ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Trung Sơn

PHƯƠNG PHÁP PHÂN CỤM VÀ ỨNG DỤNG

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số : 60.48.01

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
1. PGS. TS VŨ ĐỨC THI

Thái Nguyên – 2009

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Trung Sơn

PHƯƠNG PHÁP PHÂN CỤM VÀ ỨNG DỤNG

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số : 60.48.01

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
1. PGS. TS VŨ ĐỨC THI

Thái Nguyên – 2009

MỤC LỤC

LỜI CẢM ƠN

Em xin chân thành cảm ơn PGS. TS Vũ Đức Thi đã tận tình hướng dẫn khoa học, giúp đỡ em hoàn thành tốt luận văn tốt nghiệp này.

Em cũng xin gửi lời cảm ơn tới các thầy, cô giáo đã dạy dỗ, và truyền đạt kiến thức cho em trong suốt quá trình học tập và nghiên cứu

HỌC VIÊN

NGUYỄN TRUNG SƠN

LỜI MỞ ĐẦU

Trong những năm gần đây, sự phát triển mạnh mẽ của CNTT đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Bên cạnh đó, việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Hàng triệu CSDL đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lý..., trong đó có nhiều CSDL cực lớn cỡ Gigabyte, thậm chí là Terabyte.

Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền CNTT thế giới hiện nay nói chung và Việt Nam nói riêng. Khai phá dữ liệu đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet… Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất kinh doanh của mình và thu được những lợi ích to lớn.

Các kỹ thuật khai phá dữ liệu thường được chia thành 2 nhóm chính:

- Kỹ thuật khai phá dữ liệu mô tả: có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có.

- Kỹ thuật khai phá dữ liệu dự đoán: có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời.

Bản luận văn này trình bày một số vấn đề về Phân cụm dữ liệu, một trong những kỹ thuật cơ bản để Khai phá dữ liệu. Đây là hướng nghiên cứu có triển vọng chỉ ra những sơ lược trong việc hiểu và khai thác CSDL khổng lồ, khám phá thông tin hữu ích ẩn trong dữ liệu; hiểu được ý nghĩa thực tế của dữ liệu.

Luận văn được trình bày trong 3 chương và phần phụ lục :

Chương 1 : Trình bày tổng quan lý thuyết về Phân cụm dữ liệu, các kiểu dữ liệu, Phép biến đổi và chuẩn hóa dữ liệu.

Chương 2 : Giới thiệu, phân tích, đánh giá các thuật toán dùng để phân cụm dữ liệu

Chương 3 : Trình bày một số ứng dụng tiêu biểu của phân cụm dữ liệu.

Kết luận : Tóm tắt các vấn đề được tìm hiểu trong luận văn và các vấn đề liên quan trong luận văn, đưa ra phương hướng nghiên cứu tiếp theo.

CHƯƠNG I :
TỔNG QUAN LÝ THUYẾT VỀ PHÂN CỤM DỮ LIỆU

1. Phân cụm dữ liệu

1.1 Định nghĩa về phân cụm dữ liệu

Phân cụm dữ liệu(Data Clustering) hay phân cụm, cũng có thể gọi là phân tích cụm, phân tích phân đoạn, phân tích phân loại, là quá trình nhóm một tập các đối tượng thực thể hay trừu tượng thành lớp các đối tượng tương tự. Một cụm là một tập hợp các đối tượng dữ liệu mà các phần tử của nó tương tự nhau cùng trong một cụm và phi tương tự với các đối tượng trong các cụm khác. Một cụm các đối tượng dữ liệu có thể xem như là một nhóm trong nhiều ứng dụng.

1.2 Một số ví dụ về phân cụm dữ liệu

1.2.1 Phân cụm dữ liệu phục vụ cho biểu diễn dữ liệu gene

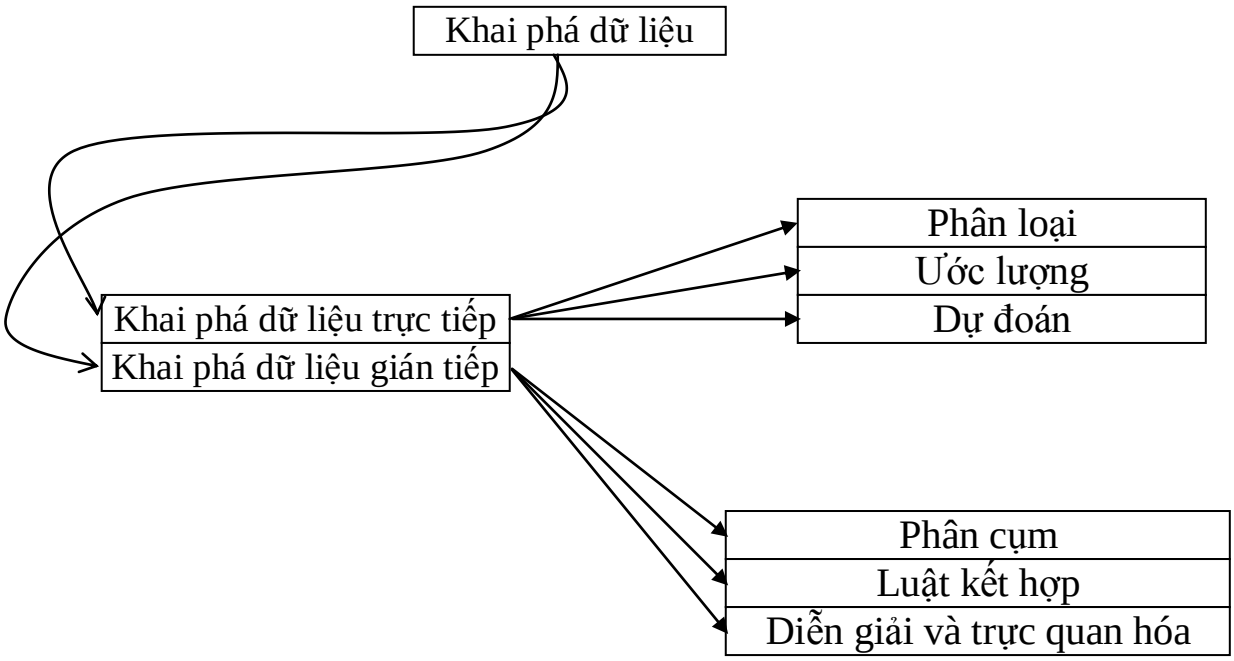
Phân cụm là một trong những phân tích được sử dụng thường xuyên nhất trong biểu diễn dữ liệu gene (Yeung et al., 2003; Eisen at al., 1998). Dữ liệu biểu diễn gene là một tập hợp các phép đo được lấy từ **DNA microarray** (còn gọi là **DNA chip** hay **gene chip**) là một tấm thủy tinh hoặc nhựa trên đó có gắn các đoạn DNA thành các hàng siêu nhỏ. Các nhà nghiên cứu sử dụng các con chip như vậy để sàng lọc các mẫu sinh học nhằm kiểm tra sự có mặt hàng loạt trình tự cùng một lúc. Các đoạn DNA gắn trên chip được gọi là probe (mẫu dò). Trên mỗi điểm của chip có hàng ngàn phân tử probe với trình tự giống nhau. Một tập hợp dữ liệu biểu diễn gene có thể được biểu diễn thành một ma trận giá trị thực :

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix},$$

Trong đó :

- n là số lượng các gen
- d là số lượng mẫu hay điều kiện thử
- x_{ij} là thước đo biểu diễn mức gen i trong mẫu j

Bởi vì các biểu ma trận gốc chứa nhiễu, giá trị sai lệch, hệ thống biến thể, do đó tiền xử lý là đòi hỏi cần thiết trước khi thực hiện phân cụm.



Hình 1 Tác vụ của Khai phá dữ liệu

Dữ liệu biểu diễn gen có thể được phân cụm theo hai cách. Cách thứ nhất là nhóm các các mẫu gen giống nhau, ví dụ như gom các dòng của ma trận D. Cách khác là nhóm các mẫu khác nhau trên các hồ sơ tương ứng, ví dụ như gom các cột của ma trận D.

1.2.2 Phân cụm dữ liệu phục trong sức khỏe tâm lý

Phân cụm dữ liệu áp dụng trong nhiều lĩnh vực sức khỏe tâm lý, bao gồm cả việc thúc đẩy và duy trì sức khỏe, cải thiện cho hệ thống chăm sóc sức khỏe, và công tác phòng chống bệnh tật và người khuyết tật (Clatworthy et al., 2005). Trong sự phát triển hệ thống chăm sóc sức khỏe, phân cụm dữ liệu được sử dụng để xác định các nhóm của người dân mà có thể được hưởng lợi từ các dịch vụ cụ thể (Hodges và Wotring, 2000). Trong thúc đẩy y tế, nhóm phân tích được sử dụng để lựa chọn nhắm mục tiêu vào nhóm sẽ có khả năng đem lại lợi ích cho sức khỏe cụ thể từ các chiến dịch quảng bá và tạo điều kiện thuận lợi cho sự phát triển của quảng cáo. Ngoài ra, phân cụm dữ liệu

được sử dụng để xác định các nhóm dân cư bị rủi ro do phát triển y tế và các điều kiện những người có nguy cơ nghèo.

1.2.3 Phân cụm dữ liệu đối với hoạt đông nghiên cứu thị trường

Trong nghiên cứu thị trường, phân cụm dữ liệu được sử dụng để phân đoạn thị trường và xác định mục tiêu thị trường (Chrisoppher, 1969; Saunders, 1980, Frank and Green, 1968). Trong phân đoạn thị trường, phân cụm dữ liệu thường được dùng để phân chia thị trường thành nhưng cụm mang ý nghĩa, chẳng han như chia ra đối tượng nam giới từ 21-30 tuổi và nam giới ngoài 51 tuổi, đối tượng nam giới ngoài 51 tuổi thường không có khuynh hướng mua các sản phẩm mới.

1.2.4 Phân cụm dữ liệu đối với hoạt động Phân đoạn ảnh

Phân đoạn ảnh là việc phân tích mức xám hay mầu của ảnh thành các lát đồng nhất (Comaniciu and Meer, 2002). Trong phân đoạn ảnh, phân cụm dữ liệu thường được sử dụng để phát hiện biên của đối tượng trong ảnh.

Phân cụm dữ liệu là một công cụ thiết yếu của khai phá dữ liệu, khai phá dữ liệu là quá trình khám phá và phân tích một khối lượng lớn dữ liệu để lấy được các thông tin hữu ích (Berry and Linoff, 2000). Phân cụm dữ liệu cũng là một vấn đề cơ bản trong nhận dạng mẫu (pattern recognition). Hình 1.1 đưa ra một danh sách giản lược các tác vụ đa dạng của khai phá dữ liệu và chứng tỏ vai trò của phân cụm dữ liệu trong khai phá dữ liệu.

Nhìn chung, Thông tin hữu dụng có thể được khám phá từ một khối lượng lớn dữ liệu thông qua phương tiện tự động hay bán tự động (Berry and Linoff, 2000). Trong khai phá dữ liệu gián tiếp, không có biến nào được chọn ra như một biến đích, và mục tiêu là để khám phá ra một vài mối quan hệ giữa tất cả các biến. Trong khi đó đối với khai phá dữ liệu gián tiếp một vài biến lại được chọn ra như các biến đích. Phân cụm dữ liệu là khai phá dữ liệu gián tiếp, bởi vì trong khai phá dữ liệu, ta không đảm bảo chắc chắn chính xác cụm dữ liệu mà chúng ta đang tìm kiếm, đóng vai trò gì trong việc hình thành các cụm dữ liệu đó, và nó làm như thế nào.

Vấn đề phân cụm dữ liệu đã được quan tâm một cách rộng rãi, mặc dù chưa có định nghĩa đồng bộ về phân cụm dữ liệu và có thể sẽ không bao giờ là một và đi đến thống nhất.(Estivill-Castro,2002; Dubes, 1987; Fraley and Raftery, 1998). Nói một cách đại khái là : Phân cụm dữ liệu, có nghĩa là ta