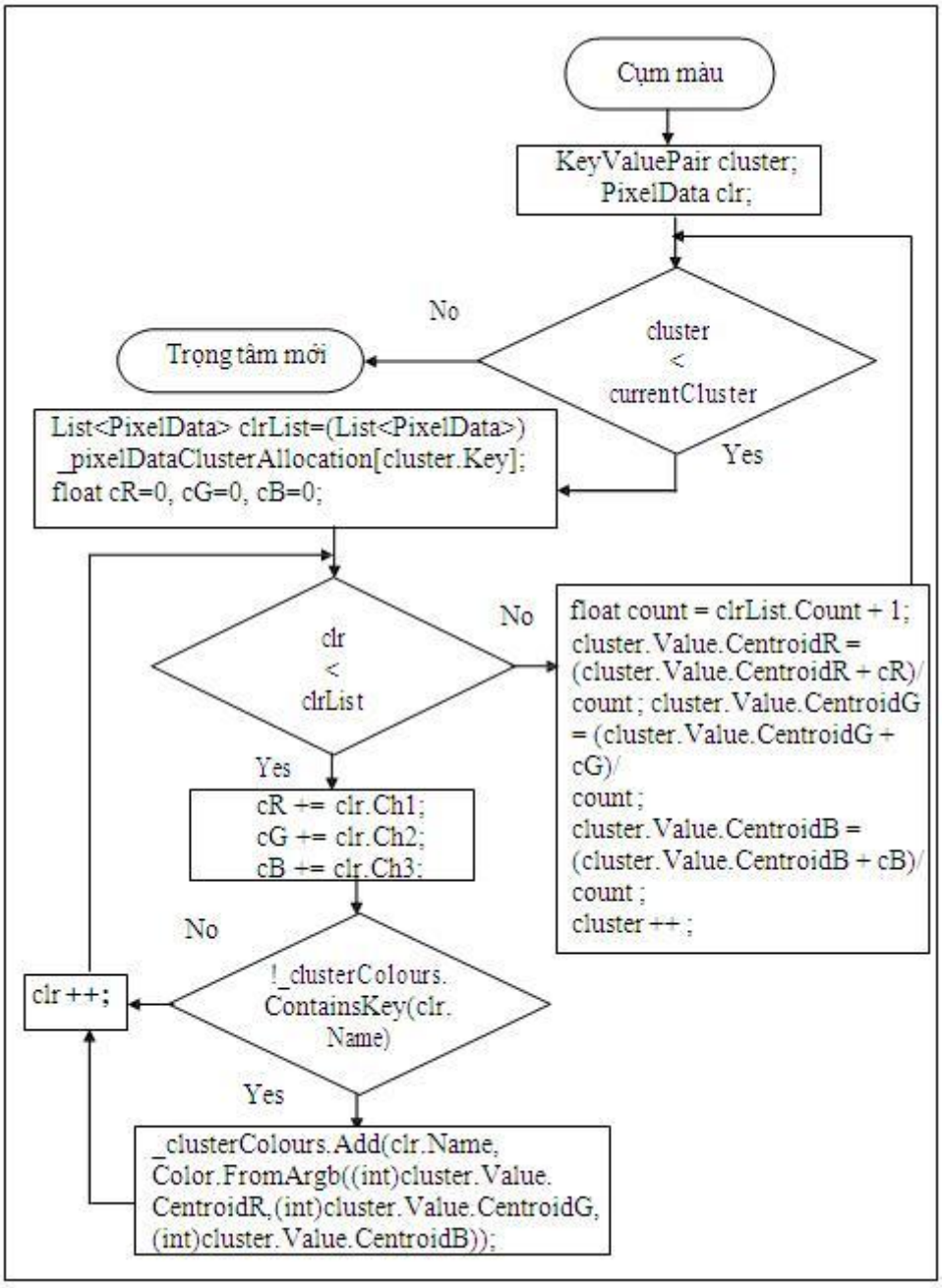
- **Tìm Top X color gán làm trung tâm**



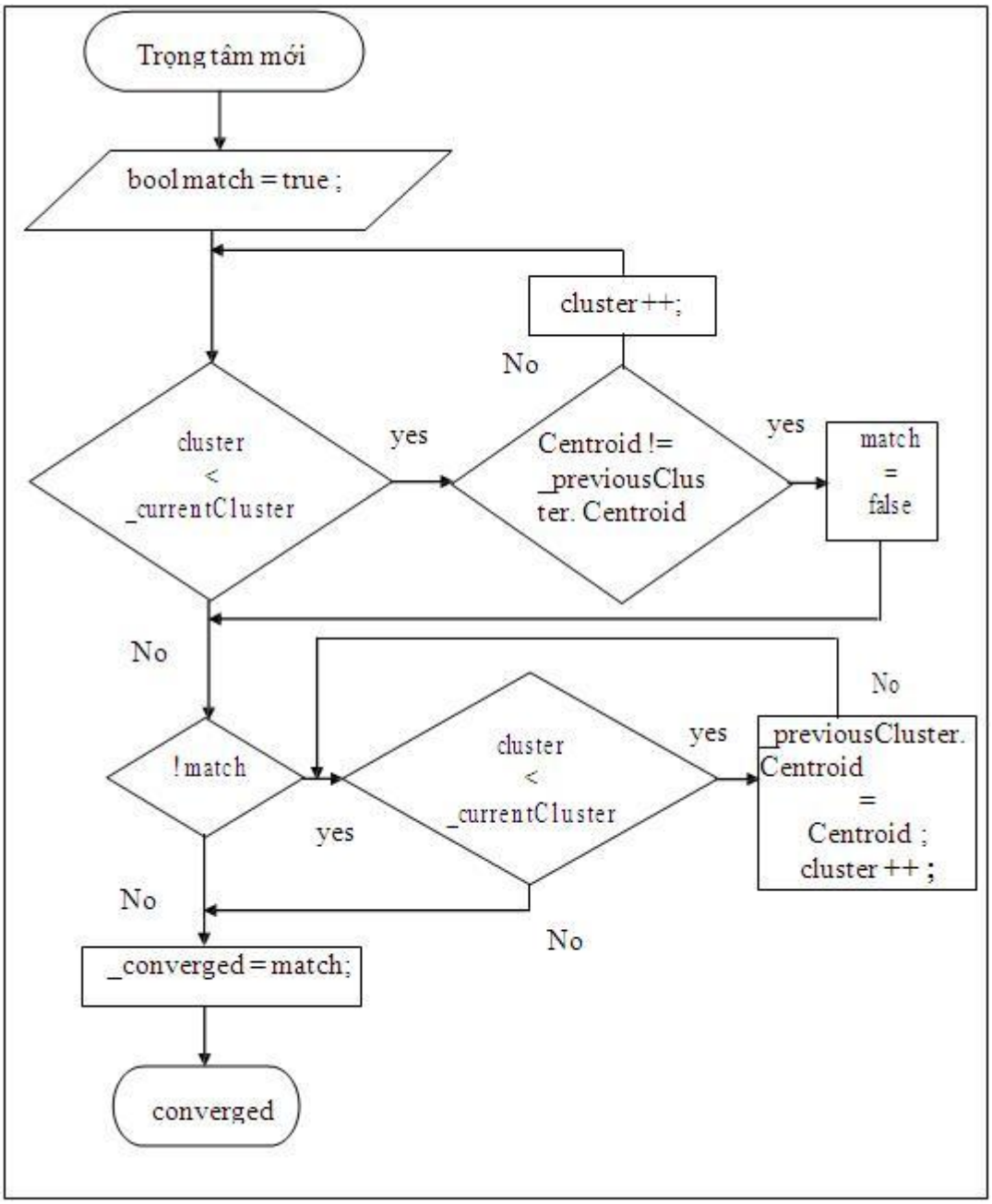
- **Tính khoảng cách và phân cụm**



- **Tính trọng tâm mới**

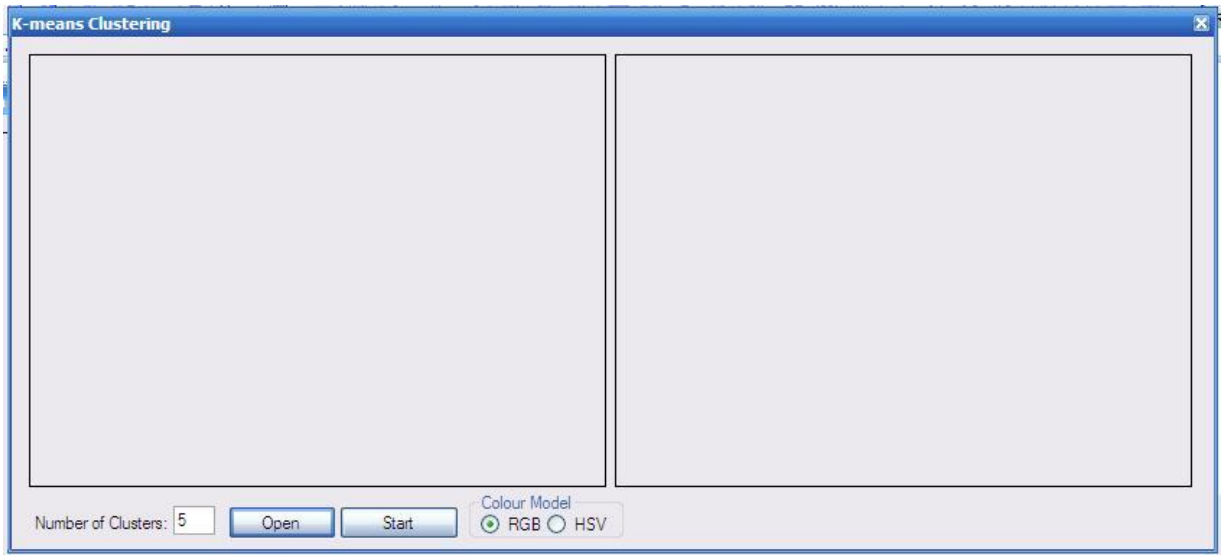


- **Kiểm tra hội tụ**

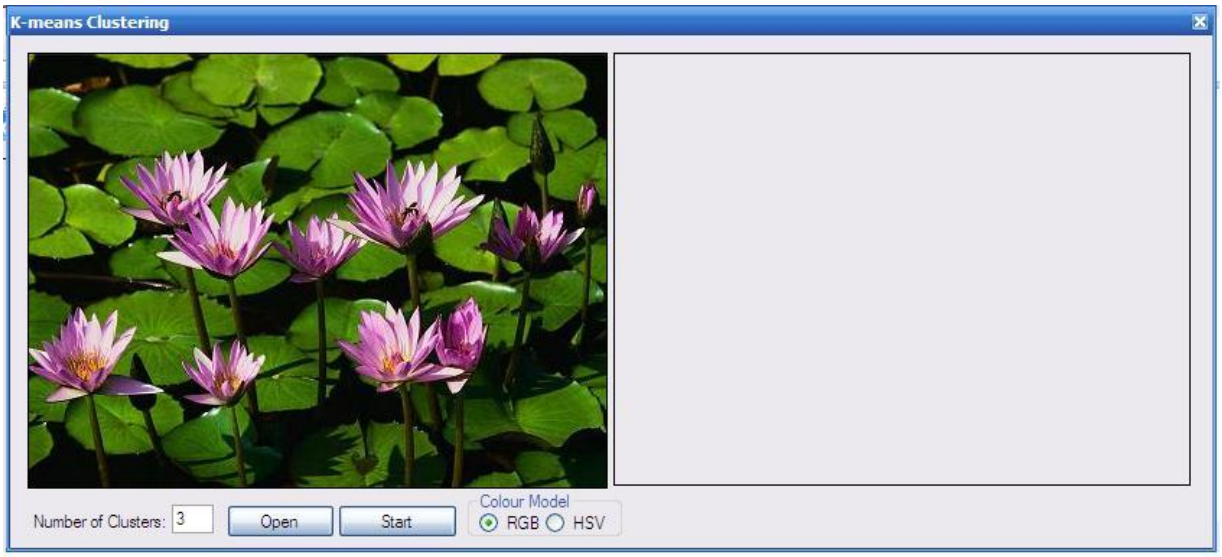


### 3.2.2. Một số giao diện

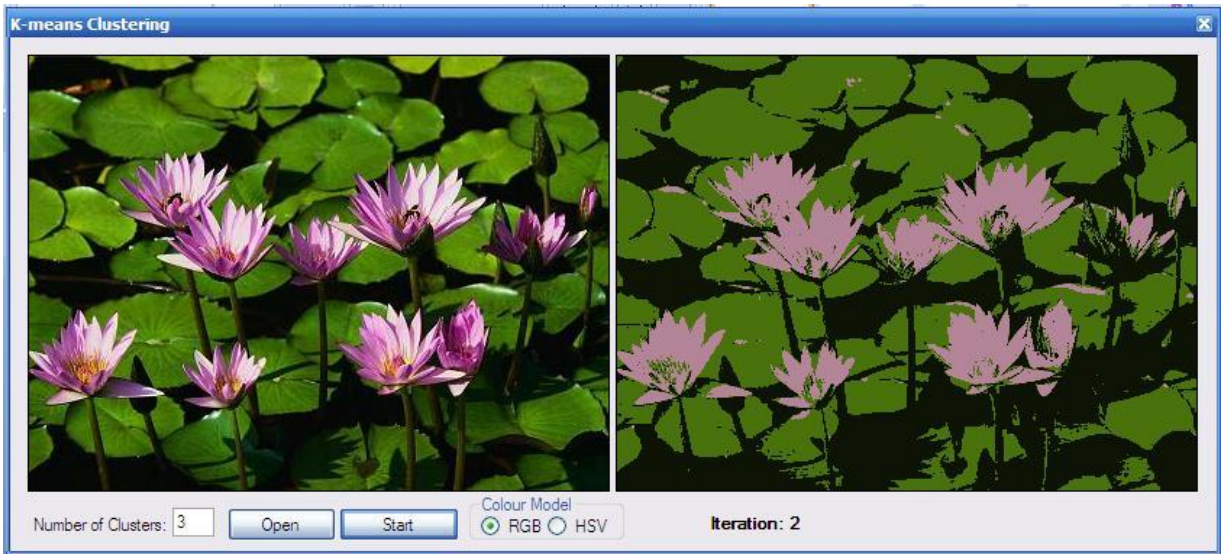
- **Giao diện khởi động**

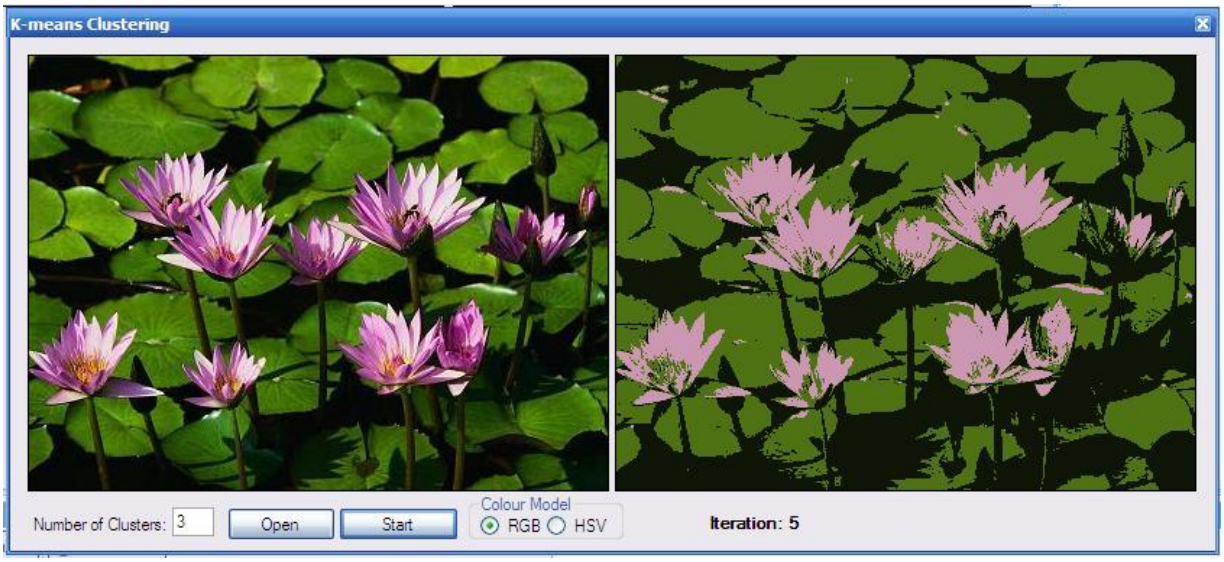



- **Đưa dữ liệu vào xử lý**



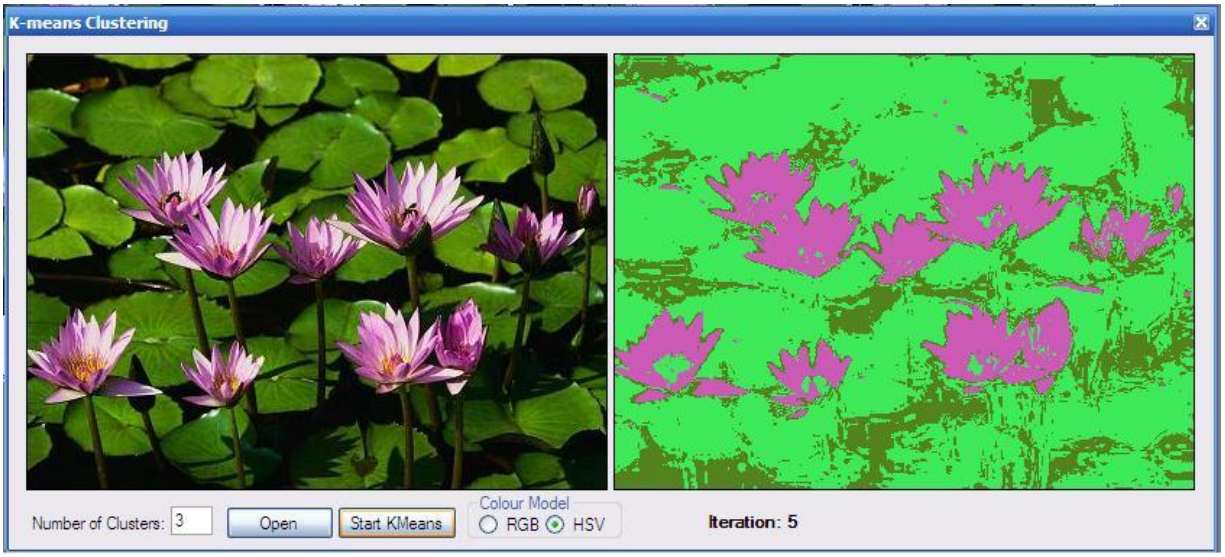
- **Quá trình xử lý dữ liệu.**



- **Quá trình xử lý kết thúc**


- **Chạy Thuật toán K-Means với hệ HSV**



# KẾT LUẬN

Phân cụm dữ liệu là nhiệm vụ quan trọng trong khai phá dữ liệu, thu hút sự quan tâm của nhiều nhà nghiên cứu. Các kỹ thuật phân cụm đã và đang được ứng dụng thành công trong nhiều lĩnh vực khoa học, đời sống xã hội. Hiện nay, do sự phát triển không ngừng của công nghệ thông tin và truyền thông, 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, áp dụng các phương pháp khai phá dữ liệu, trong đó có kỹ thuật phân cum dữ liệu là việc làm rất cần thiết có nhiều ý nghĩa.

Trong đồ án này, trước tiên em đã trình bày những hiểu biết của mình về khai phá dữ liệu sau đó là phần nội dung chính của đồ án: Bài toán phân cụm dữ liệu và một số giải thuật theo tiếp cận phân cấp. Ở phần nội dung chính em đã trình bày được thế nào là bài toán phân cụm dữ liệu, các cách tiếp cận, các ứng dụng, các kiểu dữ liệu có thể phân cụm, các độ đo độ tương tự. Đặc biệt, em tập trung đi sâu nghiên cứu về kỹ thuật phân cụm dữ liệu phân cấp và hai thuật toán điển hình của kỹ thuật này là K-Means và K-Medoids với cách thức tổ chức dữ liệu, thuật toán, đánh giá ưu nhược điểm của mỗi thuật toán.

Do thời gian thực hiện hạn chế nên em mới chỉ tìm hiểu đươc một số kỹ thuật cơ bản trong phân cụm dữ liệu, cài đặt thử nghiệm với thuật toán K- means. Nhưng còn một số các kỹ thuật em vẫn chưa tìm hiểu, khai thác và ứng dụng cho các bài toán … Trong thời gian tới em sẽ cố gắng tiếp tục nghiên cứu, tìm hiểu thêm một sô kỹ thuật phân cụm và nhất là có thể tìm hiểu và phát triển các kỹ thuật phân đoạn ảnh để có thể xử lý với ảnh động.

Tìm hiểu và thử nghiệm thuật toán với một số ứng dụng 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] Nguyễn Trung Sơn, *Phương pháp phân cụm và ứng dụng*, Luận văn thạc sĩ khoa học máy tính, Khoa công nghệ thông tin trường Đại học Thái Nguyên.

[5] Nguyễn Thị Hướng, *Phân cụm dữ liệu trong dataming,* Luận văn tốt nghiệp ngành công nghệ thông tin Đại học sư phạm Hà Nội.

[6] Tian Zhang, Raghu Ramakrishnan, Miron Livny. BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery, 1, 141–182 (1997), Kluwer Academic Publishers, 1997

[7] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, CURE: an efficient clustering algorithm for large databases, Information Systems Vol. 26, No. 1, pp. 35-58, Elsevier Science, 2001.

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