

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

**NGUYỄN MINH TÂM**

**NGHIÊN CỨU MỘT SỐ THUẬT TOÁN  
PHÂN CỤM, PHÂN LỚP DỮ LIỆU VÀ ỨNG DỤNG**

**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**

**NGUYỄN MINH TÂM**

**NGHIÊN CỨU MỘT SỐ THUẬT TOÁN  
PHÂN CỤM, PHÂN LỚP DỮ LIỆU VÀ ỨNG DỤNG**

**Chuyên ngành: Khoa học máy tính  
Mã số: 84 80 101**

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

**Người hướng dẫn khoa học: TS. NGUYỄN VĂN NÚI**

**THÁI NGUYÊN - 2019**

## LỜI CẢM ƠN

Em xin chân thành cảm ơn Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên đã tạo điều kiện cho em thực hiện luận văn này.

Em xin gửi lời cảm ơn sâu sắc tới thầy giáo TS. Nguyễn Văn Núi, Bộ môn công nghệ phần mềm - Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên đã trực tiếp hướng dẫn em trong quá trình thực hiện luận văn.

Em cũng xin gửi lời cảm ơn tới các thầy, cô đã có những ý kiến đóng góp bổ ích và đã tạo mọi điều kiện tốt nhất cho em trong suốt thời gian thực hiện luận văn. Xin cảm ơn các bạn học đồng khóa đã thường xuyên động viên, giúp đỡ tôi trong quá trình học tập.

Cuối cùng, em xin gửi lời cảm ơn đến gia đình và đồng nghiệp vì sự ủng hộ và động viên đã dành cho em trong suốt quá trình học tập cũng như thực hiện luận văn này.

Thái Nguyên, tháng 05 năm 2019

Học viên

Nguyễn Minh Tâm

## LỜI CAM ĐOAN

Em xin cam đoan về nội dung đồ án tốt nghiệp với tên đề tài “ **Nghiên cứu một số thuật toán phân cụm, phân lớp dữ liệu và ứng dụng**” không sao chép nội dung từ các luận văn khác, hay các sản phẩm tương tự mà không phải do em làm ra. Sản phẩm luận văn là do chính bản thân em tìm hiểu và xây dựng nên.

Nếu có gì sai em xin chịu mọi hình thức kỷ luật của Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

Thái Nguyên, tháng 05 năm 2019

Học viên

Nguyễn Minh Tâm

## MỤC LỤC

<b>LỜI CẢM ƠN</b> .....	1
<b>LỜI CAM ĐOAN</b> .....	2
<b>MỞ ĐẦU</b> .....	7
<b>CHƯƠNG 1 TỔNG QUAN</b> .....	9
1.1 Giới thiệu chung.....	9
1.2 Các bước trong khai phá dữ liệu .....	10
1.3 Các kỹ thuật áp dụng trong khai phá dữ liệu .....	12
1.4 Ứng dụng của khai phá dữ liệu .....	14
1.5 Những thách thức trong khai phá dữ liệu .....	15
<b>CHƯƠNG 2 PHÂN CỤM DỮ LIỆU VÀ MỘT SỐ THUẬT TOÁN CƠ BẢN</b> ..	17
2.1 Định nghĩa về phân cụm dữ liệu .....	17
2.2 Mục tiêu của phân cụm dữ liệu.....	18
2.3 Bài toán phân cụm dữ liệu .....	20
2.4 Một số kiểu dữ liệu .....	20
2.5 Một số kỹ thuật phân cụm dữ liệu .....	23
2.5.1 Phương pháp phân cụm dữ liệu dựa trên phân cụm phân cấp .....	23
2.5.2 Phương pháp phân cụm dữ liệu dựa trên mật độ.....	25
2.5.3 Phương pháp phân cụm phân hoạch.....	29
2.6 Kết luận .....	33
<b>CHƯƠNG 3 PHÂN LỚP DỮ LIỆU VÀ MỘT SỐ THUẬT TOÁN CƠ BẢN</b> .....	34
3.1 Định nghĩa về phân lớp dữ liệu.....	34
3.2 Các vấn đề quan tâm của phân lớp dữ liệu .....	34
3.2.1 Quá trình phân lớp dữ liệu: .....	34
3.2.2 So sánh các phương pháp phân lớp.....	36
3.3 Phân lớp bằng cây quyết định.....	36
3.3.1 Khái niệm về cây quyết định.....	36
3.3.2 Ưu, nhược điểm của cây quyết định .....	39
3.3.3 Một số thuật toán của cây quyết định.....	40

3.4 Phân lớp bằng Bayesian.....	48
3.5 Phân lớp dựa trên sự kết hợp .....	51
3.5.1 Các khái niệm quan trọng về luật kết hợp.....	51
3.5.2 Một số thuật toán về luật kết hợp .....	52
3.6 Độ chính xác classifier .....	57
3.7 Kết luận .....	59
CHƯƠNG 4 MỘT SỐ KẾT QUẢ THỬ NGHIỆM.....	60
4.1. Giới thiệu về công cụ phân cụm, phân lớp dữ liệu Weka.....	60
4.2. Ứng dụng phân cụm dữ liệu để phân nhóm khách hàng .....	62
4.3. Ứng dụng phân lớp dữ liệu để phân lớp .....	68
4.3.1. Phân lớp dữ liệu với thuật toán Apriori .....	68
4.3.2. Phân lớp dữ liệu với thuật toán Naive Bayes .....	71
KẾT LUẬN .....	75
TÀI LIỆU THAM KHẢO.....	76
Tiếng Việt: .....	76
Tiếng Anh: .....	76

## DANH MỤC CÁC HÌNH VẼ

Hình 1.1. Các bước trong khai phá dữ liệu .....	10
Hình 2.1. Mô phỏng vấn đề phân cụm dữ liệu.....	17
Hình 2.2. Cụm dữ liệu được khám phá bởi giải thuật DBSCAN .....	26
Hình 2.3. Thứ tự phân cụm các đối tượng theo OPTICS .....	29
Hình 2. 4. Phân cụm dựa trên phương pháp k-means.....	31
Hình 4. 1. Giao diện chính của phần mềm.....	61
Hình 4.2. Thông tin dữ liệu cơ bản của file bank-k.arff hiển thị bởi Weka ...	63
Hình 4.3. Lưu đồ thuật toán K-Means .....	64
Hình 4.4. Bảng tham số sử dụng cho thuật toán K-Means: Hình (a) K=3; Hình (b): K=5 .....	65
Hình 4.5. Kết quả phân cụm với thuật toán K-Means (K=3) .....	66
Hình 4.6. Kết quả phân cụm với thuật toán K-Means (K=5) .....	67
Hình 4.7. Giao diện Weka chọn thuật toán Apriori .....	68
Hình 4.8. Giao diện Weka thiết lập tham số cho thuật toán Apriori .....	69
Hình 4.9. Kết quả sinh luật bởi thuật toán Apriori .....	70
Hình 4.10. Giao diện Weka lựa chọn thuật toán Naive Bayes .....	71
Hình 4.11. Kết quả sinh luật bởi thuật toán Naive Bayes.....	72
Hình 4.12. Giao diện Weka lựa chọn thuật toán C4.5 .....	73
Hình 4.13. Kết quả sinh luật bởi thuật toán C4.5.....	74

### DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Viết đầy đủ	Ghi chú
1	PLDL	Phân lớp dữ liệu	
2	CSDL	Cơ sở dữ liệu	
3	KPDL	Khai phá dữ liệu	
4	AGNES	Agglomerative Nesting	Thuật toán tích đồng lồng
5	BIRCH	Blanced Iterative Reducing and Clustering using Hieachies	
6	CF	Clustering Feature	Đặc trưng của phân cụm
7	DBSCAN	Density Based Spatial Clustering of Application with Noise	
8	OPTICS	Ordering Point to Identify the Clustering Structure	
9	PAM	Partitioning Around Medoids	
10	ID3	Iterative Decision 3	
11	NBC	Native Bayes Classification	Phân lớp dữ liệu Naive Bayes
12	FP	Frequent Pattern	Mẫu thường xuyên



## MỞ ĐẦU

Trong thời gian gần đây, sự phát triển mạnh mẽ của công nghệ thông tin, thương mại điện tử vào nhiều lĩnh vực của đời sống, kinh tế xã hội đã sinh ra một lượng dữ liệu lưu trữ khổng lồ. Sự bùng nổ này đã dẫn tới một nhu cầu cấp thiết cần có những kỹ thuật và công cụ để tự động chuyển đổi dữ liệu thành các tri thức có ích mà các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống không còn đáp ứng được nữa. Trong những khuynh hướng kỹ thuật mới có kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD – Knowledge Discovery and Data Mining). Nhưng để có thể khai phá dữ liệu một cách hiệu quả và chính xác, ta cần có những mô hình toán học, các giải thuật đáp ứng được điều đó. Vì vậy, trong luận văn này có trình bày một số vấn đề về phân cụm, phân lớp dữ liệu một trong những kỹ thuật cơ bản để khai phá dữ liệu nhưng lại được sử dụng rộng rãi và đem lại hiệu quả cao.

### **Bố cục của luận văn**

Nội dung chính của luận văn được chia thành 4 chương như sau:

*Chương 1. Tổng quan:* Chương này giới thiệu một cách tổng quát về quá trình phát hiện tri thức nói chung và khai phá dữ liệu nói riêng. Đặc biệt, chương trình còn liệt kê một số điểm chính về ứng dụng cũng như thách thức của khai phá dữ liệu và phát hiện tri thức.

*Chương 2. Phân cụm dữ liệu và một số thuật toán cơ bản:* Chương này trình bày các nội dung chính liên quan đến phân cụm dữ liệu. Một số thuật toán phân cụm dữ liệu cơ bản cũng được trình bày chi tiết trong chương này.

*Chương 3. Phân lớp dữ liệu và một số thuật toán cơ bản:* Chương này trình bày các nội dung chính liên quan đến phân lớp dữ liệu và ứng dụng. Một số

thuật toán phân lớp dữ liệu bao gồm: ID3, C.4.5, Naive Bayes, Apriori, ... cũng sẽ được trình bày chi tiết trong chương này.

*Chương 4. Một số kết quả thử nghiệm:* Chương này trình bày và phân tích một số kết quả thử nghiệm các thuật toán phân cụm, phân lớp dữ liệu cơ bản. Kết quả phân tích chủ yếu được triển khai thực hiện dựa trên phần mềm Weka (Waikato Environment for Knowledge Analysis) - một bộ phần mềm học máy được trường Đại học Waikato, New Zealand phát triển bằng Java. Weka là phần mềm tự do phát hành theo Giấy phép Công cộng GNU, hiện đang được sử dụng rất rộng rãi bởi cộng đồng những người làm về lĩnh vực khai phá dữ liệu và phát hiện tri thức.