ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LÊ TRƯỜNG GIANG

PHƯƠNG PHÁP GIA TĂNG RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH SỬ DỤNG ĐỘ ĐO KHOẢNG CÁCH

LUẬN VĂN THẠC SĨ KỸ THUẬT

Thái Nguyên - 2014

LỜI CẢM ƠN

Lời cảm ơn trân trọng đầu tiên em muốn dành tới ***TS. Nguyễn Long Giang,*** người thầy đã dìu dắt và hướng dẫn tôi trong suốt quá trình làm luận văn, sự chỉ bảo và định hướng của thầy giúp tôi tự tin nghiên cứu những vấn đề mới và giải quyết bài toán một cách khoa học.

Em xin trân trọng cảm ơn *Ban giám hiệu và các thầy cô Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái nguyên* đã tạo các điều kiện cho chúng tôi được học tập và làm khóa luận một cách thuận lợi.

Lời cảm ơn sâu sắc muốn được gửi tới các thầy giáo *Viện Công nghệ Thông tin - Viện hàn lâm khoa học và Công nghệ Việt Nam*, những người thầy đã dạy dỗ và mở ra cho chúng tôi thấy chân trời tri thức mới, hướng dẫn chúng tôi cách khám phá và làm chủ công nghệ mới.

Xin được cảm ơn Trung tâm Quản lý Chất lượng – Trường Đại học Công nghiệp Hà Nội đã tạo mọi điều kiện để tôi được đi học và hoàn thành tốt khoá học

Mặc dù đã cố gắng rất nhiều, nhưng chắc chắn trong quá trình học tập cũng như luận văn không khỏi những thiếu sót. Em rất mong được sự thông cảm và chỉ bảo tận tình của các thầy cô và các bạn.

Thái Nguyên, tháng ..... năm 2014

Lê Trường Giang

MỤC LỤC

Danh mục các thuật ngữ

Thuật ngữ tiếng Việt	**Thuật ngữ tiếng Anh**
Tập thô	*Rough Set*
Hệ thông tin	*Information System*
Bảng quyết định	*Decision Table*
Bảng quyết định nhất quán	*Consistent Decision Table*
Bảng quyết định không nhất quán	*Inconsistent Decision Table*
Quan hệ không phân biệt được	*Indiscernibility Relation*
Xấp xỉ dưới	*Lower Approximation*
Xấp xỉ trên	*Upper Approximation*
Rút gọn thuộc tính	*Attribute Reduction*
Tập rút gọn	*Reduct*
Tập lõi	*Core*
Ma trận phân biệt	*Indiscernibility Matrix*
Hàm phân biệt	*Indiscernibility Function*
Luật quyết định	*Decision Rule*
Khoảng cách	*Distance*

Bảng các ký hiệu, từ viết tắt

Ký hiệu, từ viết tắt	Diễn giải
$IS = U, A, V, f$	Hệ thông tin
$DS = U, C \cup D, V, f$	Bảng quyết định
$\|U\|$	Số đối tượng
$\|C\|$	Số thuộc tính điều kiện trong bảng quyết định
$\|A\|$	Số thuộc tính trong hệ thông tin
$u\ a$	Giá trị của đối tượng u tại thuộc tính a
$IND\ B$	Quan hệ $B-$ không phân biệt
u_B	Lớp tương đương chứa u của quan hệ $IND\ B$
U / B	Phân hoạch của U sinh bởi tập thuộc tính B.
$\underline{B}X$	$B-$xấp xỉ dưới của X
$\overline{B}X$	$B-$xấp xỉ trên của X
$BN_B\ X$	B - miền biên của X
$POS_B\ D$	$B-$miền dương của D
$RED\ C$	Họ tất cả các tập rút gọn của bảng quyết định
$CORE\ C$	Tập lõi của bảng quyết định
$K\ P$	Tri thức sinh bởi tập thuộc tính P trong hệ thông tin.

Danh sách bảng

MỞ ĐẦU

Lựa chọn thuộc tính, còn gọi là trích chọn đặc trưng, là một trong những bài toán quan trọng trong khai phá dữ liệu và học máy. Lựa chọn thuộc tính sử dụng lý thuyết tập thô [9] được gọi là rút gọn thuộc tính. Rút gọn thuộc tính trong bảng quyết định là bài toán tìm tập con nhỏ nhất của tập thuộc tính điều kiện mà bảo toàn thông tin phân lớp của bảng quyết định, gọi là tập rút gọn. Trong hai thập kỷ trở lại đây, chủ đề nghiên cứu về rút gọn thuộc tính theo tiếp cận lý thuyết tập thô đã thu hút đông đảo cộng đồng nghiên cứu về tập thô tham gia [1]. Có rất nhiều phương pháp rút gọn thuộc tính khác nhau đã được đề xuất sử dụng các độ đo khác nhau như miền dương, ma trận phân biệt, các độ đo entropy trong lý thuyết thông tin, các độ đo trong tính toán hạt, độ đo khoảng cách. Tuy nhiên, hầu hết các nghiên cứu về rút gọn thuộc tính đều được thực hiện trên các bảng quyết định với tập đối tượng và tập thuộc tính cố định, không thay đổi. Trong thực tế, các bảng quyết định luôn bị cập nhật và thay đổi với các trường hợp: bổ sung hoặc loại bỏ tập đối tượng, bổ sung hoặc loại bỏ tập thuộc tính, cập nhật tập đối tượng đã tồn tại. Mỗi khi thay đổi như vậy, chúng ta lại phải thực hiện lại các thuật toán tìm tập rút gọn trên toàn bộ tập đối tượng, do đó chi phí về thời gian thực hiện thuật toán tìm tập rút gọn sẽ rất lớn.

Trong mấy năm gần đây, một số công trình nghiên cứu đã xây dựng các phương pháp gia tăng rút gọn thuộc tính trên bảng quyết định thay đổi dựa trên các độ đo khác nhau [3, 4, 6, 10, 11, 12]. Trong [3, 4, 12], các tác giả đã xây dựng phương pháp gia tăng tìm tập rút gọn dựa trên miền dương và ma trận phân biệt khi bổ sung tập đối tượng mới. Trong [10], các tác giả đã xây dựng các công thức tính các độ đo entropy (entropy Shannon, entropy Liang, entropy kết hợp) khi bổ sung, loại bỏ các thuộc tính. Tuy nhiên, các công thức tính toán entropy trong [10] còn phức tạp. Về hướng tiếp cận rút gọn thuộc

tính sử dụng độ đo khoảng cách được định nghĩa qua các khái niệm trong lý thuyết tập thô, trong [1, 7] tác giả đã sử dụng độ đo khoảng cách Jaccard để giải quyết bài toán rút gọn thuộc tính trong bảng quyết đinh. Tuy nhiên, tác giả trong [1, 7] mới giải quyết bài toán rút gọn thuộc tính trong trường hợp bảng quyết định cố định, không thay đổi.

Mục tiêu của luận văn là xây dựng phương pháp rút gọn thuộc tính trong bảng quyết định thay đổi dựa vào độ đo khoảng cách trong hai trường hợp: bổ sung đối tượng mới và loại bỏ đối tượng đã có.

Đối tượng nghiên cứu của luận văn là các bảng quyết định với dữ liệu thay đổi khi bổ sung và loại bỏ các đối tượng.

Phạm vi nghiên cứu: Với công cụ là lý thuyết tập thô, đề tài tập trung nghiên cứu phương pháp gia tăng tìm tập rút gọn của bảng quyết định khi bổ sung và loại bỏ tập đối tượng.

Phương pháp nghiên cứu của đề tài là nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

Về nghiên cứu lý thuyết: Nghiên cứu các kết quả đã công bố và xây dựng các công thức tính toán gia tăng khi bổ sung và loại bỏ đối tượng, trên cơ sở đó đề xuất các thuật toán hiệu quả.

Về nghiên cứu thực nghiệm: Cài đặt và thử nghiệm các thuật toán, các thuật toán gia tăng tìm tập rút gọn sử dụng khoảng cách trên các bộ số liệu mẫu lấy từ kho dữ liệu UCI [14] nhằm đánh giá tính hiệu quả của phương pháp gia tăng so với phương pháp truyền thống.

Bố cục của luận văn gồm phần mở đầu, ba chương nội dung, phần kết luận và các mục tài liệu tham khảo.

Chương 1: Trình bày một số khái niệm cơ bản trong lý thuyết tập thô và các kết quả nghiên cứu về các phương pháp rút gọn thuộc tính trong bảng

quyết định theo tiếp cận heuristic, các kết quả nghiên cứu về phân nhóm, so sánh và đánh giá các phương pháp.

Chương 2: Trình bày các bước xây dựng phương pháp rút gọn thuộc tính sử dụng độ đo khoảng cách, bao gồm định nghĩa độ đo khoảng cách, định nghĩa tập rút gọn và độ quan trọng của thuộc tính dựa trên khoảng cách và thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng khoảng cách. Trên cơ sở đó, chương 2 trình bày nội dung chính là xây dựng thuật toán tìm tập rút gọn của bảng quyết định thay đổi trong trường hợp bổ sung và loại bỏ đối tượng theo hướng tiếp cận tính toán gia tăng.

Chương 3: Trình bày kết quả thử nghiệm và đánh giá các thuật toán tìm tập rút gọn theo hướng tiếp cận gia tăng trong trường hợp bổ sung và loại bỏ đối tượng. So sánh kết quả thực hiện so với các phương pháp truyền thống là tính toán lại tập rút gọn trên toàn bộ tập đối tượng để thấy rõ tính hiệu quả của phương pháp gia tăng.

Phần kết luận: Tóm tắt kết quả đạt được của luận văn và hướng phát triển tiếp theo của tác giả luận văn.