ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ THỊ UYÊN

KHAI PHÁ LUẬT QUYẾT ĐỊNH TRÊN BẢNG DỮ LIỆU CÓ THUỘC TÍNH THAY ĐỔI

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: GS.TS VŨ ĐỨC THI

THÁI NGUYÊN - 2013

LỜI CAM ĐOAN

Tôi xin cam đoan rằng đây là công trình nghiên cứu của tôi, có sự hỗ trợ từ Giáo viên hướng dẫn là **GS.TS Vũ Đức Thi**. Các nội dung nghiên cứu và kết quả trong đề tài này là trung thực và chưa từng được ai công bố trong bất cứ công trình nghiên cứu nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi trong phần tài liệu tham khảo. Ngoài ra, đề tài còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả, cơ quan tổ chức khác, và cũng được thể hiện trong phần tài liệu tham khảo. Nếu sai tôi xin hoàn toàn chịu trách nhiệm.

Thái Nguyên, ngày 15 tháng 9 năm 2013

Tác giả luận văn

Lê Thị Uyên

LỜI CẢM ƠN

Với lòng biết ơn sâu sắc nhất, em xin gửi đến các Thầy Cô ở Trường Đại Học Công Nghệ Thông Tin và Truyền thông cùng các Thầy ở Viện Khoa Học và Công Nghệ Việt Nam đã cùng với tri thức và tâm huyết của mình để truyền đạt vốn kiến thức quý báu cho chúng em trong suốt khóa học vừa qua.

Luận văn được hoàn thành dưới sự hướng dẫn, chỉ bảo tận tình của GS.TS Vũ Đức Thi. Xin bày tỏ lòng biết ơn chân thành và sâu sắc tới Thầy đã quan tâm, nghiêm khắc và tạo mọi điều kiện để em có thể hoàn thành những mục tiêu của đề tài.

Sau cùng, em xin kính chúc các Thầy Cô thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp của mình là truyền đạt kiến thức cho thế hệ mai sau.

Em xin chân thành cảm ơn!

Thái Nguyên, ngày 15 tháng 9 năm 2013

Tác giả luận văn

Lê Thị Uyên

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VIẾT TẮT

Ký hiệu	Ý nghĩa
$BN_p(X)$	P – miền biên của X
$\overline{P}X$	P – Xấp xỉ trên của X
$\underline{P}X$	P – Xấp xỉ dưới của X
IND(P)	P – Quan hệ bất khả phân biệt
$Sup(C_i, D_j)$	Độ hỗ trợ của luật quyết định $C_i \rightarrow D_j$
$Cov(C_i, D_j)$	Độ phủ của luật quyết định $C_i \rightarrow D_j$
$Acc(C_i, D_j)$	Độ chính xác của luật quyết định $C_i \rightarrow D_j$
$Acc^{(t)}(C, D)$	Ma trận Độ chính xác tại thời điểm t của tất cả luật quyết định $C_i \rightarrow D_j$
$Sup^{(t)}(C, D)$	Ma trận Độ hỗ trợ tại thời điểm t của tất cả luật quyết định $C_i \rightarrow D_j$
$Cov^{(t)}(C, D)$	Ma trận Độ phủ tại thời điểm t của tất cả luật quyết định $C_i \rightarrow D_j$

DANH MỤC HÌNH

CHƯƠNG 1:
TỔNG QUAN

1.1. Khai phá dữ liệu

Khám phá tri thức là một quá trình tìm kiếm trong cơ sở dữ liệu các mẫu đúng đắn, có ích tiềm tàng và có thể hiểu được đối với người sử dụng. Quá trình khám phá tri thức gồm nhiều pha, mỗi pha có vai trò và tầm quan trọng riêng. Khai phá dữ liệu (DM) là một pha quan trọng trong toàn bộ tiến trình khai phá tri thức, sử dụng các thuật toán đặc biệt để chiết xuất các mẫu từ dữ liệu. Về bản chất đây là giai đoạn duy nhất để rút trích và tìm ra được các mẫu, các mô hình, các tri thức tiềm ẩn có trong cơ sở dữ liệu phục vụ cho việc mô tả và dự đoán.

Quá trình khai phá dữ liệu trải qua ba bước:

Bước một: Lọc dữ liệu (giai đoạn tiền xử lý). Khi dữ liệu được thu thập từ nhiều nguồn khác nhau, nên sẽ có những sự sai sót, dư thừa và trùng lặp. Lọc dữ liệu nhằm loại bỏ những dư thừa để có được dữ liệu ở định dạng thống nhất. Dữ liệu sau khi lọc và chỉnh sửa sẽ gọn hơn, do vậy có thể xử lý nhanh chóng hơn.

Ví dụ, trong bài toán tìm quy luật mua hàng của khách hàng trong một siêu thị, ta cần phải xem khách hàng thường cùng mua những mặt hàng nào, dựa trên đó ta sẽ sắp xếp những món hàng để thuận tiện cho việc mua hàng của khách hàng. Từ dữ liệu nguồn do siêu thị cung cấp, có thể có nhiều thuộc tính không cần thiết cho khai phá dữ liệu như: Mã khách hàng, nhà cung cấp, đơn giá hàng, người bán hàng, … Các dữ liệu này cần cho quản lý bán hàng nhưng không cần cho khai phá dữ liệu, vì vậy có thể loại bỏ các thuộc tính này trước khi tiến hành công việc khai phá dữ liệu.

Bước hai: Khai phá dữ liệu (là công việc chính) sử dụng các thuật toán khác nhau để khai phá các tri thức tiềm ẩn trong dữ liệu.

Bước ba (giai đoạn hậu xử lý) là quá trình đánh giá kết quả khai phá theo yêu cầu của người dùng. Các kỹ thuật khai phá dữ liệu khác nhau được đánh giá theo các quy tắc, trong số các kết quả thỏa mãn yêu cầu đánh giá, giữ lại kết quả phù hợp nhất với yêu cầu của người sử dụng.

Có nhiều kỹ thuật khai phá dữ liệu được nghiên cứu, trong đó có hai kỹ thuật được các nhà nghiên cứu sử dụng nhiều nhất là: Kỹ thuật phân lớp dữ liệu và kỹ thuật phân nhóm dữ liệu.

1.1.1. Kỹ thuật phân lớp dữ liệu

Phân lớp dữ liệu là kỹ thuật nhằm xây dựng mô hình cho phép phân các đối tượng vào lớp được biết trước nào đó. Kỹ thuật này phép dự đoán giá trị bị thiếu của thuộc tính trong dữ liệu hay dự đoán giá trị của dữ liệu sẽ xuất hiện trong tương lai. Phân lớp dữ liệu là kỹ thuật được xem là một trong những kỹ thuật hay được dùng nhất trong học máy và khai phá dữ liệu.

Quá trình phân lớp dữ liệu được thực hiện qua hai bước. Thứ nhất dựa vào tập hợp dữ liệu huấn luyện (các đối tượng dữ liệu đã được gán nhãn lớp) để xây dựng một mô hình mô tả những đặc trưng của các lớp hoặc các khái niệm tương ứng với lớp. Thứ hai dựa trên mô hình phân lớp dữ liệu hoặc mô hình diễn giải và phân biệt các khái niệm đã được xác định để gán nhãn lớp cho những đối tượng được quan tâm.

1.1.2. Một số kỹ thuật phân lớp phổ biến

Cây quyết định là một cấu trúc cây, trong đó mỗi nút trong của cây biểu thị một phép phân nhánh tương ứng với một thuộc tính; mỗi nhánh biểu thị một điều kiện; các nút lá tương ứng với các lớp. Để phân lớp một đối tượng chưa biết, các giá trị thuộc tính của đối tượng đó được kiểm tra bám theo cây quyết định. Đường dẫn đi từ gốc đến một nút lá nào đó tương ứng với đối tượng cho phép xác định lớp tương ứng của nó. Cây quyết định có thể dễ dàng chuyển thành một tập các luật phân lớp.

Tập thô được sử dụng trong việc phân lớp nhằm mục đích khám phá các quan hệ có cấu trúc đối với dữ liệu không chính xác hoặc dữ liệu có các giá trị thuộc tính đã được rời rạc hóa. Do đó, đối với các thuộc tính có giá trị liên tục thì nó phải được rời rạc hóa trước khi sử dụng.

1.1.3. Kỹ thuật phân nhóm dữ liệu

Phân nhóm dữ liệu là quá trình nhóm những đối tượng thành các lớp. Các đối tượng trong một lớp tương đồng nhau, nhưng độ tương đồng của chúng phải lớn hơn độ tương đồng với các đối tượng trong các lớp khác. Trong phân nhóm, không đòi hỏi biết được số lớp cần cấu tạo ra. Mặt khác, với kỹ thuật này, các đối tượng được nhóm lại trong cùng một lớp dựa vào sự giống nhau của chúng, được xác định bởi những đặc trưng giống nhau. Thông thường, người ta sử dụng sự giống nhau định lượng dưới dạng khoảng cách. Độ đo giống nhau có thể xác định dựa trên ý kiến chuyên gia trong lĩnh vực.

1.2. Khai phá luật quyết định

Khai phá các luật quyết định là quá trình xác định những luật quyết định trên bảng quyết định cho trước, phục vụ cho việc phân lớp của các đối tượng mới. Khai phá luật quyết định đã được nhiều chuyên gia trong và ngoài nước quan tâm trên cả hai phương diện lý thuyết và ứng dụng, các nghiên cứu này chủ yếu xem xét trên các bảng dữ liệu tĩnh.

Trong thực tế, dữ liệu thường xuyên thay đổi theo thời gian. Đã có một số nghiên cứu về các khía cạnh khác nhau để cập nhật tri thức trên các bảng dữ liệu động, tập trung chủ yếu vào ba trường hợp sau đây:

+ Tập các giá trị thuộc tính thay đổi trong khi tập các đối tượng và tập các thuộc tính không đổi.

+ Tập các đối tượng thay đổi trong khi tập các thuộc tính và tập các giá trị thuộc tính không đổi.