

**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

LỜI CAM ĐOAN .....	i
LỜI CẢM ƠN .....	ii
DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT.....	v
DANH MỤC BẢNG BIỂU .....	vi
DANH MỤC HÌNH VẼ.....	vii
MỞ ĐẦU .....	viii
<b>CHƯƠNG 1. PHÂN LỚP VÀ PHƯƠNG PHÁP XÂY DỰNG CÂY</b>	
<b>PHÂN LỚP THEO TẬP MẪU HỌC .....</b>	<b>1</b>
1.1. Tổng quan về kỹ thuật khai phá dữ liệu.....	1
1.1.1. Khái niệm về khai phá dữ liệu .....	1
1.1.2. Một số phương pháp khai phá dữ liệu hiện đại và thông dụng.....	2
1.1.3. Các ứng dụng khai phá dữ liệu .....	3
1.2. Những vấn đề chung nhất về phân lớp và phương pháp phân lớp cơ bản.	7
1.2.1 Khái niệm phân lớp dữ liệu.....	7
1.2.2. Các bước tiến hành phân lớp dữ liệu .....	7
1.2.3. Phân lớp theo cây quyết định.....	9
1.2.4. Phân lớp kiểu Bayes.....	12
1.2.5. Phân lớp dựa trên các quy tắc IF-THEN.....	13
1.2.6. Phân lớp dựa trên luật kết hợp .....	16
1.2.7. Phân lớp dựa vào K-lân cận gần nhất .....	18
1.2.8. Phân lớp dựa vào giải thuật di truyền .....	19
1.2.9. Phân lớp theo cách tiếp cận tập thô.....	20
1.2.10. Phân lớp theo cách tiếp cận tập mờ .....	21
1.3. Khái niệm về tập mẫu học và phương pháp xây dựng cây phân lớp.....	24
1.3.1. Định nghĩa tập mẫu học .....	24
1.3.2. Xây dựng cây phân lớp dựa theo Khóa.....	24

1.3.3. Xây dựng cây phân lớp nhờ các luật kết hợp phân lớp (Class Association Rules) trong bảng mẫu học .....	27
<b>CHƯƠNG 2. MỘT SỐ PHƯƠNG PHÁP TÌM CÁC LUẬT KẾT HỢP PHÂN LỚP TRÊN TẬP MẪU HỌC .....</b>	<b>29</b>
2.1. Phương pháp phân lớp dựa trên luật kết hợp .....	29
2.1.1. Các bước tiến hành phân lớp dựa trên luật kết hợp .....	29
2.1.2. Tạo luật kết hợp bằng cây quyết định .....	29
2.2. Một số thuật toán cổ điển xây dựng cây phân lớp dựa trên luật kết hợp. 29	
2.2.1. Thuật toán CBA-RG .....	30
2.2.2. Thuật toán CBA-CB.....	32
2.3. Thuật toán hiện đại.....	34
2.3.1. Thuật toán CBA cải tiến.....	34
2.3.2. Ví dụ áp dụng thuật toán cải tiến .....	37
<b>CHƯƠNG 3. CHƯƠNG TRÌNH THỬ NGHIỆM TÌM CÁC LUẬT KẾT HỢP PHÂN LỚP DỰA TRÊN TẬP MẪU HỌC.....</b>	<b>42</b>
3.1. Bài toán thử nghiệm.....	42
3.1.1. Bài toán và tập mẫu học đầu vào .....	42
3.1.2. Chọn thuật toán thử nghiệm.....	46
3.2. Môi trường thử nghiệm .....	47
3.2.1. Chọn môi trường chứa dữ liệu đầu vào .....	47
3.2.2. Chọn ngôn ngữ lập trình .....	47
3.3. Nội dung và kết quả thử nghiệm.....	47
3.3.1. Mô hình thuật toán thử nghiệm.....	47
3.3.3. Một số giao diện chính của chương trình thử nghiệm .....	50
3.4. Đánh giá chương trình thử nghiệm .....	51
3.5. Mở rộng bài toán .....	51
<b>KẾT LUẬN.....</b>	<b>60</b>
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>62</b>

## **DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT**

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

## DANH MỤC BẢNG BIỂU

Bảng 1.1.	Ví dụ về tập mẫu học.....	15
Bảng 1.2.	Các bộ huấn luyện đã được phân lớp trong CSDL.....	20
Bảng 1.3.	Ví dụ tập mẫu học được phân lớp dựa theo khóa.....	33
Bảng 2.1.	Ví dụ tập mẫu học để tìm các luật kết hợp phân lớp theo thuật toán cải tiến.....	47
Bảng 2.2.	Bảng tổng hợp.....	49
Bảng 2.3a.	Khoản mục.....	50
Bảng 2.3b.	Các luật kết hợp phân lớp phổ biến 1 – Khoản mục.....	50
Bảng 2.3c.	Các luật kết hợp phân lwps 2 – Khoản mục.....	50
Bảng 3.1.	Tập mẫu học.....	55
Bảng 3.2.	Bảng mẫu học được số hóa.....	56
Bảng 3.3.	Bảng tổng hợp kết quả thu được.....	59
Bảng 3.4.	Bảng mẫu học (mở rộng) đầu vào.....	60
Bảng 3.5.	Bảng mẫu học mở rộng được số hóa.....	64



**DANH MỤC HÌNH VẼ**

Hình 1.1. Cây quyết định cho việc chơi Gold.....	16
Hình 1.2. Một tập thô xấp xỉ tập các bộ của C khi dùng các tập xấp xỉ trên và dưới của C. Các vùng hình chữ nhật biểu diễn các lớp tương đương.....	27
Hình 1.3. Các giá trị mờ thật với thu nhập, biểu diễn mức thành viên các giá trị thu nhập theo các loại {thấp, trung bình, cao}.....	28
Hình 1.4. Cây phân lớp xây dựng với 2 trường hợp.....	34

## 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 liệu truyền thống chỉ khai thác được một lượng thông tin nhỏ không còn đáp ứng đủ những yêu cầu, 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 dữ 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