ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

PHẠM THANH TUẤN

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP TÌM CÁC LUẬT KẾT HỢP PHÂN LỚP TRÊN TẬP MẪU HỌC VÀ ỨNG DỤNG TRONG CHẨN ĐOÁN BỆNH

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên, 2019

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

PHẠM THANH TUẤN

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP TÌM CÁC LUẬT KẾT HỢP PHÂN LỚP TRÊN TẬP MẪU HỌC VÀ ỨNG DỤNG TRONG CHẨN ĐOÁN BỆNH

Chuyên ngành: Khoa học máy tính
Mã số: 8 48 01 01

Người hướng dẫn khoa học: TS. Lê Văn Phùng

Thái Nguyên, 2019

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu do tôi thực hiện, dưới sự hướng dẫn khoa học của TS. Lê Văn Phùng. Các số liệu và kết quả trình bày trong luận văn là trung thực, chưa được công bố bởi bất kỳ tác giả này hay ở bất kỳ công trình nào khác.

LỜI CẢM ƠN

Trong quá trình thực hiện đề tài "Nghiên cứu một số phương pháp tìm các luật kết hợp phân lớp trên tập mẫu học và ứng dụng trong chẩn đoán bệnh", tôi đã nhận được rất nhiều sự giúp đỡ, tạo điều kiện của tập thể Ban Giám hiệu, Phòng Đào tạo, khoa Công nghệ thông tin và các phòng chức năng của trường Đại học Công nghệ thông tin và truyền thông, Đại học Thái Nguyên. Tôi xin bày tỏ lòng cảm ơn chân thành về sự giúp đỡ quý báu đó.

Tôi xin được bày tỏ lòng biết ơn sâu sắc đến TS. Lê Văn Phùng là thầy giáo trực tiếp hướng dẫn, chỉ bảo giúp tôi hoàn thành luận văn này.

TÁC GIẢ LUẬN VĂN

Phạm Thanh Tuấn

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT

1. DM – Data Mining.
2. CSDL – Cơ sở dữ liệu.
3. CBA - Classification-Based Associon
4. CMAR - Classification based on Multiple Asociation Rule

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Thế kỷ XXI được xem là một kỷ nguyên của công nghệ thông tin. Cùng với việc ứng dụng công nghệ thông tin ở hầu hết các lĩnh vực trong nhiều năm qua dẫn đến lượng dữ liệu, thông tin của nhân loại được lưu trữ ngày một tăng. Nguồn dữ liệu khổng lồ ấy được tích lũy với tốc độ bùng nổ từ rất nhiều lĩnh vực: khoa học, kinh doanh, giao dịch, thương mại, chứng khoán,… Vậy chúng ta có thể khai thác được gì từ "núi" dữ liệu tưởng chừng như bỏ đi ấy.

Cùng với việc tăng không ngừng khối lượng dữ liệu, các hệ thống thông tin cũng được chuyên môn hóa, phân hạch hóa theo các lĩnh vực như sản xuất, tài chính, buôn bán thị trường .v.v, tuy nhiên các hệ quản trị cơ sở dữ liêu truyền thống chỉ khai thác được một lượng thông tin nhỏ không còn đáp ứng đủ những yêu câu, những thách thức mới. Do vậy một khuynh hướng mới được ra đời đó là kỹ thuật phát hiện tri thức trong cơ sở dữ liệu. Khai phá dữ liệu (Data Mining – DM) ra đời phần nào đó đã giải quyết hữu hiệu những yêu cầu, thách thức đó.

Một trong những lĩnh vực nghiên cứu các phương pháp ứng dụng khai phá dữ liệu, tìm kiếm tri thức, kết xuất tri thức… từ dữ liệu là tìm kiếm các Luật kết hợp phân lớp (Class Association Rules) cũng được nghiên cứu từ nhiều năm trước đây và đã có những kết quả khả quan và mang lại hướng ứng dụng có hiệu quả cao. Ngày nay, kỹ thuật khai phá dữ liệu dựa trên việc tìm kiếm các luật kết hợp phân lớp đã được áp dụng và mang lại hiệu quả cho nhiều ngành, nhiều lĩnh vực như: Kinh tế, tài chính, khoa học - kỹ thuật, ngân hàng, thương mại, giáo dục, y tế… các kỹ thuật khai phá dự liệu bằng Luật kết hợp phân lớp rất đa dạng và phong phú như các kỹ thuật dựa trên các thuật toán CBA-RG, CBA-CB,…

Với mong muốn nắm vững hơn các quá trình phát hiện tri thức từ dữ liệu