

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

---

**NGUYỄN ĐĂNG NGUYỄN**

**PHƯƠNG PHÁP XÂY DỰNG CÂY QUYẾT ĐỊNH  
DỰA TRÊN TẬP PHỤ THUỘC HÀM XẤP XỈ**

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**THÁI NGUYÊN - 2017**

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

---

**NGUYỄN ĐĂNG NGUYỄN**

**PHƯƠNG PHÁP XÂY DỰNG CÂY QUYẾT ĐỊNH  
DỰA TRÊN TẬP PHỤ THUỘC HÀM XẤP XỈ**

**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: TS. LÊ VĂN PHÙNG**

**THÁI NGUYÊN - 2017**

## LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do chính tôi thực hiện, dưới sự hướng dẫn khoa học của TS. Lê Văn Phùng, số liệu và kết quả nghiên cứu trong luận văn này hoàn toàn trung thực và chưa sử dụng để bảo vệ một công trình khoa học nào, các thông tin, tài liệu trích dẫn trong luận văn đã được chỉ rõ nguồn gốc. Mọi sự giúp đỡ cho việc hoàn thành luận văn đều đã được cảm ơn. Nếu sai tôi hoàn toàn chịu trách nhiệm.

*Thái Nguyên, tháng 05 năm 2017*

**Học viên**

**Nguyễn Đăng Nguyên**

## LỜI CẢM ƠN

Trước hết em xin trân trọng cảm ơn các thầy giáo, cô giáo trường Đại học Công nghệ Thông tin và Truyền thông đã giảng dạy em trong quá trình học tập chương trình sau đại học. Dù rằng, trong quá trình học tập có nhiều khó khăn trong việc tiếp thu kiến thức cũng như sưu tầm tài liệu học tập, nhưng với sự nhiệt tình và tâm huyết của thầy cô cùng với những nỗ lực của bản thân đã giúp em vượt qua được những trở ngại đó.

Em xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo TS.Lê Văn Phùng người hướng dẫn khoa học, đã tận tình hướng dẫn em trong suốt quá trình làm luận văn.

Xin chân thành cảm ơn các bạn bè, đồng nghiệp, các bạn học viên lớp cao học CK14A, những người thân trong gia đình đã động viên, chia sẻ, tạo điều kiện giúp đỡ trong suốt quá trình học tập và làm luận văn.

*Một lần nữa em xin chân thành cảm ơn!*

*Thái Nguyên, tháng 05 năm 2017*

**Học viên**

**Nguyễn Đăng Nguyên**

## MỤC LỤC

LỜI CAM ĐOAN .....	i
LỜI CẢM ƠN .....	ii
MỤC LỤC .....	iii
DANH MỤC TỪ VIẾT TẮT VÀ KÍ HIỆU SỬ DỤNG .....	vi
DANH MỤC CÁC BẢNG.....	vii
DANH MỤC CÁC HÌNH.....	viii
THUẬT NGỮ TIẾNG ANH.....	ix
<b>MỞ ĐẦU .....</b>	<b>1</b>
<b>Chương 1: TỔNG QUAN VỀ CÂY QUYẾT ĐỊNH VÀ PHỤ THUỘC HÀM XẤP XỈ.....</b>	<b>3</b>
1.1. Tổng quan về khai phá dữ liệu và cây quyết định .....	3
1.1.1. Khái niệm về khai phá dữ liệu, quá trình phát triển và ứng dụng trong việc phát hiện tri thức .....	3
1.1.2. Khái quát về các phương pháp khai phá dữ liệu phổ biến.....	5
1.2. Phụ thuộc hàm xấp xỉ.....	7
1.2.1. Khái niệm về phụ thuộc hàm trong mô hình CSDL quan hệ.....	7
1.2.2. Khái niệm về phụ thuộc hàm xấp xỉ và các đặc trưng của chúng.....	13
1.3. Kết luận chương 1 .....	18
<b>Chương 2: MỘT SỐ THUẬT TOÁN XÁC ĐỊNH PHỤ THUỘC HÀM XẤP XỈ VÀ XÂY DỰNG CÂY QUYẾT ĐỊNH .....</b>	<b>17</b>
2.1. Thuật toán TANE xác định phụ thuộc hàm xấp xỉ từ quan hệ.....	19
2.1.1. Khái niệm lớp tương đương và phân hoạch.....	19
2.1.2. Phân hoạch mịn hơn.....	20
2.1.3. Thuật toán TANE cải tiến .....	24
2.1.4. Chiến lược tìm kiếm.....	24
2.2. Thuật toán xác định phụ thuộc hàm xấp xỉ dựa trên luật kết hợp.....	38

2.2.1. Luật kết hợp .....	38
2.2.2. Biểu diễn PTH xấp xỉ qua LKH.....	41
2.2.3. Độ hỗ trợ của PTH xấp xỉ và tính không tầm thường.....	45
2.2.4. Định nghĩa PTH xấp xỉ mạnh [14].....	47
2.2.5. Biểu diễn độ đo, độ hỗ trợ, độ chính xác qua lý thuyết PTH xấp xỉ.....	48
2.2.6. Thuật toán xác định PTH xấp xỉ dựa trên LKH.....	52
2.3. Thuật toán xác định phụ thuộc hàm xấp xỉ dựa trên phủ tối thiểu và lớp tương đương .....	54
2.3.1. Khái niệm về Phủ tối thiểu và các mệnh đề liên quan .....	54
2.3.2. Thuật toán tìm Phủ tối thiểu.....	56
2.3.3. Thuật toán khai phá PTH xấp xỉ nhờ phủ tối thiểu và lớp tương đương ....	57
2.3.4. Độ phức tạp của thuật toán khai phá PTH xấp xỉ sử dụng phủ tối thiểu và lớp tương đương .....	60
2.4. Thuật toán xây dựng cây quyết định dựa trên phụ thuộc hàm xấp xỉ.....	61
2.4.1. Giải thuật chung xây dựng cây quyết định .....	61
2.4.2. Giải thuật xây dựng cây quyết định dựa trên tập PTH xấp xỉ phân lớp ...	67
2.5. Kết luận chương 2 .....	69
<b>Chương 3: CHƯƠNG TRÌNH THỬ NGHIỆM XÂY DỰNG CÂY QUYẾT ĐỊNH CHẨN ĐOÁN BỆNH TẠI BỆNH VIỆN ĐA KHOA TRUNG ƯƠNG THÁI NGUYÊN DỰA TRÊN VIỆC KHAI PHÁ TẬP PTH XẤP XỈ.....</b>	<b>70</b>
3.1. Mô tả Bài toán chẩn đoán bệnh cúm tại bệnh viện đa khoa Trung ương Thái Nguyên và yêu cầu chương trình.....	70
3.1.1. Giới thiệu về bệnh Cúm .....	70
3.1.2. Quy trình chẩn đoán xác định bệnh cúm .....	71
3.2. Tập dữ liệu huấn luyện (input).....	74
3.3. Ứng dụng hai thuật toán 2.3 và 2.4 để xác định tập phụ thuộc hàm xấp xỉ và xây dựng cây quyết định chẩn đoán bệnh .....	75

3.4. Thiết kế chương trình.....	76
3.5. Các giao diện chính của chương trình.....	77
3.6. Đánh giá kết quả thử nghiệm .....	82
3.7. Kết luận chương 3 .....	83
<b>KẾT LUẬN CHUNG .....</b>	<b>84</b>
1. Kết quả đạt được trong luận văn .....	84
2. Hướng phát triển của đề tài .....	84
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>85</b>

## DANH MỤC TỪ VIẾT TẮT VÀ KÍ HIỆU SỬ DỤNG

Từ và Ký hiệu	Diễn giải
$R(U)$	Quan hệ trên tập thuộc U
$U = \{A_1, \dots, A_m\}$	Tập m thuộc tính.
$S = \langle U, F \rangle$	Lược đồ quan hệ với U là tập thuộc tính, F là tập các phụ thuộc hàm trên U
LĐQH	Lược đồ quan hệ
CSDL	Cơ sở dữ liệu
PTH	Phụ thuộc hàm
KPDL	Khai phá dữ liệu



## DANH MỤC CÁC BẢNG

Bảng 1.1. Ví dụ về quan hệ.....	9
Bảng 1.2. Các thuật toán khám phá phụ thuộc hàm.....	12
Bảng 1.3: Bảng quan hệ ví dụ.....	17
Bảng 1.4: Bảng quan hệ ví dụ về phụ thuộc hàm điều kiện.....	18
Bảng 2.1. Bảng quan hệ minh họa cho phân hoạch.....	20
Bảng 2.2. Bảng quan hệ ví dụ cho phân hoạch mịn hơn.....	21
Bảng 2.3: Bảng quan hệ minh họa cho PTH xấp xỉ.....	22
Bảng 2.4. Ví dụ về CSDL giao tác D.....	38
Bảng 2.5. Ví dụ về các tập phổ biến với độ hỗ trợ tương ứng, $\text{minsupp} = 50\%$ .....	39
Bảng 2.6. Một quan hệ R.....	43
Bảng 2.7. Tập các giao tác TD của R.....	45
Bảng 2.8. Một số LKH trong TD tương ứng với PTH xấp xỉ trong R.....	45

**DANH MỤC CÁC HÌNH**

Hình 1.1. Quá trình phát hiện tri thức .....	5
Hình 1.2. Các loại phụ thuộc dữ liệu .....	9
Hình 1.3. Kỹ thuật phát hiện phụ thuộc hàm .....	12
Hình 2.1. Dàn cho các thuộc tính (A, B, C, D, E) .....	24
Hình 2.2. Một tập đã được cắt tia chứa dàn cho {A,B,C,D}. .....	26
Hình 2.3. Cây trước khi cắt tia .....	65
Hình 2.4. Cây sau khi cắt tia .....	67