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

DANH MỤC CÁC HÌNH VẼ

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 đống lồng
5	BIRCH	Blanced Iterative Reducing and Clustering using Hỉeachies	
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	Interative 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.