BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

ISO 9001: 2008

ĐỒ ÁN TỐT NGHIỆP

NGÀNH CÔNG NGHỆ THÔNG TIN

HẢI PHÒNG 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

ÁP DỤNG MỘT SỐ THUẬT TOÁN KHAI PHÁ DỮ LIỆU TRONG QUẢN LÝ ĐỊA CHỈ INTERNET

ĐỒ ÁN TỐT NGHIỆP LIÊN THÔNG

Ngành:Công nghệ thông tin

HẢI PHÒNG- 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

ÁP DỤNG MỘT SỐ THUẬT TOÁN KHAI PHÁ DỮ LIỆU TRONG QUẢN LÝ ĐỊA CHỈ INTERNET

ĐỒ ÁN TỐT NGHIỆP LIÊN THÔNG

Ngành:Công nghệ thông tin

Sinh viên thực hiện: Nguyễn Văn Tuyên

Giáo viên hướng dẫn: Nguyễn Trịnh Đông

Mã số sinh viên: 1513101002

HẢI PHÒNG- 2016

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

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do –Hạnh phúc

-------o0o-------

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

Sinh viên: Nguyễn Văn Tuyên

Mã số: 1513101002

Lớp: CTL901

Ngành: Công Nghệ Thông tin

Tên đề tài:

Áp dụng một số thuật toán khai phá dữ liệu trong quản lý địa chỉ Internet

NHIỆM VỤ ĐỀ TÀI

1. Nội dung và yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp

a. Nội dung.

- Tìm hiểu các phương pháp phân cụm.
- Tìm hiểu một số phương pháp tạo các luật cơ bản và các giải thuật liên quan.
- Đề ra phương pháp xâp dựng hệ thống.
- Thử nghiệm với các công cụđể gải quyết bài toán.

b. Các yêu cầu cần giải quyết

2. Các số liệu thống kê, tính toán

3. Địa điểm thực tập

CÁN BỘ HƯỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

Người hướng dẫn thứ nhất:

Họ và tên: Nguyễn Trịnh Đông

Học hàm, học vị: Thạc sĩ

Cơ quan công tác: Trường Đại Học Dân Lập Hải Phòng

Nối dung hướng dẫn:

Tìm hiểu các phương pháp phân cụm.

- Tìm hiểu một số phương pháp tạo các luật cơ bản và các giải thuật liên quan.
- Đề ra phương pháp xâp dựng hệ thống.
- Thử nghiệm với các công cụ để gải quyết bài toán.

Người hướng dẫn thứ hai:

Họ và tên : .....

Học hàm, học vị:.....

Cơ quan công tác: .....

Nội dung hướng dẫn: .....

.....

.....

.....

Đề tài tốt nghiệp được giao ngày 03 tháng 10 năm 2016

Yêu cầu hoàn thành trước ngày 30 tháng 12 năm 2016

Đã nhận nhiệm vụ: Đ. T. T. N	Đã nhận nhiệm vụ: Đ. T. T. N
Sinh viên	Cán bộ hướng dẫn Đ. T. T. N

Hải Phòng,ngày . . . tháng. . . năm 2016

HIỆU TRƯỞNG

GS. TS. NGƯT Trần Hữu Nghị

PHẦN NHẬN XÉT TÓM TẮT CỦA CÁN BỘ HƯỚNG DẪN

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

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

2. Đánh giá chất lượng 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 hướng dẫn:(*Điểm ghi bằng số và chữ*)

. . . . . .
. . . . . .
. . . . . .
. . . . . .

Ngày. . . . . tháng. . . . . năm 2016

Cán bộ hướng 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 như cơ sở lý luận, thuyết minh chương trình, giá trị thực tế, . . .)

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

2.Cho điểm của cán bộ phản biện(*điểm ghi bằng số,chữ*)

. . . . . .
. . . . . .
. . . . . .

Ngày. . . . . tháng. . . . . năm 2016

Cán bộ chấm phản biện

(ký,ghi rõ họ tên)

MỤC LỤC

MỤC LỤC HÌNH ẢNH

LỜI CẢM ƠN

Em xin chân thành cảm ơn thầy giáo Ths. Nguyễn Trịnh Đông đã tận tình chỉ bảo, định hướng, góp ý cho em trong suốt thời gian qua. Để em có thể hoàn thành đồ án tốt nghiệp.Cũng như em xin chân thành cảm ơn các thầy, cô trongKhoa công nghệ thông tin trường ĐHDL Hải Phòng giúp đỡ em. Em cũng xin gửi lời cảm ơn tới gia đình, bạn bè, những người luôn động viên, quan tâm và giúp đỡ em trong suốt thời gian em làm đồ án.

Trong đồ án này chắc còn nhiều thiếu sót. Em rất mong nhận được những lời nhận xét, góp ý từ các thầy, cô giáo và các bạn.

Hải phòng, ngày 23 tháng 12 năm 2016

Sinh viên

Nguyễn Văn Tuyên

GIỚI THIỆU

Sự phát triển của khoa học và công nghệ, cũng như sự phát triển củacông nghệ thông tin đã và đang được áp dụng trong nhiều lĩnh vực trong đời sống, như kinh tế, xã hội, y tế, giáo dục,....Ở mỗi lĩnh vực lại có những bước tiến khác nhau, nhằm phục vụ cho đời sống con người ngày một tốt lên.

Khi khoa học và công nghệ phát triển đã tạo ra những bước tiến to lớn cho con người. Những phát minh ngày càng phongphú và đa dạng. Một trong số đó là mạng Interet. Mạng Internet từ khi được giới thiệu cũng như được sử dụngrộng rãi đến mọi người đã tạo ra một cuộc cách mạng. Và khi đó cần có các chuẩn để mọi người có thể nhìn vào đó để xây dựng lên hệ thống của mình mà có thể trao đổi với hệ thống khác. Từ đó các giao thức được sinh ra như: TCP/IP. Trong đó dịch vụ World Wide Web đã được sinh ra và đã trở thành dịch vụ khá phổ biến trên Internet.

Mỗi quốc gia sẽ có sẽ có những nhà cung cấp khác nhau để có thể phục vụ các nhu cầu đăng ký sử dụng của người dùng. Mỗi ngày có rất nhiều tên miền được đăng ký. Mỗi tên miền sẽ chứa những nội dung có thể giống hoặc khác nhau tùy theo mục đích của người tạo. Khi đó sẽ mỗi nhà cung cấp sẽ có một khối dữ liệu khổng lồ. Và dưới khối dữ liệu khổng lồ đó tiềm ẩn rất nhiều thông tin hữu ích, phục vụ cho việc kinh doanh cũng như đánh giá sự phát triển của xã hội. Nhất là trong việc kinh doanh, khi mà thông tin là một phần cực kỳ quan trọng cho việc đưa ra các định hướng cho việc kinh doanh. Khi đó các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống không thể đáp ứng được, từ đó các nhà khoa học sẽ phải suy nghĩ và đưa ra các cách quản lý và khai thác mới nhằm có thể khai thác dữ liệu một cách tối đa. Khai phá tri thức đã đượcxây dựng nhằm phá tri thức và khai phá dữ liệu phục vụ cho mục đích tìm kiếm thông bên dưới dữ liệu.

Xuất phát từ các lý do trên em chọn đề tài: **“ÁP DỤNG MỘT SỐ THUẬT TOÁN KHAI PHÁ DỮ LIỆU TRONG QUẢN LÝ ĐỊA CHỈ INTERNET. ”**

Mục tiêu của đề tàiáp dụng một số thuật toán khai phá dữ liệu, trong quản lý địa chỉ Internet.

Đề tài được trình bày như sau:

Giới thiệu: *Phát biểu bài toán*

Chương 1: *Trình bày các khái niệm và kiến thức cơ bản trong lĩnh vực khai phá dữ liệu.*

Chương 2: *Chương này tập trung trình bày các thuật toán phục vụ cho việc khai phá dữ liệu.*

Chương 3: *Áp dụng một số thuật toán cho khai phá dữ liệu.*

Kết luận

Tài liệu tham khảo

CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ KHAI PHÁ DỮ LIỆU

1. Giới thiệu

1.1. Mở đầu

Hiện nay, sự phát triển nhanh chóng của Internet đã giúp cho việc trao đổi thôngtin giữa các tổ chức, công ty, cá nhân ngày càng gia tăng. Khi đó, mỗi công ty, tổ chức,cá nhân sẽ có rất nhiều thông tin.Sau một thời gian,các thông tin quá nhiều. Khi đó sẽ cần các cách quản lý tốt hơn, nhằm phục vụ cho mục đích đó đã hình thành các khái niệm DATAMINING vàWEBMINING. Trong đồán chúng ta chỉ quan tâm đến DATA MINING.

1.2.Khai phá dữ liệu

Khai phá dữ liệu được định nghĩa là sử dụng các hệ chuyên gia, hệ lập lịch, hệhọc máy,… và CSDL hoặc kho dữ liệu. Nhằm phân tích đánh giá rút, trích tri thức để đưa ra các quy luật, dự đoán để hỗ trợ cho việc quyết định.

1.3. Phạm vi của khai phá dữ liệu

Khai phá dữ liệu được sử dụng rộng rãi ở nhiều lĩnh vực khác nhau. Như thống kê, học máy cơ sở dữ liệu.

Trong học máy, khai phá dữ liệu đưa ra những thông tin cụ thể khá chính xác, để từ đó đưa vào các thuật toán được xây dựng sẵn trên máy nhằm trích chọn đưa ra các dự đoán trong tương lai. Học máy và khai phá dữ liệu luôn song hành với nhau, mục tiêu tuy khác nhau, nhưng lại có liên quan mật thiết với nhau.

Trong lĩnh vực thống kê, khai phá dữ liệu là tiền đề để đưa ra các thông tin cụ thể tùy theo mục đích của người thống kê. Tuy trong thống kê chỉ cần những thông tin chưa đầy đủ chưa tìm ra hết những thông tin, nhưng với những thông tin chi tiết từ bước khai phá sẽ giúp việc thống kê dễ dàng hơn. Độ tin cậy cao hơn. Tuy cơ sở dữ liệu truy vấn truyền thống (SQL) có thể phần nào đáp ứng được nhu cầu, nhưng vẫn có những thông chưa được tìm ra. Dữ liệu có nhiều loại khác nhau và mỗi loại dữ liệu là các môi trường khác nhau để khai phá.

1.4. Mục tiêu của khai phá dữ liệu

Từ những gì được trình bày ở trên chúng ta có thể thấy các mục đích của khai phá dữ liệu như sau:

- *Khai phá thông tin tìm kiếm tri thức nhỏ được dấu kín trong kho thông tin. Trích rút thông tin, dựa trên các thông tin đã rút trích để đưa ra dự báo dữ liệu tương lai. Chỉ ra xu hướng có thể xuất hiện cho việc kinh doanh, hay sự thay đổi của xã hội.*
- *Tìm ra các quy luật mô tả sao cho con người có thể hiểu được dữ liệu đó. Thông qua việc rút trích phân tích dữ liệu.*

1.5. Các kỹ thuật khai phá dữ liệu

- Cây quyết định.
- Luật kết hợp.
- Các phương pháp phát triển tri thức qua việc học tập mẫu.
- Khoảng cách ngắn nhất.
- Phân cụm (clustering).

1.6. Ứng dụng của khai phá dữ liệu

Các kỹ thuật khai phá dữ liệu có thể được áp dụng vào trong nhiều lĩnh vực, điển hình như sau:

- Thông tin thương mại:
  - Phân tích dữ liệu tiếp thị và bán hàng và thị trường.
  - Phân tích vốn đầu tư.
  - Quyết định cho vay vốn.
  - Phát hiện gian lận.
- Thông tin sản xuất:
  - Điều khiển và lập lịch.
  - Hệ thống quản lý.
  - Quản trị mạng.
  - Phân tích kết quả thí nghiệm.
- Thông tin khoa học:
  - Dự báo thời tiết.
  - Cơ sở dữ liệu sinh học.
  - Khoa học địa lý: tìm động đất; …
- Thông tin cá nhân

1.7. Các khó khăn trong khai phá dữ liệu

Khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực trong thực tế, vìvậy các thách thức và khó khăn ngày càng nhiều, càng lớn. Một số các thách thứcvà khó khăn cần được quan tâm:

Các cơ sở dữ liệu lớn, các tập dữ liệu cần xử lý có kích thước rất lớn,trongthực tế, kích thước của các tập dữ liệu thường ở mức tera-byte.

- Mức độ nhiễu cao hoặc dữ liệu bị thiếu (nhiều thông tin sai lệch)
- Số chiều lớn (nhiều dữ liệu giữa được khai thác)
- Thay đổi dữ liệu và tri thức có thể làm cho các mẫu đã phát hiện khôngcòn phù hợp
- Quan hệ giữa các trường phức tạp(cơ sở dữ liệu lớn, nhiều quan hệ ràng buộc)

2. Chi tiết các bước khai phá tri thức

Một tiến trình khám phá tri thức gồm 6 giai đoạn.

Bước 1: Chọn lọc dữ liệu (data selection).

Bước 2: Xóa bỏdữ liệu không cần thiết (cleaning).

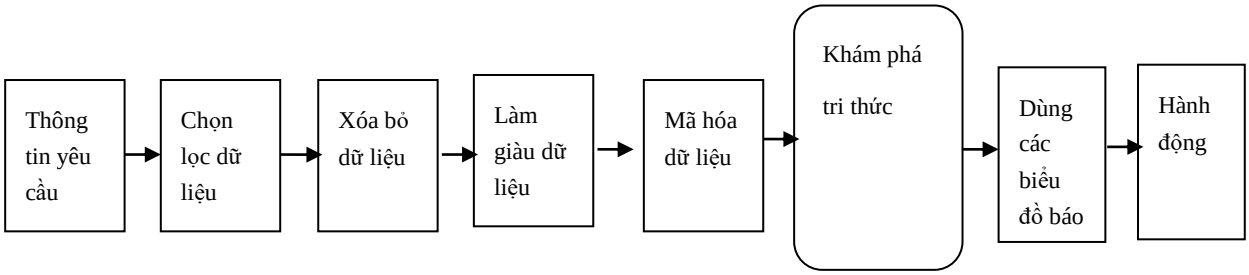
Bước 3: Làm giàu dữ liệu (enrichment).

Bước 4: Mã hóa (coding).

Bước 5: Khám phá tri thức (data mining).

Bước 6: Báo cáo kết quả (reporting).

Bên trên là 6 bước khai phá tri thức nhưng thực ra việc khai phá tri thức chỉ thực sự bắt đầu ở bước thứ 5 mà thôi.



Hình 1:Mô hình khai phá dữ liệu

2.1. Lựa chọn dữ liệu (data selection)

Trong việc quản lý dữ liệu các cơ sở dữ liệu sẽ được xây dựng ở khắp mọi nơi chúng ta cần lựa chọn, một cách tốt nhất nhằm phục vụ cho việckhai phá. Ở bước này cần có sự phân tích cao nhất, tránh sai sót để đưa ra một bộ dữ liệu hoàn chỉnh.

2.2.Xóa bỏ dữ liệu không cần thiết (cleaning)

Các cơ sở dữ liệu sau khi được tập hợp sẽ được tập trung tại một chỗ. Khi đó trong dữ liệu sẽ có các dữ liệu không cần thiết cho việc khai phá. Chúng ta cần phải xóa bỏ chúng để cơ sở dữ liệu trở lên linh hoạt và thuận tiện nhất.

Giai đoạn này có thể được thực hiện nhiều lần trong quá trình khai phá. Dữ liệu cuối cùng cần là tốt nhất tránh sai xót, để khi khai phá tránh đưa ra dữ liệu không tốt.

2.3.Làm giàu dữ liệu (enrichment)

Trong giai đoạn này chúng ta cần bổ sung thông tin cho cơ sở chính bằng cách đưa liên kết với cơ sở dữ liệu ngoài. Những cơ sở dữ liệu có liên quan đến cơ sở dữ liệu chính. Chọn lọc các cơ sở dữ liệu phù hợp bổ sung cho cơ sở dữ liệu chính. Làm cho thông tin chính rõ ràng hơn, nhằm phục vụ cho việc khai phá dữ liệu tốt nhất.

Khi có sự kết hợp giữa hai cơ sở dữ liệu chúng ta cần lưu ý đến các mối quan hệ có thể giữa hai cơ sở dữ liệu. Việc làm giàu có thể rất có ích nếu chúng ta xây dựng đúng cách. Nếu bước này làm sai sẽ gây ra việc khó lắm dữ liệu cho bước sau, làm bước sau khó đoán nhận dữ liệu.

2.4. Chuẩn hóa và mã hóa (coding and normalzation)

Mục đích chính của giai đoạn này là biến cơ sở dữ liệu về dạng mà khi triển khai các thuật toán khai phá dữ liệu là tốt nhất. Nhưng không phải loại dữ liệu nào cũng có thể mã hóa được, mà tùy loại dữ liệu mà chúng ta sử dụng các cách mã hóa khác nhau.

2.5. Khám phá tri thức (datamining)

Sử dụng các thuật toán khai phá dữ liệu để tìm kiếm tri thức trong cơ sở dữ liệu. Trong giai đoạn này chúng ta có rất nhiều các thuật toán để phù hợp với mọi loại dữ liệu chúng ta thu thập được. Giai đoạn này sẽ được đề cập cụ thể hơn ở chương tiếp theo.

2.6. Báo cáo kết quả (reporting)

Đây là giai đoạn cuối cùng của quá trình khai phá tri thức. Tổng hợp dữ liệu đã khai phá tri thức thông báo kết quả. Đưa ra tóm tắt sao cho người đọc dễ hiểu, dễ tiếp cận dữ liệu quan trọng.

3. Chi tiết mã hóa và biến đổi dữ liệu

Ngoài một số cách mã hóa như trên chúng ta còn có một số cách phương pháp biến đổi để có thể khai phá. Trong phần này đề cập đến phép biến đổi và chuẩn hóa dữ liệu.

3.1. Phép biến đổi và chuẩn hóa dữ liệu

Trong thực tế sau khi đã có dữ liệu từ nhiều nguồn khác nhau, chúng ta chưa thể khai phá ngay được. Chúng ta cần đưa về một loại dữ liệu nhất định.

3.1.1. Phép chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu sẽ làm cho dữ liệu ban đầu nhỏ đi tốt cho việc phân cụm dữ liệu. Việc chuẩn hóa sẽ biến đổi vị trí, cấu trúc dữ liệu ban đầu hoặc có thể bị mất đi[2]. Có hai phương pháp chuẩn hóa là: **Chuẩn hóa toàn cục và chuẩn hóa trong cụm.**

Chuẩn hóa toàn cục: làm chuẩn hóa các biến trên tất cả các yếu tố trong các tập dữ liệu. Trong vòng-cụm tiêu chuẩn hóa dùng để chỉ tiêu chuẩn hóa xảy ra trong các cụm biến mỗi ngày. Một số hình thức tiêu chuẩn hóa có thể được sử dụng trong các chuẩn hóa toàn cục và chuẩn hóa trong phạm vi rất tốt. Tuy nhiên trong một số trường hợp chúng ta chỉ có thể sử dụng trong chuẩn hóa toàn cục.

Chuẩn hóa trong cụm: Để khắc phục nhược điểm của chuẩn hóa toàn cục là chỉ chuẩn hóa khi dữ liệu cho trước. Khi đó tổng thể và [6]đề xuất một cách tiếp cận lặp rằng các cụm thu được đầu tiên dựa trên số ước lượng tổng thể và sau đó sử dụng kết quả của cụm này để **so sánhvới cụm khác** để xem sự chênh lệch trong cụm có lớn không.

3.2.Biến đổi dữ liệu

Biến đổi dữ liệu tác động lên dữ liệu chuẩn hoá, nhưng biến đổi dữ liệuphức tạp hơn so với chuẩn hoá dữ liệu. Chuẩn hoá dữ liệu tập trung vàocác biến, nhưng biến đổi dữ

liệu tập trung vào các dữ liệu toàn bộ thiết lập.Trong phần này, trình bày một số dữ liệukỹ thuật biến đổi có thể được sử dụng trong phân cụm dữ liệu.

3.2.1. Phân tích thành phần chính

Mục đích chính của phân tích thành phần chính là giảm chiều cao của một chiều cao của một chiều đặt dữ liệu bao gồm một lượng lớn số biến tương quan và đồng thời giữ lại càng nhiều càng tốt của biến đổi hiện diện trong tập dữ liệu. Các thành phần chính (PC) là các biến mới được không tương quan và ra lệnh như vậy là người đầu tiên giữ lại vài phần lớn các biến thể hiện diện trong tất cả các bản gốc biến.[3]

3.2.2. SVD (Singular Value Decomposition)

SVD(*phân tách giá trị riêng*) là một kỹ thuật mạnh mẽ trong tính toán ma trận và phân tích, chẳng hạn như việc giải quyết các hệ thống phương trình tuyến tính và xấp xỉ ma trận. SVD cũng là một kỹ thuật nổi tiếng chiếu tuyến tính và đã sử dụng rộng rãi trong nén dữ liệu và ảo.

3.2.3. Phép biến đổi Karhunen-Loéve

Các phép biến đổi Karhunen-Loeve (KL) có liên quan với các giải thích cấu trúc dữ liệu thông qua một số tuyến tính kết hợp của các biến. Giống như PCA, phép biến đổi KL cũng là cách tối ưu cho dự án, tính toán sao cho sai số là nhỏ nhất (tức là tổng khoảng cách bình phương (SSD) là tối thiểu [7].

4. Địa chỉ Internet

Đồ án tập trung khai phá dữ liệu địa chỉ Internet nhằm tìm ra những thông tin về loại dữ liệu người dùng thường truy cập, sở thích, thói quen,…. Những thông tin trên, sẽ cho chúng ta biết được sự quan tâm của mọi người trong một khoảng thời gian sẽ như thế nào.

4.1. Giới thiệu địa chỉ Internet

IPlàmột giao thức hướng dữ liệu được sử dụng bởi các máy chủ nguồn và đích để truyền dữ liệu trong một liên mạng chuyển mạch gói.Dữ liệu trong một liên mạng IP được gửi theo các khối được gọi là các gói (*packet* hoặc *datagram*). Cụ thể, IP không cần thiết lập các đường truyền trước khi một máy chủ gửi các gói tin cho một máy khác mà trước đó nó chưa từng liên lạc với nhau.[4]

4.2. Cấu trúc của địa chỉ Internet

Địa chỉ IP được dùng phổ biến hiện nay là IPv4, và một số nước đang sử dụng song song giữa IPv4 và IPv6.

Cấu trúc của IPv4

IPv4 sử dụng 32 bits để đánh địa chỉ, được chia thành 4 octet, theo đó, số địa chỉ tối đa có thể sử dụng là 4.294.967.296 (2^{32}). Tuy nhiên, trong thực tế chúng ta đã sử dụng gần hết địa chỉ cũng như một số địa chỉ được sử dụng cho mục đích khác.Với sự phát triển không ngừng của mạng Internet, nguy cơ thiếu hụt địa chỉ đã được dự báo, tuy nhiên, nhờ công nghệ NAT (Network Address Translation - Chuyển dịch địa chỉ mạng) tạo nên hai vùng mạng riêng biệt: **Mạng riêng** và **Mạng công cộng**, địa chỉ mạng sử dụng ở mạng riêng có thể dùng lại ở mạng công cộng mà không hề bị xung đột, giải quyết được vấn đề thiếu hụt địa chỉ .

Ban đầu, một địa chỉ IP được chia thành hai phần:

- Network ID: Xác lập bởi octet đầu tiên
- Host ID: Xác định bởi ba octet còn lại

Với cách chia này, số lượng network bị giới hạn ở con số 256, quá ít so với nhu cầu thực tế.

Để vượt qua giới hạn này, việc phân lớp mạng đã được định nghĩa, tạo nên một tập hợp lớp mạng đầy đủ (classful). Theo đó, có 5 lớp mạng (A, B, C, D và E) được định nghĩa.

Class A: 0. 0. 0. 0 - 127. 255. 255.255Default Subnet Mask: 255. 0. 0. 0

Class B: 128. 0. 0. 0 - 191. 255. 255. 255Default Subnet Mask: 255. 255. 0. 0

Class C: 192.0. 0. 0 - 223.255. 255. 255Default Subnet Mask: 255. 255. 255. 0

Class D: 224.0. 0. 0 - 239. 255. 255. 255: multicast/broadcast

Class E: 240. 0. 0. 0 - 255. 255. 255. 255: reserve (bảo tồn).

-->*các thiết bị chỉ đặt được IP trong dải A, B, C*.

IP Private: trích 1 phần nhỏ từ 3 class A, B, C.

- 10. 0. 0. 0 - 10. 255. 255. 255

- 172.16. 0. 0 - 172.31. 255. 255

- 192.168. 0. 0 - 192.168. 255. 255

Cấu trúc IPv6.

IPv6 viết tắt tiếng Anh: "Internet Protocol version 6", là "Giao thức liên mạng thếhệ 6", một phiên bản của giao thức liên mạng (IP) nhằm mục đích nâng cấp giao thức liên mạng phiên bản 4 (IPv4) hiện đang truyền dẫn cho hầu hết lưu lượng truy cập.

Internet nhưng đã hết địa chỉ. IPv6 cho phép tăng lên đến 2^{128} địa chỉ, một sự gia tăng khổng lồ so với2^{32} (khoảng 4.3 tỷ) địa chỉ của IPv4.[4]

Phiên bản địa chỉ Internet mới IPv6 được thiết kế để thay thế cho phiên bản IPv4, với hai mục đích cơ bản:

- *Thay thế cho nguồn IPv4 cạn kiệt để tiếp nối hoạt động Internet.*
- *Khắc phục các nhược điểm trong thiết kế của địa chỉ IPv4.*

Mục tiêu IPv6.

- *Không gian địa chỉ lớn hơn và dễ dàng quản lý không gian địa chỉ.*
- *Khôi phục lại nguyên lý kết nối đầu cuối-đầu cuối của Internet và loại bỏ hoàn toàn công nghệ NAT.*
- *Quản trị TCP/IP dễ dàng hơn: DHCP được sử dụng trong IPv4 nhằm giảm cấu hình thủ công TCP/IP cho host. IPv6 được thiết kế với khả năng tự động cấu hình mà không cần sử dụng máy chủ DHCP, hỗ trợ hơn nữa trong việc giảm cấu hình thủ công.*
- *Cấu trúc định tuyến tốt hơn: Định tuyến IPv6 được thiết kế hoàn toàn phân cấp.*
- *Hỗ trợ tốt hơn Multicast: Multicast là một tùy chọn của địa chỉ IPv4, tuy nhiên khả năng hỗ trợ và tính phổ dụng chưa cao.*
- *Hỗ trợ bảo mật tốt hơn: IPv4 được thiết kế tại thời điểm chỉ có các mạng nhỏ, biết rõ nhau kết nối với nhau. Do vậy bảo mật chưa phải là một vấn đề được quan tâm. Song hiện nay, bảo mật mạng Internet trở thành một vấn đề rất lớn, là mối quan tâm hàng đầu.*
- *Hỗ trợ tốt hơn cho di động: Thời điểm IPv4 được thiết kế, chưa tồn tại khái niệm về thiết bị IP di động. Trong thế hệ mạng mới, dạng thiết bị này ngày càng phát triển, đòi hỏi cấu trúc giao thức Internet có sự hỗ trợ tốt hơn.* [4]

IPv6 có độ dài 128bit, biểu diễn ở dạng số Hecxa, chia thành 8 octet. Mỗi ocet có 4 số hecxa, và cách nhau bởi dấu ":".

Ví dụ:0123: 4567: 89AB: CDEF: 0123: 4567: 89AB: CDEF

Tương thích giữa IPv4 và IPv6 khi chuyển đổi:

- Dual-stack: thiết bị vừa chạy được IPV4, vừa chạy được IPv6

- Tunneling: khi 2 đoạn IPv6 bị chia cắt bởi đoạn IPv4.

- Translation: các đoạn IPv4 và IPv6 nối liên tiếp nhau.

Một số cách viết IPv6:

- *Trong 1 octet, ta có thể xóa số 0 ở ngoài cùng bên trái.*

Ví dụ 1:

0123: 4567: 89AB: CDEF: 0123: 4567: 89AB: CDEF

123: 4567: 89AB: CDEF: 123: 4567: 89AB: CDEF

Ví dụ 2:

1234: 0010: 3456: 7890: 000A: ABCE: 1234: 4567

1234: 10: 3456: 7890: A: ABCE: 1234: 4567

Trong 1 octet toàn 0, ta có thể giữ lại một số 0

Ví dụ:

1234: 0000: 3456: 7890: 0000: ABCE: 1234: 4567

1234: 0: 3456: 7890: 0: ABCE: 1234: 4567

Nếu có từ 2 octet trở lên toàn 0, thì ta có thể viết gọn thành dấu": :"

Nhưng chú ý là 1 IPv6 chỉ được viết ": :"một lần.

Ví dụ:

1234: 0000: 0000: 1234: 0000: 0000: 0000: 1234

1234: : 1234: 0: 0: 0: 1234

1234: 0: 0: 1234: : 1234

Một số không gian IPv6:

- Global Unicast: giống như IPv4 Public.

→chạy từ 2000: → 3FFF:

- Link Local: giống như IPv4 APIPA (169. 254.0. 0/16)

→ thiết bị dùng IPv6 lúc nào cũng tự sinh ra Link-local address, không quan tâm tới việc nó đã được đặt IP hay có DHCP hay chưa.

→ chạy từ FE80: → FEBF:

- Loopback: giống như IPv4 127. 0. 0. 0/8

→với IPv6 là ": :"1

- Dải đặc biệt:(0: 0: 0: 0: 0: 0: 0: 0)

- Unique Local: giống như IPv4 Private

→ chạy từ FC00: → FDFF:

4.3.Hệ thống tên miền (DNS)

Hệ thống tên miền là một hệ thống cho phép thiết lập tương ứng giữa địa chỉ IP và tên miền trên Internet.

Hệ thống tên miền về căn bản là một hệ thống giúp cho việc chuyển đổi các tên miền mà con người dễ ghi nhớ (dạng kí tự, ví dụ www. example. com) sang địa chỉ IP vật lý (dạng số, ví dụ 123.11.5.19) tương ứng của tên miền đó. Hệ thống tên miền giúp liên kết với các trang thiết bị mạng cho các mục đích định vị và địa chỉ hóa các thiết bị trên Internet.

Phép so sánh thường được sử dụng để giải thích cho hệ thống tên miền, nó phục vụ như một "Danh bạ điện thoại", có khả năng tìm kiếm và dịch tên miền thành địa chỉ IP. Ví dụ, www.example.com dịch thành 208.100.188.166. Tên miền Internet dễ nhớ hơn các địa chỉ IP, là 208.100.188. 166 (IPv4) hoặc 2001:db8:1f70: :999:de8:7648:6e8 (IPv6).

4.4.Chức năng hệ thống tên miền

Mỗi website có một tên (là tên miền hay đường dẫn URL: Uniform Resource Locator) và một địa chỉ IP. Địa chỉ IP gồm 4 nhóm số cách nhau bằng dấu chấm(IPv4). Khi mở một trình duyệt Web và nhập tên website, trình duyệt sẽ đến thẳng website mà không cần phải thông qua việc nhập địa chỉ IP của website. Quá trình "dịch" tên miền thành địa chỉ IP để cho trình duyệt hiểu và truy cập được vào website là công việc của một máy chủ hệ thống tên miền.

Các máy chủ DNS trợ giúp qua lại với nhau để **dịch địa chỉ "IP" thành "tên" và ngược lại**. Người sử dụng chỉ cần nhớ "tên", không cần phải nhớ địa chỉ IP (địa chỉ IP là những con số rất khó nhớ).

4.4 Tổ chức quản lý IP và Hệ thống tên miền

Tổ chức cấp phát số hiệu Internet (tên tiếng Anh là **Internet Assigned Numbers Authority –IANA**) là một cơ quan giám sát việc chỉ định địa chỉ IP, quản lý khu vực gốc của DNS toàn cầu, và cấp phát giao thức Internet khác. Tổ chức này được điều này bởi ICANN.

Trước khi **ICANN** được thành lập với mục đích này, IANA chủ yếu do Jon Postel quản lý tại Viện Khoa học Thông tin của trường Đại học Nam California, dưới một hợp đồng USC/ISI với Bộ Quốc phòng Hoa Kỳ, cho đến khi ICANN được thành lập để nhận trách nhiệm dưới hợp đồng của Bộ Thương mại Hoa Kỳ.

Tập đoàn Internet cấp số và tên miền (*Internet Corporation for Assigned Names and Numbers*- **ICANN**) là một tổ chức phi lợi nhuận đặt trụ sở tại Marina del Rey, California, United States. ICANN được thành lập ngày 18 tháng 9 năm 1998 và hợp nhập vào ngày 30 tháng 9 năm 1998 để giám sác một số nhiệm vụ liên quan tới Internet mà trước đây được thực hiện trực tiếp bởi các tổ chức khác trên danh nghĩa của chính phủ Mỹ, mà đáng chú ý trong số đó là IANA ICANN chịu trách nhiệm trong việc quản lý không gian địa chỉ IP(IPv4 và IPv6) và việc phân phối các khối địa chỉ tới các cơ quan đăng ký Internet khu vực. Duy trì các cơ quan đăng ký tên định danh IP; Quản lý không gian tên miền cấp cao nhất(miền DNS gốc), bao gồm việc điều hành của những máy phục vụ tên gốc. Phần lớn các công việc của ICANN liên quan tới việc giới thiệu của những miền cấp cao mới (top-level domains (TLDs)). Công việc kĩ thuật của ICANN giống như chức năng của IANA.

Những nguyên tắc cơ bản hàng đầu trong việc điều hành của ICANN được mô tả như việc giúp đỡ duy trì sự hoạt động ổn định của Internet; thúc đẩy việc cạnh tranh; đạt được sự đại diện rộng rãi của cộng đồng Internet toàn cầu và xây dựng chính sách phù hợp với nhiệm vụ của ICANN thông qua các quá trình từ dưới lên, dựa trên sự nhất trí ý kiến. Vào ngày 29 tháng 9 năm 2006, ICANN đã ký một thỏa thuận với Bộ Thương mại Hoa Kỳ về việc đưa tổ chức tư nhân vào sự quản lý toàn diện của hệ thống các tên định danh được điều phối tập trung của Internet thông qua mô hình nhiều phía cùng có lợi trong việc trao đổi ý kiến mà ICANN đại diện.[4]

Ở Việt Nam **Trung tâm Internet Việt Nam** hay tên viết tắt là VNNIC là một đơn vị trực thuộc Bộ Thông tin và Truyền thông, nước Cộng hòa Xã hội Chủ nghĩa Việt Nam được thành lập chính thức vào ngày 28 tháng 04 năm 2000. Theo đó, Trung tâm Internet Việt Nam chịu trách nhiệm quản lý về tên miền Internet trong lãnh thổ Việt Nam cũng như thống kê về tình hình sử dụng Internet tại Việt Nam.

Một số nhiệm vụ của VNNIC. Theo quyết định số 02/2008/QĐ-BTTTT do Bộ trưởng Bộ Thông tin và Truyền thông, Lê Doãn Hợp ban hành ngày 05/03/2008, quy định các chức năng, nghĩa vụ và quyền hạn của VNNIC như sau: [5]

- *Quy hoạch, quản lý và phân bổ địa chỉ (IP) và số hiệu mạng (ASN) ở cấp quốc gia.*
- *Quản lý tên miền Internet cấp quốc gia bao gồm tên miền các cấp dưới .vn.*
- *Quy hoạch, đầu tư xây dựng cơ sở hạ tầng kỹ thuật, công nghệ và nhân lực để phát triển Trung tâm Internet Việt Nam phù hợp với yêu cầu thực tiễn.*
- *Thiết lập, khai thác và duy trì hoạt động hệ thống máy chủ tên miền (DNS) quốc gia .vn; trạm trung chuyển Internet quốc gia;đăng ký và duy trì địa chỉ IP, số hiệu mạng cho Internet Việt Nam; tham gia khai thác các công nghệ mới liên quan đến tài nguyên Internet, công nghệ DNS và giao thức IP và hệ thống chứng thực CA trên Internet.*
- *Kiểm tra, giám sát việc cấp, đăng ký, sử dụng địa chỉ IP, số hiệu mạng và tên miền đối với các tổ chức, cá nhân tham gia hoạt động Internet.*
- *Nghiên cứu đề xuất và tham gia với các đơn vị chức năng trực thuộc Bộ để xây dựng các văn bản quy phạm pháp luật về công tác quản lý nhà nước về tài nguyên Internet, về khai thác, sử dụng dịch vụ và chất lượng Internet trên phạm vi cả nước.*
- *Phối hợp với các đơn vị chức năng của Bộ trong công tác quản lý nhà nước đối với các hoạt động của hội và tổ chức phi chính phủ trong lĩnh vực Internet.*
- *Tham gia đại diện chính thức về Internet của Việt Nam, tham gia các hoạt động của các tổ chức Internet quốc tế liên quan đến tài nguyên mạng Internet và công nghệ IP.*
- *Được quyền yêu cầu các tổ chức, cá nhân hoạt động trên mạng Internet cung cấp các thông tin và các số liệu thống kê liên quan tới hoạt động Internet. Thực hiện báo cáo thống kê tình hình phát triển Internet trong nước.*
- *Được thu phí và lệ phí các hoạt động theo chức năng, nhiệm vụ, quyền hạn của Trung tâm và theo quy định của pháp luật.*
- *Tham gia việc đào tạo, bồi dưỡng chuyên môn nghiệp vụ theo quy định của Bộ Thông tin và Truyền thông.*
- *Được phối hợp, hợp tác với các tổ chức quốc tế để khai thác dự phòng hệ thống cho tên miền quốc gia .vn, đăng ký và duy trì tài nguyên Internet Việt Nam, quảng bá quốc tế về Internet Việt Nam và phát triển sử dụng tên miền. vn.*
- *Được tham gia cung cấp các dịch vụ liên quan đến tài nguyên Internet, công nghệ IP, công nghệ thông tin và tham gia các hoạt động có liên quan để tạo thêm các nguồn thu khác nhằm mở rộng phạm vi và quy mô hoạt động phù hợp với chức năng, nhiệm vụ, quyền hạn của Trung tâm và theo quy định của pháp luật, bảo toàn và phát triển các nguồn lực được giao.*
- *Quản lý về tổ chức, cán bộ, viên chức và tài sản của Trung tâm theo quy định của pháp luật và phân cấp của Bộ trưởng.*
- *Thực hiện các nhiệm vụ khác do Bộ trưởng giao.*

CHƯƠNG 2: CÁC THUẬT TOÁN TRONG KHAI PHÁ DỮ LIỆU

1. Giới thiệu phân cụm dữ liệu

1.1. Định nghĩa phân cụm

Phân cụm dữ liệu (Data Clustering) là quá trình nhóm các điểm dữ liệu trong cơ sở dữ liệu thành cáccụm sao cho những điểm dữ liệu trong cùng một cụm có độ tương đồng lớn và những điểm không cùng một cụm có sự tương đồng là rất nhỏ. Một cụm các đối tượng dữ liệu có thể xem như là một nhóm trong nhiều ứng dụng.

Quá trình phân cụm là quá trình tìm ra các đối tượng trong cơ sở dữ liệu mộtcách tự động. Không giống như phân lớp (clasification), phân cụm không cần những thông tin được xác định trước. Nói cách khác, phân cụm là phương pháp học từ quan sát (learning from obversation) hay còn gọi là học không thầy (unsupervised learning or automatic classfication) trong trí tuệ nhân tạo. Phân cụm đặc biệt hiệu quả khi không biết về thông tin các cụm, hoặc khi ta quan tâm tới các thuộc tính của cụm mà chưa biết hoặc biết rất ít về các thông tin đó.

Đã có rất nhiều thuật toán cũng như hệ thống được phát triển cho bài toán phân cụm trong cơ sở dữ liệu lớn. Sự phát triển của lĩnh vực này đã được áp dụngvào nhiều lĩnh vực ứng dụng như xử lý ảnh, nhận dạng, đánh giá kinh doanh….Sựđa dạng của thuật toán phân cụm là do sự khác nhau của những ứng dụng thực tếcũng dẫn tới những yêu cầu về dữ liệu khác nhau và đòi hỏi những thuật toán phân cụm khác nhau.

Một trong những câu hỏi lớn đặt ra trong bài toán phân cụm là đo độ tương đồng không gian giữa các đối tượng dữ liệu (spatial similarity). Trong dữ liệu không gian thì độ đo tương đồng được xem như sự quan hệ về vị trí không gian giữa các đối tượng dữ liệu. Nói cách khác thì hai đối tượng dữ liệu được gọi là tương đồng nếu "khoảng cách không gian" giữa chúng là nhỏ. Một trong những phương pháp đo độtương đồng giữa hai đối tượng là bằng nghịch đảo của hàm không tương đồng (dissimilarity function).Hàm không tương đồng, hàm dựa trên những thuộc tính không gian của các đối tượng dữ liệu như: toạđộ của các đối tượng, độ cao của các đối tượng… Trong nhiều trường hợp thì hàm không tương đồng được xem như là hàm khoảng cách không gian giữa các đối tượng như hàm khoảng cách Euclid, hàm khoảng cách Manhattan, hàm khoảng cách Minkowski…

Bài toán phân cụm là quá trình nhóm một cơ sở dữ liệu thành những nhóm đối tượng dữ liệu phục vụ cho mục đích cụ thể của từng ứng dụng thực tế. Không có một thuật toán phân cụm nào là tốt nhất và thích hợp cho tất cả mọi ứng dụng mà với mỗi ứng dụng khác nhau thì người sử dụng phải lựa chọn ra một thuật toán phân cụm cụ thể thích ứng với ứng dụng đó. Kết quả đánh giá cho từng thuật toán cũng phụ thuộc vào những yêu cầu của từng ứng dụng.

1.2.Mục đích của phân cụm

Mục đích của phương pháp phân cụm dữ liệu là quá trình nhóm các điểm dữ liệu trong cơ sở dữ liệu thành các cụm sao cho những điểm dữ liệu trong cùng một cụm có độ tương đồng lớn và những điểm không cùng một cụm có sự tương đồng là rất nhỏ. Điểm mạnh của phân cụm dữ liệu là đưa ra được những cấu trúc có ích hoặc những cụm các đối tượng tìm thấy trực tiếp từ dữ liệu mà không cần bất kì một tri thức cơ sở nào. Giống như cách tiếp cận học máy, phân cụm dữ liệu được hiểu như là phương pháp "học không có thầy" (*unsupervised learning*). Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát (*learning by observation*), trong khi phân lớp dữ liệu là học bằng ví dụ (*learning by example*).

Trong phương pháp này sẽ không thể biết kết quả các cụm thu được sẽ như thế nào khi bắt đầu quá trình. Vì vậy, cần có một chuyên gia đểđánh giá các cụm thu được. Phân cụm dữ liệu được sử dụng nhiều trong các ứng dụng về phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, phân loại trang Web… Ngoài ra, phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lí cho các thuật toán khai phá dữ liệu khác.

1.3.Những lĩnh vực áp dụng phân cụm

- Word Wide Web: Phân loại tài liệu. Phân loại người dung web …

- Marketing: Phân tích đánh giá người dung, xu hướng mua sắm, sử dụng dịch vụ. Nhằm đưa ra các chính sách, các ưu đái, các chính sách kinh doanh sau này.

- Tài chính, bảo hiểm: Phân nhóm khách hàng sử dụng các dịch vụ bảo hiểm, tài chính. Phát hiện xu hướng đầu tư, các gian lận trong tài chính, bảo hiểm.

- Thư viện: Theo dõi, độc giả, sách, dự đoán nhu cầu của độc giả, …

- Giáo dục: Theo dõi sinh viên, học sinh. Tìm ra việc học tập và dạy học sao cho tốt nhất.

1.4.Các yêu cầu về thuật toán phân cụm

Theo các nghiên cứu cho thấy hiện này chưa có một phương pháp phân cụmtổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cơ sở dữ liệu. Hơnnữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của các cơ sở dữ liệu,với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng thuật toán phân cụm phùhợp. Vì vậy, phân cụm dữ liệu vẫn đang là một 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 khai phá dữ liệu.

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 thướccủa dữ liệu.*
- *Thuật toán phải xử lý và áp dụng được với cơ sở dữ liệu nhiều nhiễu, phứctạp gồm cả dữ liệu không gian, phi không gian, dữ liệu số, phi số, kiểu nhịphân, 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 kỳbao gồm cả những cụm có hình dạng lồng nhau, cụm có hình dạng lõm, hìnhcầu, hình que…*

- *Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào. Do các giá trịđầu vào thường ảnh hưởng rất lớn đến thuật toán phân cụm và rất phức tạpđể xác định các giá trị vào thích hợp đối với các cơ sở dữ liệu lớn.*
- *Thuật toán phải thực hiện với mọi thứ tự đầu vào dữ liệu. Nói cách khác kếtquả của thuật toán nên độc lập với dữ liệu đầu vào (cùng một tập dữ liệu, khiđưa vào xử lý cho thuật toán phân cụm dữ liệu với các thứ tự vào của các đối tượng dữliệu ở các lần thực hiện khác nhau thì không ảnh hưởng lớn đến kết quả phâncụm).*
- *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 dữliệu phức tạp và có tính chất khác nhau.*
- *Thuật toán phải thích nghi với dữ liệu đa chiều: Thuật toán có khả năng ápdụng hiệu quả cho dữ liệu có số khác chiều nhau.*
- *Thuật toán phải dễ hiểu, dễ cài đặt và khả thi: Người sử dụng có thể chờ đợinhững kết quả phân cụm dễ hiểu, dễ lý giải và dễ sử dụng. Nghĩa là, sự phâncụm có thể cần được giải thích ý nghĩa và ứng dụng rõ ràng. Việcnghiên cứucách để một ứng dụng đạt được mục tiêu rất quan trọng có thể gây ảnhhưởng tới sự lựa trọn các phương pháp phân cụm.*

1.5. Các kiểu dữ liệu 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 dữliệu" 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 bảng 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 nhau:

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Ma trận không giống nhau(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,jkhá khác nhau. Do d(i,j)=d(j,i)=0 nên ta có thể biểu diễn ma trận cách tự như nhau:

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

Phần lớn các thuật toán phân cụm sử dụng cấu trúc ma trận khác nhau. Dovậy, nếu dữ liệu cần phân cụm được tổ chức dưới dạng ma trận dữ liệu thì cần biếnđổi 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 cơ sở dữ liệu D chứa n đối tượng trong không gian k chiều; x, y, z là các đốitượng thuộc D:

$$\mathbf{x=(x_1,x_2,\dots,x_k);y=(y_1,y_2,\dots y_k);z=(z_1,z_2,\dots z_k)}$$

trong đó $\mathbf{x_i}$, $\mathbf{y_j}$, $\mathbf{z_j}$với **i = 1,. . , k** là các đặc trưng hoặc thuộc tính tương ứng củacác đối tượng **x, y, z;** như vậy sẽ có các kiểu dữ liệu sau:

1.5.1. 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 tínhnhị phân mà miền giá trị chỉ có hai phân tử (ví dụ: Yes/No, True/False,On/Off. .).

1.5.2.Kiểu dữ liệu dựa trên hệ đo

Thuộc tính định danh: Là dạng thuộc tính khái quát hoá của thuộc tính nhịphân, trong đó có miền giá trị là rời rạc không phân biệt thứ tự và có nhiềuhơn hai phần tử. Nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xácđịnh là x ≠ y hoặc x =y.

Thuộc tính có thứ tự: Là thuộc tính định danh nhưng có thêm tính thứ tựnhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thìcó thể xác định là x ≠ 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 tínhkhoảng có thể xác định một thuộc tính là đứng trược hoặc đứng sau thuộctính khác với một khoảng là bao nhiêu. Nếu $\mathbf{x_i} > \mathbf{y_i}$thì có thể nói x cách ymột khoảng $\mathbf{x_i} - \mathbf{y_i}$tương ứng với thuộc tính thứ *i*.

Việc lựa chọn đơn vị đo cho các thuộc tính cũng ảnh hưởng đến chất lượngphân cụm. Nếu đơn vị độ đo của một thuộc tính càng được chia nhỏ, thì khoảngcách xác định của thuộc tính đó càng lớn và ảnh hưởng nhiều hơn đến kết quả phâncụm. 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 tínhưu tiên.

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

- Tính độ lệch trung bình:

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

trong đó $\mathbf{x_{1f}}$ **,…,**$\mathbf{x_{nf}}$là giá trị thuộc tính f của n phần tử dữ liệu, và $\mathbf{m_f}$là giá trị trungbình của f, được cho như sau:

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

Độ đo được chuẩn hóa:

$$z_{if} = \frac{x_{if} - m_f}{S_f}$$

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 cáchtương đối so với điểm mốc.

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

Đặc biệt, còn có **dữ liệu không gian** là loại dữ liệu có thuộc tính số khái quáttrong không gian nhiều chiều, dữ liệu không gian mô tả các thông tin liênquan đếnkhông 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 nhiềuchiều và cho phép xác định khoảng cách giữa các đối tượng dữ liệu trong không gian.

1.5.3.Phép đo độ tương tự và khoảng cách đối với các kiểu dữ liệu

a. 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 kỳmột metric nào cũng là một độ đo, nhưng điều ngược lại không đúng. Để tránh sựnhầm lẫn, thuật ngữ độ đo ở đây đề cập đến hàm tính độ tương tự hoặc hàm tính độphi 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 đốitượng dữ liệu trong cơ sở dữ liệu 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 δ(x,y) được gọi là khoảng cách giữa x và y.
- Quy tắc nói trên thỏa mãn hệ tính chất sau:
  - (i) $\boldsymbol{\delta(x,y) > 0}$ nếu $\boldsymbol{x \neq y}$;
  - (ii) $\boldsymbol{\delta(x,y) = 0}$ nếu $\boldsymbol{x = y}$;
  - (iii) $\boldsymbol{\delta(x,y) = \delta(y,x)}$với mọi x,y;
  - (iv) $\boldsymbol{\delta(x,y) \leq \delta(x,z) + \delta(z,y)}$;

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

b. Thuộc tính khoảng

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

Tuy nhiên chú ý rằng đây không phải vấn đề đồ thị: vấn đề phát sinh từ côngthức toán học được sử dụng để kết hợp khoảng cách giữa các thành phần đơn đặctính dữ liệu vectơ vào trong một độ đo khoảng duy nhất mà có thể được sử dụngcho 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 đốitượng để thực hiện phân cụm. Kiến thức miền phải được sử dụng để để trình bày rõràng phép đo khoảng thích hợp cho mỗi ứng dụng. Hiện nay, phép đo có nhiều mứcđộ khách nhau tùy theo từng trường hợp.

- *Khoảng cách Minkowski được định nghĩa như sau*

$$dist_q(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q\right)^{1/q}, q \geq 1;$$

Trong đó x, y là hai đối tượng với n số lượng thuộc tính =(x_1,x_2,…,x_n) và y=(y_1,y_2,…,y_n); dist là kích thước của dữ liệu.

- *Khoảng cách Euclidean*

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

Là khoảng cách giữa hai đối tượng trong trường hợp đặc biệt q=2.

- *Khoảng cách Manhattan*

$$dist_q(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|\right)$$

- *Khoảng cách Chebychev*

$$dist_{\infty}(x,y) = max_{i=1}^{n}|x_i - y_i|;$$

Trong trường hợp q=∞,hữu ích để định nghĩa các đối tượng phi tương tự nếu chúng khác nhau chỉ trong một kích thước biến đổi.

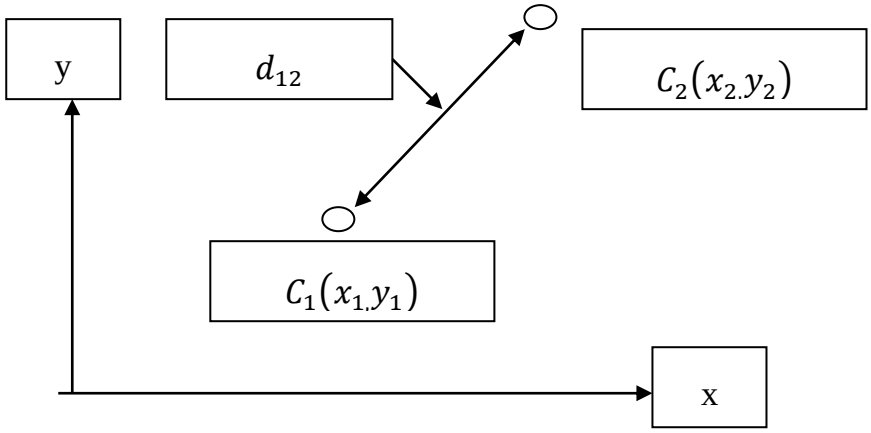
- *Bình phương khoảng cách Euclidean*

$$dist_p(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

Tỷ lệ khác nhau. Giả sử các biến là tuyệt đối.

$$dist(x,y) = \big(Number(x_i \neq y_i)\big)/i$$

Khoảng cách Euclidean được sử dụng phổ biến nhất để đo độ tương tự củakhoảng cách Minkowski. Giả sử có hai trường hợp C1 và C2 có các biến liên tục xvà y, lấy lần lượt các giá trị (x1, y1) và (x2, y2) tương ứng, có thể vẽ đồ thị hai trườnghợp trong không gian x-y:



Hình 2: Tính khoảng cách

Tuy nhiên không có nguyên tắc tổng quát để chọn phép đo áp dụng cho bấtcứ bài toán nào. Một cách đơn giản để đo độ tương tự giữa các nhóm trong khungtương tự bằng cách thay thế nhóm cho thuộc tính thứ i của đối tượng đo chẳng hạnnhư khoảng cách Euclidean, khoảng cách Manhattan, hoặc bình phươngMahalanobis. Ví dụ, Giả sử rằng nhóm A có vectơ trung bình$\overline{\boldsymbol{A}}$=[$\overline{\boldsymbol{x\,a1}}$, $\overline{\boldsymbol{x\,a2}}$,. . . , $\overline{\boldsymbol{x\,an}}$]vànhóm B có vectơ trung bình

$\overline{\boldsymbol{B}}$ =$\left[\overline{\boldsymbol{x}_{b1}}, \overline{\boldsymbol{x}_{b2}},\ldots, \overline{\boldsymbol{xbn}}\right]$, thì cách đo bằng khoảng cáchEuclidean giữa hai nhóm có thể được định nghĩa là:

$$dist(A,B) = \left(\sum_{i=1}^{n}(\overline{x_{ai}} - \overline{x_{bi}})^2\right)^{1/2}$$

Cách tiếp cận khác để khoảng cách giữa phần tử gần nhất hoặc phần tử xanhất. Cách tiếp này sử dụng các thuật toán phân cụm phân cấp chẳng hạn như liênkết đơn và liên kết đầy đủ. Vấn đề chính với hai cách tiếp cận này giống nhau làkhông cảm nhận được mâu thuẫn định lượng và không tính toán cho các yếu tố củacác phần tử trong một nhóm.

Cách tiếp cận khác, là trung bình nhóm, có thể sử dụng phép đo tương tựgiữa các nhóm. Cách tiếp cận này, sự giống nhau giữa các nhóm được đo bằng cáchlấy giá trị trung bình cả tất cả các phép đo giữa các đối tượng cho từng cặp đốitượng trong các nhóm khác nhau. Ví dụ, trung bình phi tương tự giữa nhóm A và Bcó thể được định nghĩa là:

$$dist(A,B) = \left[\sum_{i=1}^{n_x}\sum_{j=1}^{n_b} d(x_i, y_i)\right]/n$$

trong đó, n là tổng số các đối tượng cùng cặp, n = n_x×n_y, n_x và n_y lần lượt là số cácđối tượng trong đối tượng x_i và y_i, d(x_i, y_i) là phi tương tự của một cặp đối tượng $x_i$và y_i, x_i∈A, y_i∈B. Hàm phi tương tự có thể dễ dàng chuyển đổi sang hàm tươngtự bằng cách thay đổi cho nhau.

c. Thuộc tính nhị phân

Tất cả các phép đo được định nghĩa ở trên là đa số thích hợp cho các biếnliên tục. Cho các biến danh nghĩa, "phép đo khoảng cách" là 0 nếu các trường hợpcó cùng giá trị danh nghĩa, và 1 nếu các trường hợp có các giá trị danh nghĩa khácnhau, hoặc với độ đo tương tự 1 (nếu các trường hợp có cùng giá trị danh nghĩa) và0 (nếu không giống nhau).

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

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

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

Các phép đo độ tương tự của các trường hợp với dữ liệu thuộc tính nhị phân được thực hiện bằng các cách sau:

- Hệ số đối sánh đơn giản**:** d(x,y)=$\frac{\boldsymbol{a+d}}{\boldsymbol{p}}$; 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 cógiá trị 1 của đối tượng dữ liệu cao hơn nhiều so với các thuộc tính có giá trị 0. Nhưvậy thuộc tính nhị phân ở đây là không đối xứng.

$$d(x,y)=\frac{a}{p}\quad d(x,y)=\frac{a}{b+c}\quad d(x,y)=\frac{a}{2a+b+c}$$

Các giá trị được định nghĩa trong khoảng [0, 1] và có thể biến đổi sang độ đophi tương tự bằng biểu thức: ds(x,y)=1-d(x,y).

d. 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 cácthuộc tính.

e. Thuộc tính có thứ tự

Phép đo độ phi tương tự giữa các đối tượng dữ liệu với thuộc tính thứ tựđược thực hiện như sau: Giả sử i là thuộc tính thứ tự có M_i giá trị (M_i là kích thướcmiền giá trị):

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

Mỗi thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy phải chuyểnđổi chúng về cùng miền giá trị [0, 1] bằng cách thực hiện phép biến đổi sau cho mỗithuộc 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 cácgiá trị $Z^{(j)}_i$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

f. Thuộc tính tỷ lệ

Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Mộttrong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính x_i, ví dụ qi=log(x_i), q_iđóng vai trò như thuộc tính khoảng. Phép biến đổi logarit này thích hợptrong trường hợp các giá trị của thuộc tính là số mũ.

Trong thực tế, khi tính độ tương tự dữ liệu, chỉ xem xét một phần các thuộctính đặc trưng đối với các kiểu dữ liệu hoặc là đánh trọng số cho tất cả các thuộctính dữ liệu. Trong một số trường hợp, loại bỏ đơn vị đo của các thuộc tính dữ liệubằng cách chuẩn hóa 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, vídụ với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng wi ($1 \leq i \leq k$), độtương đồng dữ liệu được xác định như sau:

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

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

Tóm lại, tùy từng trường hợp dữ liệu cụ thể mà có thể sử dụng các mô hìnhtính độ tương tự khác nhau. Việc xác định độ tương đồng dữ liệu thích hợp, chínhxác đảm bảo khách quan là rất quan trọng, góp phần xây dựng thuật toán phân cụm dữ liệucóhiệu quả cao trong việc đảm bảo chất lượng cũng như chi phí tính toán.

1.5.4.Các phương pháp tiếp cận của bài toán phân cụm dữ liệu

Các phương pháp phân cụm dữ liệu được phân thành các nhóm: phương pháppphân hoạch (**partitioning**), phương pháp phân cấp (**hierarchical**), phương pháp dựatrên mật độ (**density-based**), phương pháp dựa trên lưới (**grid-based**). Trong khuônkhổ đề tài này chỉ giới thiệu một vài thuật toán đại diện cho mỗi phương pháp.

a. Phương pháp phân hoạch(Partitioning Methods)

Thuật toán phân hoạch là một thuật toán phân cụm có từ rất lâu và khá phổ biến trước khi xuất hiện lĩnh vực khai phá dữ liệu[1]. Phân cụm không thứ bậc hoặc phân cụm theo phân hoạch (non-hierarchy or partition clustering) chia cơ sở dữ liệu bằng cách xác định trước các đối tượng đại diện (đối tượng nhân) của các cụm. Kế tiếp mỗi đối tượng dữ liệu sẽ được đưa vào cụm mà khoảng cách từ đối tượng dữ liệu đến đối tượng đại diện của cụm là nhỏ nhất. Sau mỗi bước thì đối tượng đại diện của mỗi cụm có thể được xác định lại dựa vào các đối tượng dữ liệu thuộc cụm đó. Mặc dù biểu diễn các cụm dữ liệu khác nhau, tuy nhiên các thuật toán đều có cách tiếp cận chung khi tính toán các giải pháp.

Ý tưởng của phương pháp phân hoạch như sau:

Cho tập D gồm n đối tượng, và một tham số đầu vào k được xác định bởingười dùng. Thuật toán phân hoạch sẽ **chọn** k **đối tượng đại diện cho** k **cụm** (k đốitượng đại diện có thể được chọn ngẫu nhiên hoặc theo một tiêu chuẩn của người sửdụng). Với một đối tượng dữ liệu q sẽ được đưa vào cụm có đối tượng đại diện gầnvới q nhất. Sau đó, đối tượng đại diện của mỗi cụm sẽ được tính lại dựa vào nhữngđiểm dữ liệu thuộc cụm đó. Thông thường thì đối tượng đại diện được xác định saocho khoảng cách từ đối tượng đại diện đến điểm xa nhất là nhỏ nhất có thể được.

Hình dưới mô tả quá trình phân hoạch với k=3.Khởi tạo bởi hình A với 3 đốitượng đại diện là 3 điểm đậm được lựa chọn ngẫu nhiên. Kế tiếp mỗi đối tượng dữliệu được đưa vào cụm mà khoảng cách từ điểm đó tới đối tượng đại diện của cụmlà nhỏ nhất. Với mỗi cụm tìm đối tượng đại diện cho cụm đó (lấy đối tượng dữ liệumới là điểm trung bình của tất cả các đối tượng dữ liệu thuộc cụm). Quá trình trênđược lặp lại cho đến khi các đối tượng đại diện của tất cả các cụm là không thayđổi.

Mô hình thuật toán phân cụm phân hoạch

Đầu vào: Số cụm k và cơ sở dữ liệu D gồm n đối tượng.

Đầu ra: Tập các cụm. Partition(D, k);

1. Chọn ngẫu nhiên k tâm bất kỳ O^0. Đặt i = 0.

2.Với mỗi điểm dữ liệu p $\in$D thì tìm đối tượng đại diện gần nhất và đưa p vàocụm đó.

3.Tính lại đối tượng đại diện của các cụm O^{i+1} dựa vào các điểm dữ liệu thuộccụm.

4.Nếu $O^i+1 = O^i$ thì dừng lại. Trong trường hợp ngược lại i = i+1 và quay lại 2, $O^i = \{o_1^{(i)}, o_2^{(i)},\ldots, o_k^{(i)}\}$ là tập các đối tượng đại diện của k cụm.

Với phương pháp này, số cụm được thiết lập là đặc trưng được lựa chọn trước. Phương pháp phân hoạch thích hợp với bài toán tìm các cụm trong không gian 2D. Ngoài ra, phương pháp xem xét đến khoảng cách cơ bản giữa các điểm dữ liệu đểxác định chúng có quan hệ gần nhau, hoặc không gần nhau hay không có quan hệ.

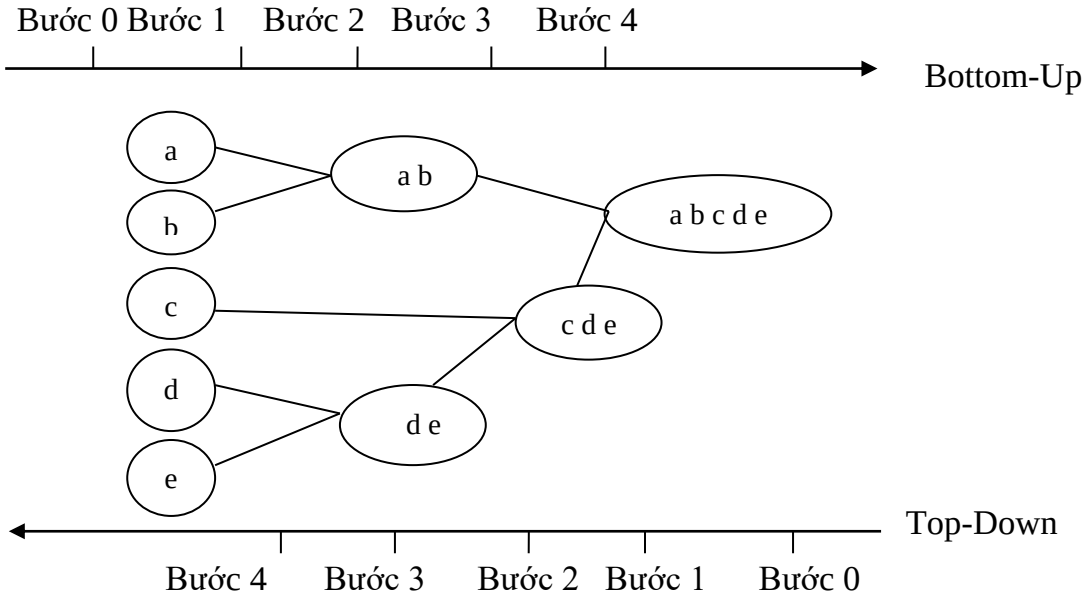
Nhược điểm của phương pháp này là đòi hỏi phải đưa vào tham số k và khôngxử lý trên bộ dữ liệu thuộc cụm có hình dạng phức tạp hoặc mật độ phân bố dàyđặc. Thêm vào đó, thuật toán có độ phức tạp tính toán lớn khi cần xác định kết quảtối ưu.

Các thuật toán trong phương pháp phân hoạch: k-means, PAM (**PartitioningAround Medoids**), CLARA (**Clustering LARge Application**), CLARANS(**Clustering Large Applications based upon RANdomized Search**), ...

b. Phương pháp phân cấp (Hierarchical Methods)

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệuđang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạnghình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Kỹ thuật này có 2 cách tiếp cận đó là:

- Tiếp cận hội tụ, thường được gọi là tiếp cận **Bottom – Up.**
- Tiếp cận phân chia nhóm,thường được gọi là tiếp cận **Top – Down.**



1) Tiếp cận **bottom-up:** bắt đầu với mỗi đối tượng thành lập một cụm riêngbiệt. Sau đó tiến hành hợp hoặc nhóm các đối tượng theo một vài tiêu chí đó nhưkhoảng cách giữa trung tâm của 2 nhóm. Thuật toán kết thúc khi tất cả các nhómđược hợp thành một nhóm (nút gốc của cây) hoặc thỏa mãn điều kiện dừng. Từ cây mới tạo được, đưa ra các cụm bằng cách chọn tập các đối tượng tại cácnút thoả mãn điều kiện dừng.

2) Tiếp cận **top-down**: Xuất phát từ gốc là một cụm với tất cả các đối tượngtrong một cơ sở dữ liệu. Tại mỗi bước lặp thì cụm được phân chia thành cụm nhỏhơn theo tiêu chí nào đó. Việc phân chia dừng khi mỗi đối tượng là một cụm hoặcthỏa mãn điều kiện dừng (kết thúc). Điều kiện kết thúc là điều kiện để xác định mộttập các đối tượng tại mỗi nút có phải là một cụm hay không. Điều kiện kết thúcđược đưa vào từ người sử dụng.

Ưu điểm của phương pháp này là kết hợplinh hoạt vào mức độ chi tiết, dễdàng xử lý với bất kỳ kiểu đo độ tương tựkhoảng cách nào, thích hợp với mọi kiểudữ liệu thuộc tính.

Nhược điểm là điều kiện để dừng vòng lặp rất mơ hồ, không cụ thể. Mặt khác,phương pháp không duyệt lại các mức trước khi xây dựng để cải tiến chất lượng cáccụm. Phương pháp này gồm có các thuật toán: **AGNES** (**Agglomerative NEsting**) và**DIANA** (**DIvisia ANAlysic**), CURE (**Clustering Using Representatives**), BIRCH(**Balance Iterative Reducing and Clustering using Hierarchies**), CHAMELEON …

c. Phương pháp dựa trên mật độ (Density-Based Methods)

Phương pháp dựa trên mật độ phân cụm các đối tượng dữ liệu dựa trên mốiquan hệ của các đối tượng dữ liệu với các điểm lân cận của các điểm dữ liệu đó. Phân cụm dựa trên mật độ (có điều kiện cụm cục bộ) giống như các điểm có khảnăng liên kết theo mật độ (**density-connected**). Một cụm được mở rộng theo hướngbất kỳ mà mật độ dẫn theo, do đó phương pháp này có khả năng tìm ra các cụm cóhình dạng phức tạp. Mặc dù chỉ duyệt tập dữ liệu một lần nhưng phương pháp nàycó khả năng loại bỏ phần tử nhiễu và phần tử ngoại lai. Phương pháp này phù hợp với các đối tượng có trường dữ liệu kiểu số, dữ liệu thuộc tính chỉ là thuộc tính môtả thêm cho các đối tượng không gian.

Phương pháp này có thể tiếp cận theo 2 hướng chính: liên kết dựa trên **mật độvà hàm mật độ**.

Các thuật toán thuộc phương pháp này bao gồm DBSCAN (**Density BasedSpatial Clustering of Application with Noise**), OPTICS (**Ordering Points to Identifythe Clustering Structure**), DENCLUE (**Density-based CLUstEring**), DBCLASD(**Distribution Based Clustering of Large Spatial Databased**)

d. Kết luận

Mỗi một phương pháp phân cụm đều có điểm mạnh điểm yếu và thích hợpcho từng ứng dụng cụ thể.

Phương pháp phân hoạch

Phương pháp phân hoạch đơn giản, dễ áp dụng và hiệu quả đối với cơsở dữ liệu nhỏ với các cụm đưa ra có hình dạng lồi. Tuy nhiên, do các cụmtrong phương pháp phân hoạch được biểu diễn bởi các tâm của cụm và mỗimột điểm dữ liệu được chia vào một cụm dựa vào khoảng cách từ điểm đótới tâm của cụm. Chính vì thế phương pháp phân hoạch chỉ có thể đưa rađược các cụm có hình dạng là đa giác lồi mà không thể đưa ra được các cụmcó dạng lõm phủ lên nhau hoặc lồng nhau. Ngoài ra, nếu cơ sở dữ liệu cónhiễu hoặc có đối tượng dữ liệu quá xa tâm thì phương pháp phâncụm phân hoạch cùng không áp dụng được vì trong các trường hợp đó, cácđối tượng dữ liệu nhiễu hoặc các đối tượng dữ liệu xa tâm sẽ làmtâm của cụm bị lệch đi. Do đó, không đưa ra được các cụm chính xác.

Phương pháp phân cấp

Thực hiện việc phân cụm bằng cách tách hoặc ghépcác nhóm đối tượng dựa vào độ tương đồng của các nhóm đối tượng đó. Phương pháp này không có khả năng phân cụm với hình dạng bất kỳ. Việcxây dựng lên cây cấu trúc tương đối phức tạp và phải duyệt cơ sở dữ liệunhiều lần, dẫn tới thời gian chạy của các thuật toán lớn. Ngoài ra, phươngpháp phân cấp đòi hỏi một không gian bộ nhớ để lưu giữ cây trong quá trìnhxây dựng. Do đó phương pháp này cũng không thích hợp với cơ sở dữ liệulớn.

Phương pháp dựa trên mật độ

Là một trong những phương pháp phân cụmhiệu quả, đặc biệt là cho những cơ sở dữ liệu không gian. Những thuật toánthuộc phương pháp này không chỉ có khả năng tìm ra những cụm với hìnhdáng bất kỳ mà còn rất hiệu qủa khi áp dụng lên những cơ sở dữ liệu cónhiễu.

Mỗi phương pháp phân cụm có ưu và nhược điểm riêng, phù hợp với từngloại ứng dụng. Tuỳ từng ứng dụng cụ thể mà người sử dụng sẽ chọn phươngpháp phân cụm thích hợp nhất cho yêu cầu ứng dụng của mình. Trong một sốtrường hợp, người ta có thể kết hợp các phương pháp khác nhau để đưa ramột thuật toán phân cụm hiệu quả hơn hoặc thích hợp hơn cho ứng dụng cụthể.

Tuy nhiên, những phương pháp và thuật toán phân cụm hiện tại đều khôngthể áp dụng và phân cụm được cơ sở dữ liệu hỗn hợp gồm nhiều lớp đốitượng dữ liệu với tính chất khác nhau và chứa nhiều đối tượng dữ liệu hỗnhợp như cơ sở dữ liệu không gian.

Phương pháp dựa trên lưới (Gird-Based Methods).

Phương pháp này thích hợp với dữ liệu nhiều chiều, được áp dụng chủ yếucho lớp cơ sở dữ liệu không gian. Ví dụ như dữ liệu được biểu diễn dướidạng cấu trúc hình học của đối tượng trong không gian cùng với các quan hệ,các thuộc tính, các hoạt động của chúng.

Ưu điểm của phương pháp dựa trên lưới là thời gian xử lý nhanh và độc lậpvới số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụthuộc vào số ô trong mỗi chiều của không gian lưới.

2.Thuật toán phân cụm dữ liệu dựa vào phân hoạch

2.1. Thuật toán K-Means

K-Means là thuật toán phân cụm mà định nghĩa các cụm bởi trung tâmcủa các phương tử. Phương pháp này dựa trên độ đo khoảng cách của các đốitượng dữ liệu trong cụm. Nó được xem như là trung tâm của cụm. Như vậy, nó cần khởi tạo một tập trung tâm các trung tâm cụm ban đầu, và thông quađó nó lặp lại các bước gồm gán mỗi đối tượng tới cụm mà trung tầm gần, vàtính toán tại trung tâm của mỗi cụm trên cơ sở gán mới cho các đối tượng. Quá trình này dừng khi các trung tâm hội tụ.

Các bước thực hiện thuật toán K-Means.

Bước 1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

Bước 2.Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean).

Bước 3. Nhóm các đối tượng vào nhóm gần nhất.

Bước 4.Xác định lại tâm mới cho các nhóm.

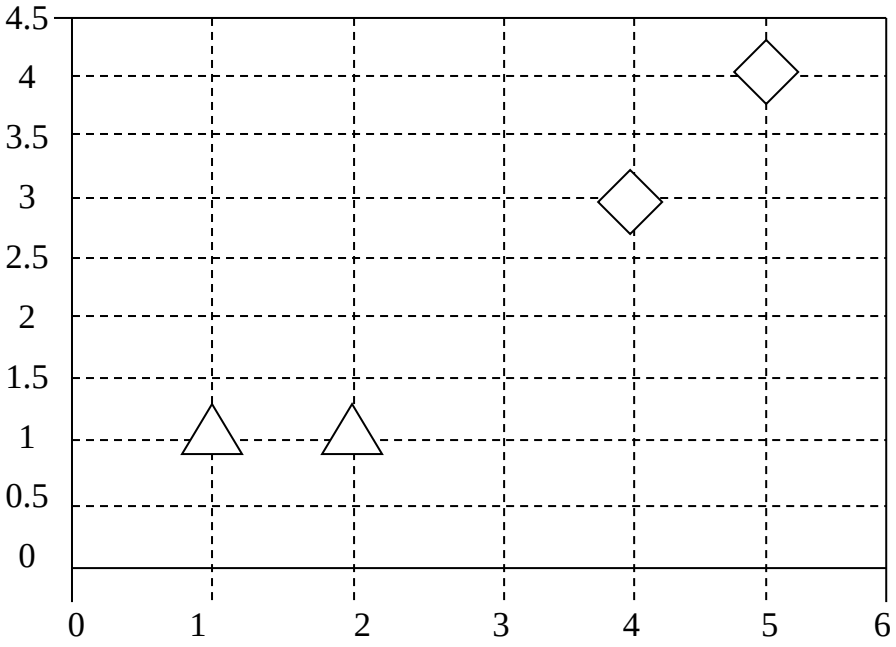
Bước 5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.

Ví dụ minh họa thuật toán K-Mean:

Giả sử ta có 4 loại thuốc A, B, C, D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng.

Object	Feature1(X): weight index	Feature 2(Y): pH
Medicine A	1	1
Medicine B	1	1
Medicine C	4	3
Medicine D	5	4

Bước 1. Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất c1(1, 1)) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai c2 (2, 1)).



Hình 3: Kmeans khởi tạo

Bước 2.Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean)

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c1 = (1,1) group - 1 \\ c2 = (1,2) group - 2 \end{matrix}$$
$$A \quad B \quad C \quad D$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Mỗi cột trong ma trận khoảng cách (D) là một đối tượng (cột thứ nhất tương ứng với đối tượng A, cột thứ 2 tương ứng với đối tượng B, …). Hàng thứ nhất trong ma trận khoảng cách biểu diễn khoảng cách giữa các đối tượng đến tâm của nhóm thứ nhất (c1) và hàng thứ 2 trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ 2 (c2).

Ví dụ, khoảng cách từ loại thuốc C=(4, 3) đến tâm c1(1, 1) là 3.61và đến tâm c2(2, 1) là 2.83 được tính như sau:

$$c_1 = (1,1)\sqrt{(4-1)^2.(3-1)^2} = 3.61$$

$$c_2 = (2,1)\sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

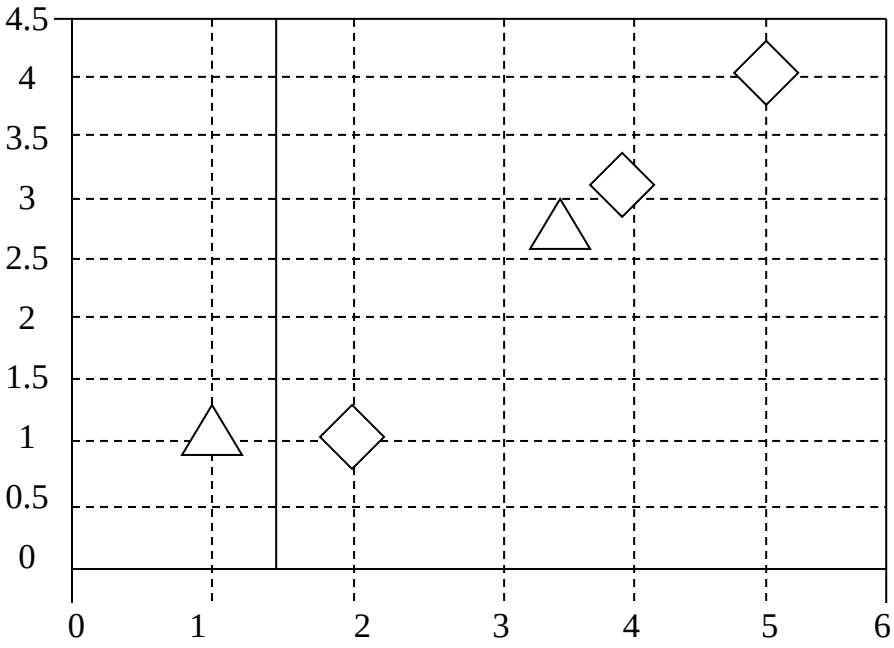
Bước 3. Nhóm các đối tượng vào nhóm gần nhất

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$

Ta thấy rằngnhóm 1 sau vòng lặp thứ nhất gồm có 1 đối tượng A và nhóm 2 gồm các đối tượng còn lại B, C, D.

Bước 4.Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, c1(1, 1). Tâm nhóm 2 được tính như sau:

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right)$$



Hình 4: Tính lại tọa độ

Bước 5. Tính lại khoảng cách từ các đối tượng đến tâm mới

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) group-1 \\ c_2 = \left(\frac{11}{3}, \frac{8}{3}\right) group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

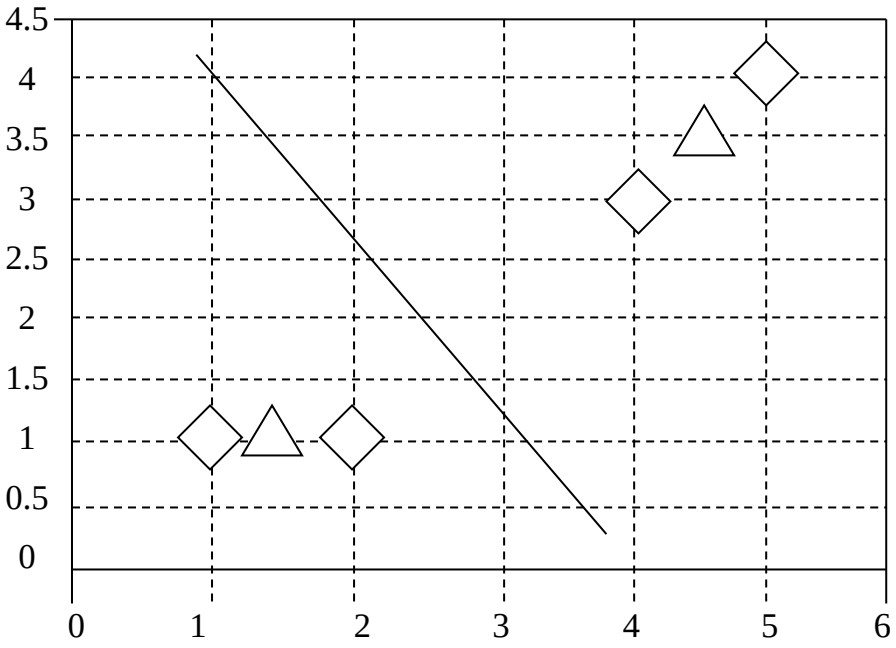
Bước 6. Nhóm các đối tượng vào nhóm

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Bước 7. Tính lại tâm cho nhóm mới

$$c_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (\frac{3}{2}, 1) c_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (\frac{9}{2}, \frac{7}{2})$$



Hình 5: Tính lại khoảng cách

Bước 8. Tính lại khoảng cách từ các đối tượng đến tâm mới

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.2 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{matrix} c_1 = (\frac{3}{2}, 1) group - 1 \\ c_2 = (\frac{9}{2}, \frac{7}{2}) group \quad 2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Bước 9. Nhóm các đối tượng vào nhóm

$$G^2 = \begin{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{matrix} group-1 \\ group-2 \end{matrix} \\ A \quad B \quad C \quad D & \end{matrix}$$

Ta thấy $G^2 = G^1$ (Không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng và kết quả phân nhóm như sau:

Object	Feature 1 (X) : weight index	Feature 2 (Y) : pH	Group(result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Thuật toán K-Means có **ưu điểm** là đơn giản, dễ hiểu và cài đặt. Tuy nhiên, một số **hạn chế** của K-Means là hiệu quả của thuật toán phụ thuộc vào việc chọn số nhóm K (phải xác định trước) và chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.

2.2. Thuật toán K-Medoids(hoặc PAM)

Mỗi cụm được biểu diễn bởi một điểm đối tượng thuộc cụm đó. Đây là giảipháp đơn giản vì phù hợp với mọi kiểu thuộc tính. Khi một đối tượng được chọnlàm trọng tâm của cụm, cụm được định nghĩa là tập con các điểm gần điểm trọngtâm đó. Mục tiêu đặt ra là tính khoảng cách trung bình hoặc sử dụng hàm tính độtương tự bất kỳ giữa các đối tượng và trọng tâm của nó.

Các bước trong thuật toán **K-Medoids** gần giống như thuật toán **K-Means**, trongđó giá trị **k** chính là **k** đối tượng được chọn ngẫu nhiên làm trọng tâm cụm. Phiênbản điển hình cho **k-medoids** là thuật toán PAM (**Partitioning Around Medoids**)gồm các bước như sau:

Phương pháp:

B1: Lấy ngẫu nhiên k đối tượng tùy ý làm trọng tâm của k cụm (**n>k**)

B2: Lặp các bước

B2.1. Gán các đối tượng vào cụm mà có độ tương tự gần với trọngtâm của cụm đó

B2.2. Chọn ngẫu nhiên đối tượng **O'** thuộc **n-k**

B2.3.Tính tổng chi phí S để chuyển từ điểm trọng tâm cũ sang **O'**

B2.4.Nếu S<0 thì chuyển điểm trọng tâm sang **O'**

B3. Thuật toán dừng khi tập các đối tượng **k** không thay đổi.

Tại bước 2.1, để tính độ đo tương tự có thể dùng khoảng cách **Euclidean**,**Manhattan** hay **Minkowski**. Thuật toán này chú tâm đến việc tìm cách thay thế cácđối tượng trọng tâm ban đầu bằng **n-k** đối tượng còn lại. Nếu không có sự thay thếxảy ra, thuật toán dừng. Bước 2.3 tính độ lệch **E** giữa trọng tâm cụm với đối tượngthuộc cụm đó. Do vậy, thuật toán này thực hiện việc tính E là **n-k** lần tương ứng vớiviệc so sánh với **n-k** điểm, nên thuật toán thực hiện tốn thời gian nếu số đối tượngcần phân cụm lớn.

Để cải tiến **nhược điểm** của thuật toán PAM, thuật toán CLARA (**ClusteringLARge Application**) ra đời vào năm 1990 bởi Kaufman và Rousseeuw. Thay vìthực hiện so sánh với **n-k** đối tượng còn lại, CLARA chỉ thực hiện trên một phần dữliệu mẫu được chọn từ tập dữ liệu ban đầu. CLARA chọn từng nhóm đối tượng, saudó dùng thuật toán PAM trên nhóm đối tượng đó. Kết quả trả về là các cụm tốt nhất.**Nhược điểm của CLARA** nếu mẫu được chọn không chứa **k** điểm trọng tâm tốt nhấtthì CLARA không đưa ra cách phân cụm tốt nhất. Năm 1994, Raymond T. Ng vàJiawei H. giới thiệu thuật toán CLARANS và đến năm 2002 thuật toán này đượccông bố là một phương pháp gom cụm hiệu quả trên cơ sở dữ liệu lớn, cơ sở dữ liệukhông gian.[1]

2.3. Thuật toán CLARA

CLARA tiến hành trích mẫucho tập dữ liệu có n phần tử, nó áp dụng thuật toán PAM cho mẫu này và tìmra các đối tượng trung tâm medoid cho mẫu được trích ra từ dữ liệu này. Nếumẫu dữ liệu được trích theo một cách ngẫu nhiên, thì các medoid của nó xấpxỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốthơn, CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗitrường hợp, sau đó tiến hành chọn kết quả phân cụm tốt nhất khi thực hiệnphân cụm trên mẫu này. Để đo chính xác, chất lượng của các cụm được đánhgiá thông qua độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệutrong tập đối tượng dữ liệu ban đầu.

Các bước thực hiện bài toán.

Bước 1. Chạy một vòng lặp: For i = 1 to 5 do

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

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

Bước 4.Tính độ phi tương tự trung bình cho phân hoạch các đối tượng thuđược ở bước trước, nếu giá rị này bé hơn giá trị tối thiểu hiện thời thìsử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, nhuvậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đếnthời điểm này.

Bước 5. Quay về bước 1.

Bước 6. Chạy cho đến khi kết thúc.

2.4.Thuật toán CLARANS

CLARANS cũng sử dụng kiểu k-medoids, nó kết hợp thuật toán PAMvới chiến lược tìm kiếm kinh nghiệm mới. Ý tưởng cơ bản của CLARANS làkhông xem xét tất cả các khả năng có thể thay thế các đối tượng tâm medoids bởi một đối tượng khác, nó ngay lập tức thay thế các đối tượng tâm này nếuviệc thay thế này có tác động tốt đến chất lượng phân cụm chứ không cần xácđịnh cách thay thế tối ưu nhất.

CLARANS lấy ngẫu nhiên một đối tượng của k đối tượng medoidtrong tâm cụm và cố gắng thay thế nó với một đối tượng chọn ngẫu nhiêntrong (n-k) đối tượng còn lại. Cụm thu được sau khi thay thế đối tượng trungtâm được gọi là một láng giềng của phân hoạch cụm trước đó. Số các lánggiềng được hạn chế bởi tham số do người dùng đưa vào là Maxneighbor, quátrình lựa chọn các láng giềng này hoàn toàn ngẫu nhiên. Tham số Numlocalcho phép người dùng xác định số vòng lặp tối ưu cục bộ được tìm kiếm. Không phải tất cả các láng giếng được duyệt mà chỉ có Maxneighbor số lánggiềng được duyệt. Nếu một láng giềng tốt hơn được tìm thấy, thì CLARANSdi chuyển láng giềng đó tới nút và quá trình bắt đầu lặp lại; nếu không kếtquả cụm hiện thời là tối ưu cục bộ. Nếu tối ưu cục bộ được tìm thấy, thìCLARANS bắt đầu với lựa chọn nút ngẫu nhiên mới trong tìm kiếm tối ưucục bộ mới.

CLARANS không thích hợp với tập dữ liệu lớn bởi vì nó lấy phần nhỏcủa toàn bộ tập dữ liệu và phần này được chọn để đại diện toàn bộ tập dữ liệuvà thực hiện sau đó. CLARANS không bị giới hạn không gian tìm kiếm nhưđối với CLARA, và trong cùng một lượng thời gian thì chất lượng của cáccụm phân được là lớn hơn CLARA.

Một số khái niệm sử dụng trong thuật toán CLARANS được định nghĩanhư sau:

Giả sử O là một tập có n đối tượng và **MO** ⊆là tập các đối tượng tâmmediod, NM = O- M là tập các đố tượng không phải tâm. Các đối tượng dữliệu sử dụng trong thuật toán CLARANS là các khối đa diện. Mỗi đối tượngđược diễn tả bằng một tập các cạnh, mỗi cạnh được xác định bằng hai điểm. Giả sử **PR** ⊆3 là một tập tất cả các điểm. Nói chung, các đối tượng ở đây làcác đối tượng dữ liệu không gian và chúng ta định nghĩa tâm của một đốitượng chính là trung bình cộng toán học của tất cả các đỉnh hay còn gọi làtrọng tâm: **center O→ P:**

Giả sử dist là một hàm khoảng cách, khoảng cách thường được chọn ởđây là khoảng cách Eucliean: **dist PxP →R^+0**hàm khoảng cách dist có thể mở rộng cho các điểm của khối đa diệnthông qua hàm tâm:

dist OxO →R^+0sao cho**disdt(o_I,o_j)=dist(center(o_i),center(o_j)).**

Mỗi đối tượng được gán cho một tâm medoid của cụm nếu khoảngcách từ trọng tâm của đối tượng đó tới tâm medoid của nó là nhỏ nhất. Vì vậy,định nghĩa tâm medoid như sau:

medoid: **O→M** sao cho**medoid(o)=m_i,m_j€M,Vm_i€M:dis(o,m_i)<=dist(o,m_j),o€O**.

Cuối cùng định nghĩamột cụm tới tâm mediod m_i tương ứng là một tập con các đối tượng trong Ovới medoid(o) = m_i. Giả sử C_o là tập tất cả các phân hoạch của O. Hàm tổng để đánh giáchất lượng một phân hoạch được định nghĩa như sau:

total_distance: **C_o→R^+_o**0 sao cho total_distance(c)=∑∑**dist**(o ,m_i) với **m_i€M,o€ cluster(m_i).**

Thuật toán chi tiết.

Input: **O,k, dist, numlocal và maxneighbor;'**

Output: k cụm dữ liệu;

CLARANS(int k, funcition dist,int numlocal,int maxneighbor)

BEGIN

For(i =1**;** i<= numlocalk;i++)

{

Current. creat_randomly(k);

j=1;

while (j<= maxneighbor)

{

Current. select_random(old,new);

Diff=Current. caculate_distance_difference(old,new);

If(Diff<0)

{

Current. Exchange(old,new);

j=1;

}

Else j++;

}

Dist=Current. caculate_total_distance();

If(dist<smallest_dist)

{

Best=Current;

Smallest_dist=dist;}

}

END

CHƯƠNG 3: THỬ NGHIỆM HỆ THỐNG

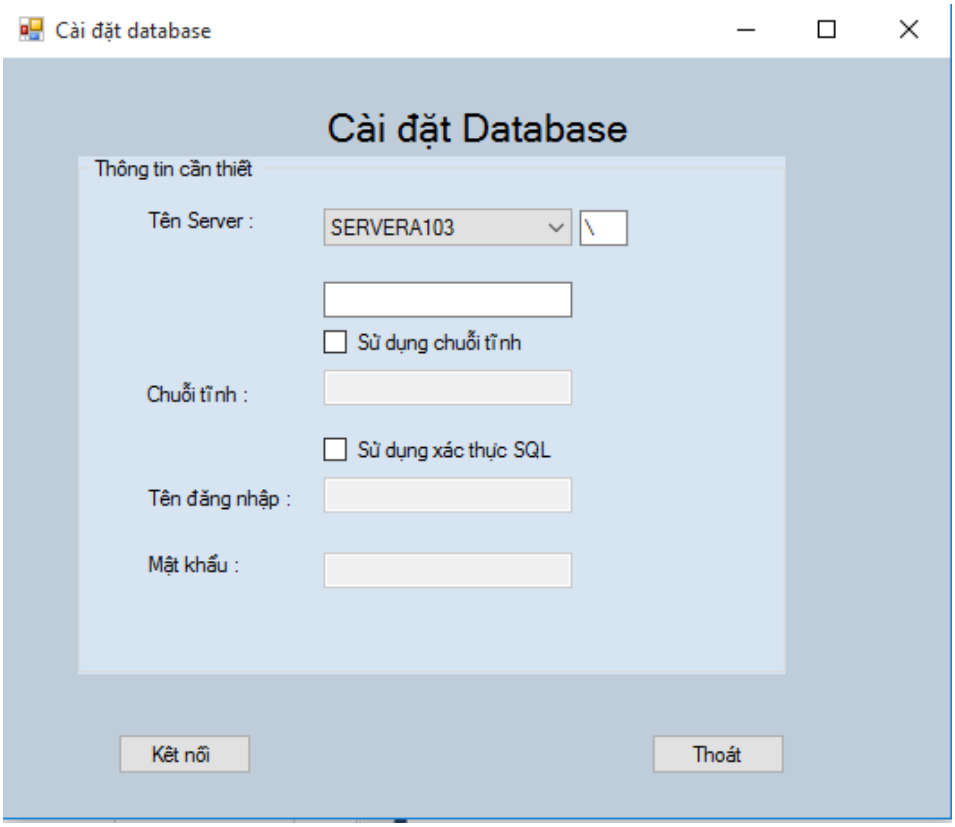
1. Phần mềm quản lý dữ liệu

Bài toán: Cần có một trung tâm quản lý dữ liệu đăng ký IP, Domain và số lượng truy cập. Xây dựng hệ thông quản lý,tìm kiếm và báo cáo sao cho đáp ứng nhu cầu của người quản lý đưa ra.

Công cụ sử dụng: Visual Studio 2010 và SQL Manager 2008(R2).

2.Các chức năng của chương trình

2.1. Thiết lập kết nối cơ sở dữ liệu



Hình 6: Kết nối cơ sở dữ liệu

Code: Nút kết nối

```
privatevoid button1_Click(object sender, EventArgs e)
        {
string strConnect = "";
if (checkBox_sudungchuoitinh. Checked==false&&checkBox_sql. Checked==false)
            {
```

```
                strConnect="DataSource="+comboBox1.Text+""+textBox4.Text+""+te
            xtBox_tendatabase.
            Text+";Database=Datamining;Trusted_Connection=True";
                }
             elseif (checkBox_sudungchuoitinh. Checked == false&& checkBox_sql.
          Checked == true)
                {
                    strConnect = "Data Source=" + comboBox1. Text + "" + textBox4.Text
+ "" + textBox_tendatabase. Text + ";Initial Catalog=Datamining;User Id=" +
textBox_usernam. Text + ";Password=" + textBox_password. Text + "";
                }
        if (checkBox_sql. Checked == false&&checkBox_sudungchuoitinh.
Checked==true)
                {
                    strConnect = "Data Source=" + textBox_chuoitnh. Text +
";Database=Datamining;Trusted_Connection=True";
        }
        else
                {
                    strConnect = "Data Source=" + textBox_chuoitnh. Text + ";Initial
Catalog=Datamining;User Id=" + textBox_usernam. Text + ";Password=" +
textBox_password. Text + "";
                }
        SqlConnection sqlcon = newSqlConnection(strConnect);
        try
                {
                    sqlcon. Open();
                    Datamining. Properties. Settings. Default. strConnect = strConnect;
                    Datamining. Properties. Settings. Default. Save();
        MessageBox. Show("Kết nối thành công ,vui lòng thử lại chương trình !!)");
        FileInfo t = newFileInfo("Config. txt");
        StreamWriter Tex = t. CreateText();
                    Tex. WriteLine("Thông tin cài đặt !!!");
                    Tex. WriteLine("Chuỗi kết nối: ");
                    Tex. WriteLine(strConnect);
                    Tex. Write(Tex. NewLine);
```

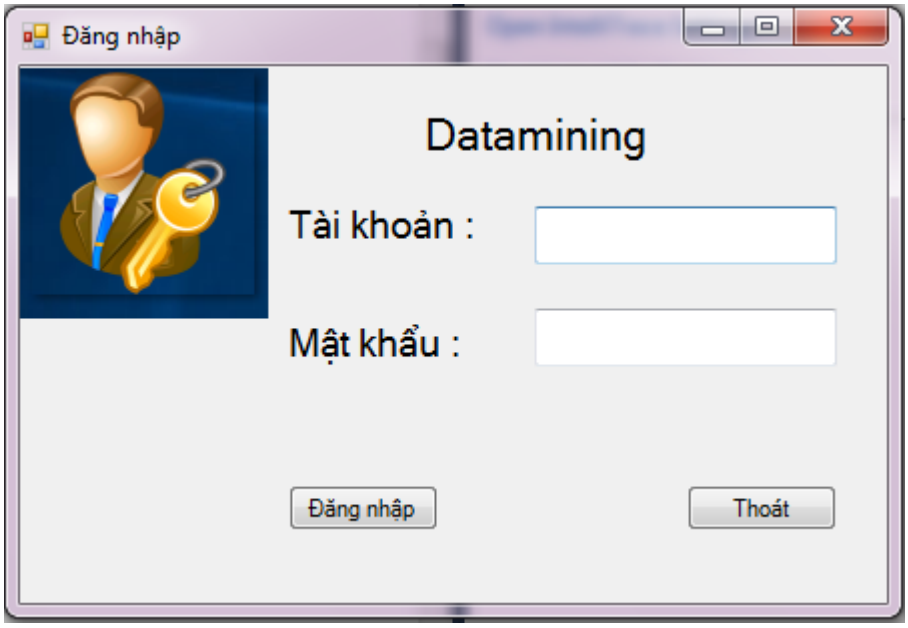
```
                Tex. Close();
Console. Beep();
Application. Exit();
            }
catch
            {
MessageBox. Show("Kết nối thất bại !!");
Application. Exit();
            }
        }
```

Code nút thoát:

```
privatevoid button_thoat_Click(object sender, EventArgs e)
{
Application. Exit();
}
```

2.2. Giao diện người dùng

2.2.1. Đăng nhập



Hình 7: Giao diện đăng nhập

Code nút "Đăng nhập"

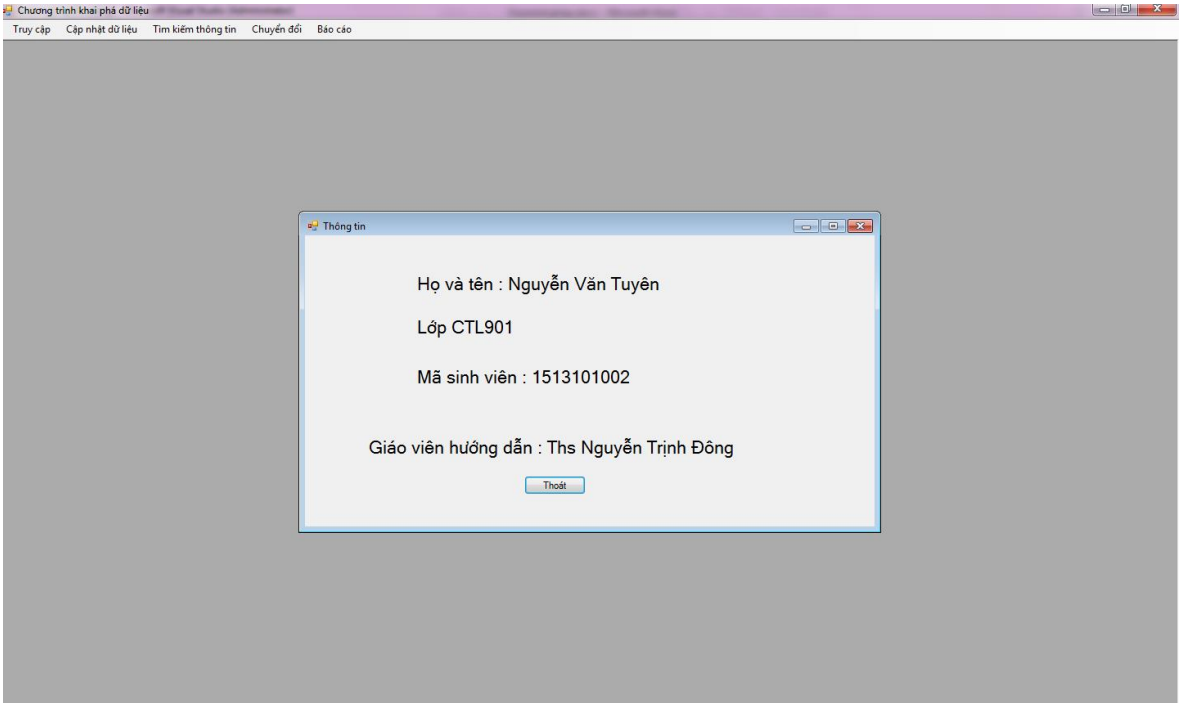
```
privatevoid button1_Click(object sender, EventArgs e)
        {
if (textBox1.Text == null&& textBox2.Text == null)
            {
MessageBox.Show("Bạn chưa nhập tài khoản hoặc mật khẩu !!! ", "Thông báo", MessageBoxButtons.OK);
                textBox1.Focus();
            }
else
            {
SqlConnection sqlcon = newSqlConnection(Datamining.Properties.Settings.Default.strConnect);
SqlDataAdapter sda = newSqlDataAdapter("Select Count(*) From Dangnhap where Ten='" + textBox1.Text + "' and Matkhau='" + textBox2.Text + "'", sqlcon);
DataTable dt = newDataTable();
```

```
            sda.Fill(dt);
        if (dt.Rows[0][0].ToString() == "1")
            {
        this.Hide();
        Form1 f = newForm1();
              f.Show();
            }
        else
            {
        MessageBox.Show("Bạn hãy kiểm tra lại tài khoản hoặc mật khẩu !!", "Thông báo", MessageBoxButtons.OK);
            }
          }
       }
```

Code nút “Thoát”

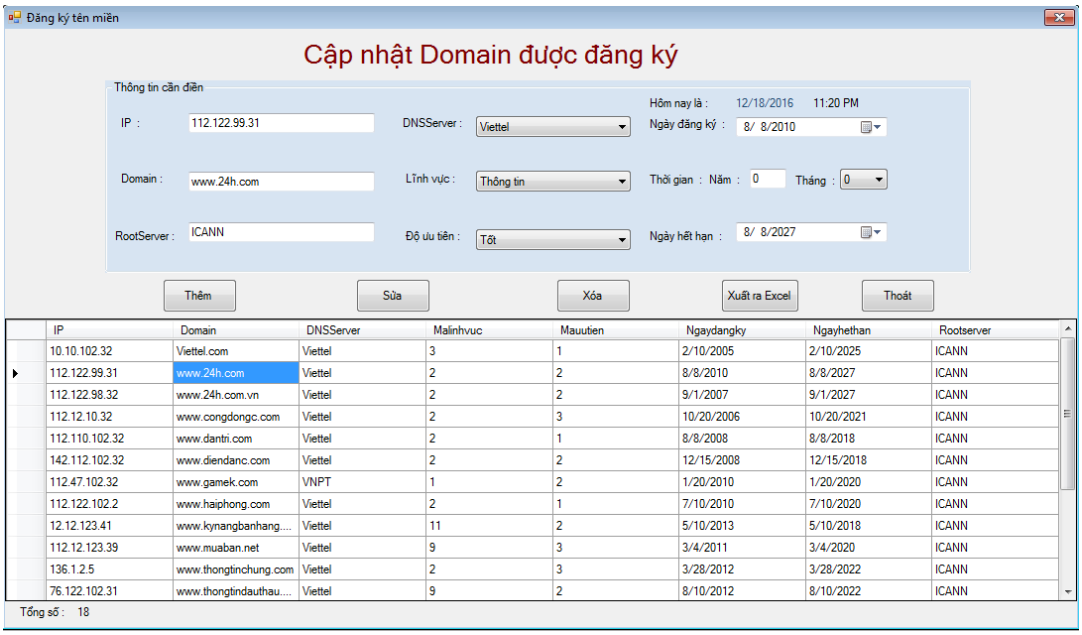
```
        privatevoid button2_Click(object sender, EventArgs e)
        {
        Application.Exit();
        }
```

2.2.2. Giao diện chính sau đăng nhập



Hình 8: Giao diện sau khi đăng nhập

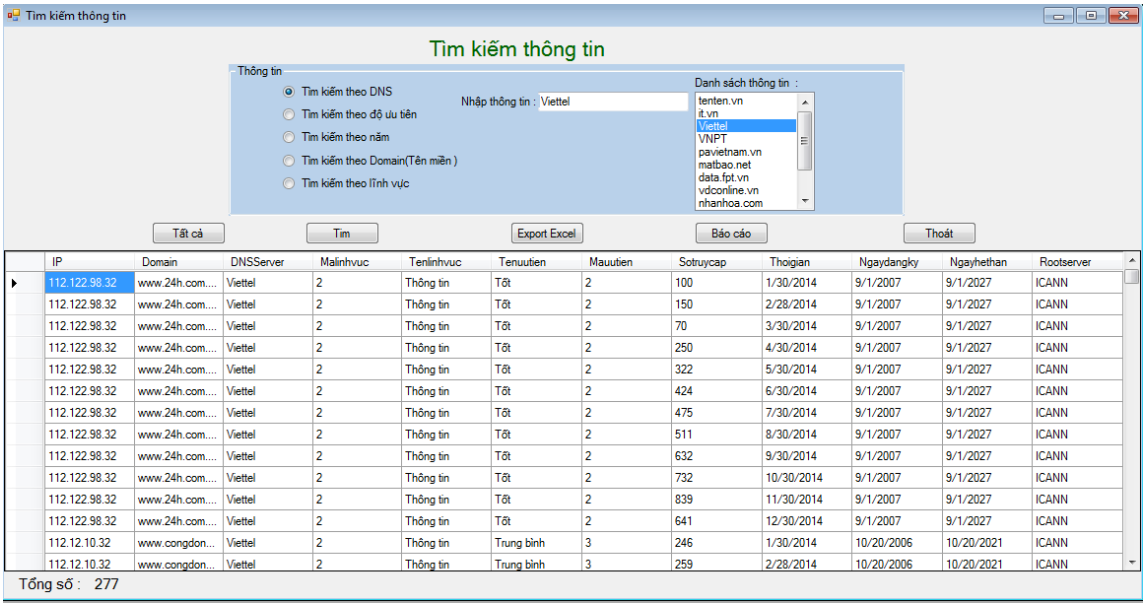
2.2.3.Cập nhật một bảng



IP	Domain	DNSServer	Malinhvuc	Mauutien	Ngaydangky	Ngayhethan	Rootserver
10.10.102.32	Viettel.com	Viettel	3	1	2/10/2005	2/10/2025	ICANN
112.122.99.31	www.24h.com	Viettel	2	2	8/8/2010	8/8/2027	ICANN
112.122.98.32	www.24h.com.vn	Viettel	2	2	9/1/2007	9/1/2027	ICANN
112.12.10.32	www.congdongc.com	Viettel	2	3	10/20/2006	10/20/2021	ICANN
112.110.102.32	www.dantri.com	Viettel	2	1	8/8/2008	8/8/2018	ICANN
142.112.102.32	www.diendanc.com	Viettel	2	2	12/15/2008	12/15/2018	ICANN
112.47.102.32	www.gamek.com	VNPT	1	2	1/20/2010	1/20/2020	ICANN
112.122.102.2	www.haiphong.com	Viettel	2	1	7/10/2010	7/10/2020	ICANN
12.12.123.41	www.kynangbanhang....	Viettel	11	2	5/10/2013	5/10/2018	ICANN
112.12.123.39	www.muaban.net	Viettel	9	3	3/4/2011	3/4/2020	ICANN
136.1.2.5	www.thongtinchung.com	Viettel	2	3	3/28/2012	3/28/2022	ICANN
76.122.102.31	www.thongtindauthau....	Viettel	9	2	8/10/2012	8/10/2022	ICANN

Hình 9: Cập nhật tên miền đăng ký

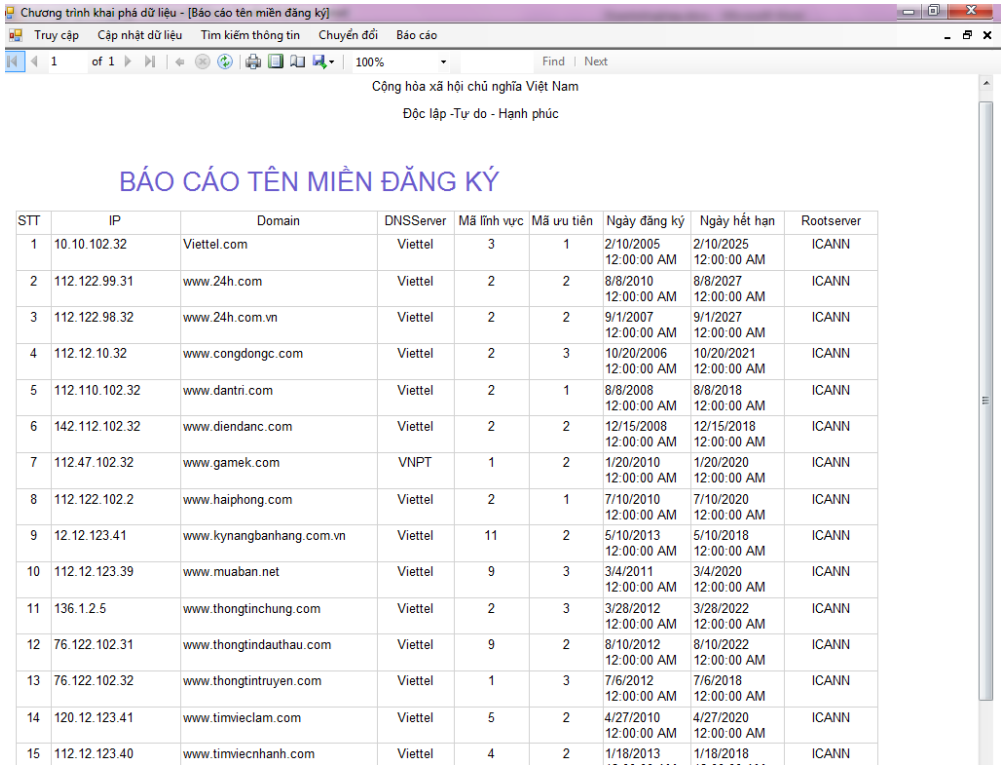
2.2.4. Tìm kiếm thông tin



IP	Domain	DNSServer	Malinhvuc	Tenlinhvuc	Tenuutien	Mauutien	Sotruycap	Thoigian	Ngaydangky	Ngayhethan	Rootserver
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	100	1/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	150	2/28/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	70	3/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	250	4/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	322	5/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	424	6/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	475	7/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	511	8/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	632	9/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	732	10/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	839	11/30/2014	9/1/2007	9/1/2027	ICANN
112.122.98.32	www.24h.com....	Viettel	2	Thông tin	Tốt	2	641	12/30/2014	9/1/2007	9/1/2027	ICANN
112.12.10.32	www.congdon...	Viettel	2	Thông tin	Trung bình	3	246	1/30/2014	10/20/2006	10/20/2021	ICANN
112.12.10.32	www.congdon...	Viettel	2	Thông tin	Trung bình	3	259	2/28/2014	10/20/2006	10/20/2021	ICANN

Hình 10: Tìm kiếm thông tin

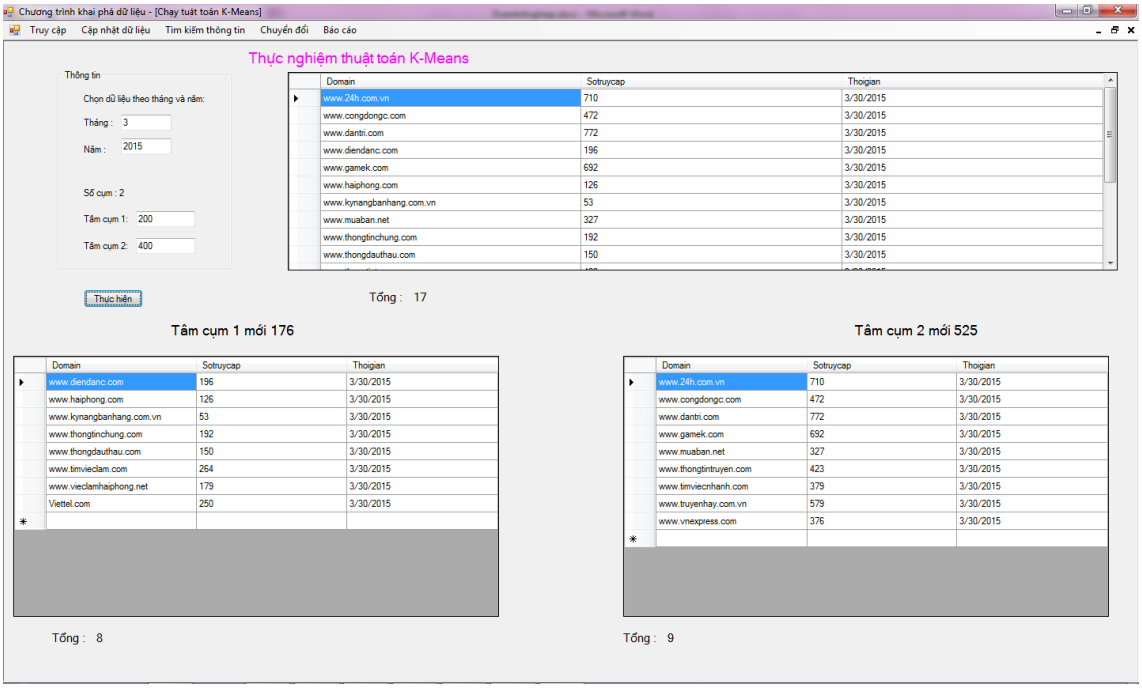
2.2.5. Báo cáo



STT	IP	Domain	DNSServer	Mã lĩnh vực	Mã ưu tiên	Ngày đăng ký	Ngày hết hạn	Rootserver
1	10.10.102.32	Viettel.com	Viettel	3	1	2/10/2005 12:00:00 AM	2/10/2025 12:00:00 AM	ICANN
2	112.122.99.31	www.24h.com	Viettel	2	2	8/8/2010 12:00:00 AM	8/8/2027 12:00:00 AM	ICANN
3	112.122.98.32	www.24h.com.vn	Viettel	2	2	9/1/2007 12:00:00 AM	9/1/2027 12:00:00 AM	ICANN
4	112.12.10.32	www.congdongc.com	Viettel	2	3	10/20/2006 12:00:00 AM	10/20/2021 12:00:00 AM	ICANN
5	112.110.102.32	www.dantri.com	Viettel	2	1	8/8/2008 12:00:00 AM	8/8/2018 12:00:00 AM	ICANN
6	142.112.102.32	www.diendanc.com	Viettel	2	2	12/15/2008 12:00:00 AM	12/15/2018 12:00:00 AM	ICANN
7	112.47.102.32	www.gamek.com	VNPT	1	2	1/20/2010 12:00:00 AM	1/20/2020 12:00:00 AM	ICANN
8	112.122.102.2	www.haiphong.com	Viettel	2	1	7/10/2010 12:00:00 AM	7/10/2020 12:00:00 AM	ICANN
9	12.12.123.41	www.kynangbanhang.com.vn	Viettel	11	2	5/10/2013 12:00:00 AM	5/10/2018 12:00:00 AM	ICANN
10	112.12.123.39	www.muaban.net	Viettel	9	3	3/4/2011 12:00:00 AM	3/4/2020 12:00:00 AM	ICANN
11	136.1.2.5	www.thongtinchung.com	Viettel	2	3	3/28/2012 12:00:00 AM	3/28/2022 12:00:00 AM	ICANN
12	76.122.102.31	www.thongtindauthau.com	Viettel	9	2	8/10/2012 12:00:00 AM	8/10/2022 12:00:00 AM	ICANN
13	76.122.102.32	www.thongtintruyen.com	Viettel	1	3	7/6/2012 12:00:00 AM	7/6/2018 12:00:00 AM	ICANN
14	120.12.123.41	www.timvieclam.com	Viettel	5	2	4/27/2010 12:00:00 AM	4/27/2020 12:00:00 AM	ICANN
15	112.12.123.40	www.timviecnhanh.com	Viettel	4	2	1/18/2013	1/18/2018	ICANN

Hình 11: Báo cáo

2.2.6. K-Means và K-Medoids(Hoặc PAM)



Hình 12: K-Means và K-Medoids

Code "Thực hiện"

```
privatevoid button1_Click(object sender, EventArgs e)
        {
try
            {
int thang = int.Parse(textBox_thang.Text);
int nam = int.Parse(textBox_năm.Text);
int cum1 = int.Parse(textBox_tamcum1.Text);
int cum2 = int.Parse(textBox_tamcum2.Text);
if (cum1 < cum2)
                {
//
//Tính tâm cụm 1
SqlConnection sql =
newSqlConnection(Datamining.Properties.Settings.Default.strConnect);
                    sql.Open();
```

```
        SqlCommand sqlCom = newSqlCommand();
                    sqlCom.CommandType = CommandType.Text;
                    sqlCom.CommandText = "SELECT
Trungbinh=(Sum(Sotruycap)/Count(Sotruycap)) FROM DBO.Chitiettruycap where
month(Thoigian)=" + thang + " and Year(Thoigian)=" + nam + "and (Sotruycap-" + cum1 +
")<=(" + cum2 + "-Sotruycap) ";
                    sqlCom.Connection = sql;
        DataTable dtble1 = newDataTable();
        SqlDataAdapter adt1 = newSqlDataAdapter(sqlCom);
                    adt1.Fill(dtble1);
        foreach (DataRow dr in dtble1.Rows)
                    {
                       label_cum1.Text = "Tâm cụm 1 mới " +
Convert.ToString(dr["Trungbinh"]);
                    }
                    sql.Close();
        // Xem cụm 1-----------------------------------
        DtKmeans dtk = newDtKmeans();
                    dtk.thang = thang;
                    dtk.nam = nam;
                    dtk.cum1 = cum1;
                    dtk.cum2 = cum2;
        ProviderKmeans prok = newProviderKmeans();
                    dataGridView2.DataSource = prok.xemcum1(dtk);
        //--------------------------------------------
        // Tính tâm cụm 2
                    sql.Open();
                    sqlCom.CommandType = CommandType.Text;
                    sqlCom.CommandText = "SELECT
Trungbinh=(Sum(Sotruycap)/Count(Sotruycap)) FROM DBO.Chitiettruycap where
```

```
month(Thoigian)=" + thang + " and Year(Thoigian)=" + nam + "and (Sotruycap-" + cum1 +
")>=(" + cum2 + "-Sotruycap) ";
                    sqlCom.Connection = sql;
        DataTable dtble2 = newDataTable();
        SqlDataAdapter adt2 = newSqlDataAdapter(sqlCom);
                    adt2.Fill(dtble2);
        foreach (DataRow dr in dtble2.Rows)
                    {
                       label_cum2.Text = "Tâm cụm 2 mới " +
Convert.ToString(dr["Trungbinh"]);
                    }
                    sql.Close();
        // Xem cụm 2----------------------------------
                    dataGridView3.DataSource = prok.xemcum2(dtk);
        //-------------------------------------------
        ////tổng số dữ liệu được lọc
        int sl;
                    sl = dataGridView1.Rows.Count - 1;
        ////tổng số dữ liệu cụm 1
                    sl = dataGridView2.Rows.Count - 1;
                    label13.Text = sl.ToString();
        ////tổng số dữ liệu cụm 2
                    sl = dataGridView3.Rows.Count - 1;
                    label15.Text = sl.ToString();
                }
        else
                {
        MessageBox.Show("Mời bạn nhập lại tâm cụm (Cụm 1< Cụm 2) !!!", "Thông báo",
MessageBoxButtons.OK, MessageBoxIcon.Warning);
```

```
            }
        }
    catch
        {
    MessageBox.Show("Hãy nhập tâm cụm trước khi thực hiện !!!", "Thông báo", MessageBoxButtons.OK, MessageBoxIcon.Warning);
        }
    }
```

KẾT LUẬN

Trong quá trình làm đồ án, đã giúp cho em học hỏi cũng như biết thêm một số kỹ năng trong việc phân tích, tìm hiểu, một vấn đề mới và cách xây dựng một đồ án hoàn chỉnh. Việc tìm hiểu nội dung đề tài đã cho em biết thêm về khai phá dữ liệu, từ những thông tin cơ bản cũng như ứng dụng của khai phá dữ liệu trong đời sống hàng ngày.

Đề tài của em đã giải quyết một số vấn đề sau :

- Giới thiệu chung về khai phá dữ liệu: Phạm vi, mục tiêu, các bước khai phá dữ liệu. Các ứng dụng trong thực tế, cũng như khó khăn thường gặp phải. Một số thông tin cơ bản về Internet và cách quản lý địa chỉ Internet.
- Các thuật toán trong khai phá dữ liệu. Đặc điểm chung, riêng của từng loại khai phá cách áp dụng trong thực tế.
- Vận dụng kiến thức đã tìm hiểu để xây dựng một chương trình nhỏ cho việc quả lý địa chỉ Internet và áp dụng thuật toán để khai phá dữ liệu.

Đề tài đã và đang được áp dụng trong cuộc sống về kinh tế (dữ liệu chứng khoán), thông tin (mọi người thường tìm gì trên Internet),....Hướng đi của đề tài trong tương lai, hoàn thiện hơn về lý thuyết cũng như xây dựng ứng dụng hoàn chỉnh để khai thác thông tin tốt hơn.

TÀI LIỆU THAM KHẢO

[1] Ng, Raymond T. and Jiawei Han. Efficient and Effective Clustering Methods for Spatial Data Mining. Proceedings of the 20th Very Large DatabasesConference (VLDB 94), Santiago, Chile. pp 144-155. (CLARAN)

[2] Milligan và Cooper, 1988Effect of Stamdardization procedure and alternative data structure

[3] Ding và He, 2004 k-Means Clustering via Principal Component Analysis

[4] Tìm hiểu trên Wiki

[5]Trung tâm Internet Việt Nam-https://www.vnnic.vn

[6] Klett – 1972Westslaven und westslavisches Erbe in Deutschland

[7] Fukunaga, 1990 Introduction to Statistical Pattern Recognition