ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

TRỊNH VĂN HÀ

LỰA CHỌN THUỘC TÍNH TRONG KHAI PHÁ DỮ LIỆU

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

THÁI NGUYÊN 2008

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

TRỊNH VĂN HÀ

LỰA CHỌN THUỘC TÍNH TRONG KHAI PHÁ DỮ LIỆU

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số ***:*** 60.48.01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hướng dẫn khoa học: TS NGUYỄN THANH TÙNG

THÁI NGUYÊN 2008

MỤC LỤC

LỜI MỞ ĐẦU

Như đã biết, trong những năm gần đây công nghệ thông tin phát triển vô cùng nhanh chóng và được ứng dụng rộng rãi trong mọi lĩnh vực đời sống xã hội, nhất là trong quản lý, một lĩnh vực mà yếu tố khoa học công nghệ có tính quyết định. Sự việc đó dẫn đến sự bùng nổ thông tin, làm cho những nhà quản lý rơi vào tình trạng "ngập lụt thông tin". Chính vì vậy, các chuyên gia cho rằng, hiện nay chúng ta đang sống trong một xã hội "*rất giàu về thông tin nhưng nghèo về tri thức*". Tình hình đó đòi hỏi phải phát triển các phương pháp khai phá, phát hiện ra những thông tin, tri thức có ích bị che giấu trong các "núi" dữ liệu phục vụ cho công việc của các nhà quản lý, các chuyên gia, từ đó thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh của các tổ chức, doanh nghiệp.

Khai phá dữ liệu (Data Mining) là một lĩnh vực khoa học liên ngành mới xuất hiện gần đây nhằm đáp ứng nhu cầu này. Các kết quả nghiên cứu cùng với những ứng dụng thành công trong khai phá dữ liệu, khám phá tri thức cho thấy khai phá dữ liệu là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống.

Hiện nay, các CSDL cần khai phá thường có kích thước rất lớn, chẳng hạn các CSDL tin-sinh-học (Bioinformatics), CSDL đa phương tiện, CSDL giao tác, … . Các CSDL này thường chứa tới hàng ngàn thuộc tính, gây rất nhiều khó khăn cho việc khai phá, thậm chí còn làm cho nhiệm vụ khai phá trở nên bất khả thi. Vấn đề đặt ra là phải tìm cách rút gọn số thuộc tính mà không làm những thông tin cần thiết phục vụ nhiệm vụ khai phá.

Mục đích của rút gọn thuộc tính là làm giảm số chiều của không gian thuộc tính, loại bỏ dữ liệu dư thừa, không liên quan. Rút gọn thuộc tính đóng vai trò quan trọng trong bước tiền xử lý dữ liệu cũng như trong quá trình khai phá. Kết quả rút gọn thuộc tính ảnh hưởng trực tiếp đến hiệu quả thực hiện các nhiệm vụ

khai phá: Gia tăng tốc độ, cải thiện chất lượng, tính dễ hiểu của các kết quả thu được.

Từ năm 1970 đến nay, rút gọn thuộc tính (hay còn gọi là rút gọn số chiều – Dimension reduction) đã trở thành đề tài được quan tâm bởi nhiều nhà nghiên cứu thuộc các lĩnh vực nhận dạng thống kê, học máy, khai phá dữ liệu.

Chính những lý do trên, chúng tôi chọn đề tài "***Lựa chọn thuộc tính trong khai phá dữ liệu***" làm đề tài nghiên cứu của mình.

Nội dung của luận văn được trình bày trong 3 chương và phần kết luận.

Chương 1: Trình bày khái quát về Khai phá dữ liệu, bao gồm: Khai phá dữ liệu là gì, quy trình khai phá, các kỹ thuật và một số ứng dụng quan trọng của khai phá dữ liệu.

Chương 2: Trình bày khái quát về nội dung, các cách tiếp cận, quy trình giải quyết vấn đề lựa chọn thuộc tính và một số ứng dụng quan trọng của lựa chọn thuộc tính.

Chương 3: Trình bày kết quả nghiên cứu một số thuật toán lựa chọn thuộc tính điển hình.

Thái nguyên, tháng 11 năm 2008.

Học viên

Trịnh Văn Hà

CHƯƠNG 1

KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU

1.1. Tại sao phải khai phá dữ liệu.

Ước tính cứ khoảng 20 tháng lượng thông tin trên thế giới lại tăng gấp đôi. Chính vì vậy, hiện nay lượng dữ liệu mà con người thu thập và lưu trữ được trong các kho dữ liệu là rất lớn, nhiều khi vượt quá khả năng quản lý. Thời gian này, người ta bắt đầu đề cập đến khái niệm khủng hoảng phân tích dữ liệu tác nghiệp để cung cấp thông tin với yêu cầu chất lượng ngày càng cao cho những người ra quyết định trong các tổ chức tài chính, thương mại, khoa học, ... Đúng như John Naisbett đã cảnh báo "*Chúng ta đang chìm ngập trong dữ liệu mà vẫn đói tri thức*".

Với một khối lượng dữ liệu tăng nhanh và khổng lồ như vậy, rõ ràng các phương pháp thủ công truyền thống áp dụng để phân tích dữ liệu sẽ không hiệu quả, tốn kém và dễ dẫn đến những sai lệch. Do đó để có thể khai phá hiệu quả các cơ sở dữ liệu lớn cần phải có những kỹ thuật mới, các kỹ thuật khai phá dữ liệu (Data Mining).

Khai phá dữ liệu là một lĩnh vực khoa học mới xuất hiện, nhằm tự động hóa khai thác những thông tin, tri thức hữu ích, tiềm ẩn trong các CSDL cho các tổ chức, doanh nghiệp, ... từ đó thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh của tổ chức, doanh nghiệp này. Các kết quả nghiên cứu cùng với những ứng dụng thành công trong khai phá dữ liệu, khám phá tri thức cho thấy khai phá dữ liệu là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống. Hiện nay, khai phá dữ liệu được ứng dụng rộng rãi trong các lĩnh vực như: Phân tích dữ liệu hỗ trợ ra quyết định, điều trị y học, tin-sinh học, thương mại, tài chính, bảo hiểm, text mining, web mining ... .

Do sự phát triển nhanh chóng về phạm vi áp dụng và các phương pháp tìm kiếm tri thức, nên đã có nhiều quan điểm khác nhau về khai phá dữ liệu. Tuy nhiên, ở một mức độ trừu tượng nhất định, chúng ta định nghĩa khai phá dữ liệu như sau :

Khai phá dữ liệu là quá trình tìm kiếm, phát hiện các tri thức mới, hữu ích tiềm ẩn trong cơ sở dữ liệu lớn.

Khám phá tri thức trong CSDL (Knowledge Discovery in Databaes – KDD) là mục tiêu chính của khai phá dữ liệu, do vậy hai khái niệm khai phá dữ liệu và KDD được các nhà khoa học xem là tương đương nhau. Thế nhưng, nếu phân chia một cách chi tiết thì khai phá dữ liệu là một bước chính trong quá trình KDD.

Khám phá tri thức trong CSDL là lĩnh vực liên quan đến nhiều ngành như: Tổ chức dữ liệu, xác suất, thống kê, lý thuyết thông tin, học máy, CSDL, thuật toán, trí tuệ nhân tạo, tính toán song song và hiệu năng cao, ... Các kỹ thuật chính áp dụng trong khám phá tri thức phần lớn được thừa kế từ các ngành này.

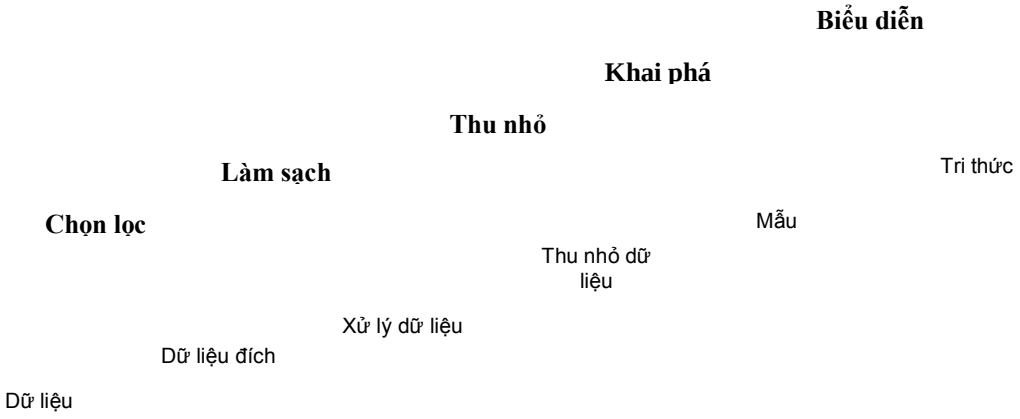
1.2. Quá trình khai phá dữ liệu

Quá trình khám phá tri thức có thể phân thành các công đoạn sau :

- *Trích lọc dữ liệu*: Là bước tuyển chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định.
- *Tiền xử lý dữ liệu*: Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán, ...), tổng hợp dữ liệu (nén, nhóm dữ liệu, tính tổng, xây dựng các histograms, lấy mẫu, ...), rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, entropy, phân khoảng, ...). Sau bước tiền sử lý này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hóa.

- *Biến đổi dữ liệu*: Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ việc áp dụng các kỹ thuật khai phá ở bước sau.
- *Khai phá dữ liệu*: Là bước áp dụng những kỹ thuật phân tích (phần nhiều là các kỹ thuật học máy) nhằm khai thác dữ liệu, trích lọc những mẫu tin (information patterns), những mối quan hệ đặc biệt trong dữ liệu. Đây được xem là bước quan trọng và tiêu tốn thời gian nhất của toàn bộ quá trình KDD.
- *Đánh giá và biểu diễn tri thức*: Những mẫu thông tin và mối quan hệ trong dữ liệu đã được phát hiện ở bước khai phá dữ liệu được chuyển sang và biểu diễn ở dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật, ... Đồng thời bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

Hình 1.1 dưới đây mô tả các công đoạn của khai phá dữ liệu:



***Hình* 1.1. Các bước thực hiện quá trình khai phá dữ liệu**

Nếu theo quan điểm của học máy (Machine Learning), thì các kỹ thuật khai phá dữ liệu bao gồm:

- *Học có giám sát* (*Supervised Learning*) : Là quá trình phân lớp các đối tượng trong cơ sở dữ liệu dựa trên một tập các ví dụ huấn luyện về các thông tin về nhãn lớp đã biết.

- *Học không có giám sát* (*Unsupervised Learning*) : Là quá trình phân chia một tập các đối tượng thành các lớp hay cụm (clusters) tương tự nhau mà không biết trước các thông tin về lớp và không có các ví dụ huấn luyện.
- *Học nửa giám sát* (*Semi-Supervised Learning*) : Là quá trình phân chia một tập các đối tượng thành các lớp dựa trên một *tập nhỏ các ví dụ huấn luyện* và một số thông tin về *một số nhãn lớp* đã biết.

1.3. Các phương pháp khai phá dữ liệu

Kỹ thuật khai phá dữ liệu thường được chia làm 2 nhóm chính:

Kỹ thuật mô tả: Các nhiệm vụ mô tả về ccác tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (clustering), tóm tắt (summerization), trực quan hóa (visualiztion), phân tích sự phát triển và độ lệch (Evolution and deviation analysis), phân tích luật kết hợp (association rules analysis)… .

Kỹ thuật dự đoán: Có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời. Các kỹ thuật này gồm: Phân lớp (classification), hồi quy (regression), … .

Với hai đích chính của khai phá dữ liệu là Dự đoán (Prediction) và Mô tả (Description), người ta thường sử dụng các kỹ thuật sau cho khai phá dữ liệu:

- *Phân lớp và dự đoán* (*classification and prediction*) : Là việc xếp các đối tượng vào những lớp đã biết trước. Ví dụ, phân lớp các bệnh nhân, phân lớp các loài thực vật, ... Hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơ-ron nhân tạo (neural network), ... Phân lớp và dự đoán còn được gọi là học có giám sát.
- *Phân cụm* (*clustering*/*segmentation*) : Là việc xếp các đối tượng theo từng cụm tự nhiên.