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

VŨ VĂN TIỆP

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN GIA TĂNG CHO VIỆC RÚT GỌN CÁC THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ

LUẬN VĂN THẠC SĨ KHOA HỌC
KHOA HỌC MÁY TÍNH

HƯỚNG DẪN: GS.TS VŨ ĐỨC THI

THÁI NGUYÊN 2015

LỜI CẢM ƠN

Em xin chân thành cảm ơn và biết ơn sâu sắc đến GS.TS Vũ Đức Thi, Viện Công nghệ thông tin – Đại học Quốc gia Hà Nội. Người đã tận tình hướng dẫn và giúp đỡ em hoàn thành luận văn này.

Em xin chân thành cảm ơn các Thầy ở Viện Công nghệ thông tin đã dạy bảo, giúp đỡ và truyền đạt kiến thức cho em trong suốt khóa học và quá trình em làm luận văn.

Em xin chân thành cảm ơn các Thầy, các Cô ở trường Đại học Công nghệ thông tin và truyền thông Thái Nguyên đã tận tình dạy bảo, động viên, giúp đỡ và tạo điều kiện cho em trong suốt thời gian học tập và nghiên cứu.

Cuối cùng xin chân thành cảm ơn bạn bè, người thân và gia đình luôn là người đồng hành, động viên, chhia sẻ những khó khăn trong suốt thời gian hoàn thành luận văn.

Học viên

Vũ Văn Tiệp

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài "*Nghiên cứu một số thuật toán gia tăng cho việc rút gọn các thuộc tính trong bảng quyết định không đầy đủ"* là công trình nghiên cứu được tôi thực hiện dưới sự hướng dẫn của giáo viên hướng dẫn khoa học.

Một số Định nghĩa, Định lý, Tính chất, Mệnh đề và Thuật toán tôi lấy từ nguồn tài liệu chính xác có trích dẫn tên tài liệu và tên tác giả rõ ràng. Tôi xin chịu trách nhiệm về luận văn của mình.

Học viên

Vũ Văn Tiệp

MỤC LỤC

Danh mục các thuật ngữ

Thuật ngữ tiếng Việt	Thuật ngữ tiếng Anh
Tập thô	*Rough Set*
Tập thô dung sai	*Tolerance Rough Set*
Hệ thông tin	*Information System*
Hệ thông tin đầy đủ	*Complete Information System*
Hệ thông tin không đầy đủ	*Incomplete Information System*
Bảng quyết định	*Decision Table*
Bảng quyết định đầy đủ	*Complete Decision Table*
Bảng quyết định không đầy đủ	*Incomplete Decision Table*
Quan hệ không phân biệt được	*Indiscernibility Relation*
Quan hệ dung sai	*Tolerance Relation*
Xấp xỉ dưới	*Lower Approximation*
Xấp xỉ trên	*Upper Approximation*
Rút gọn thuộc tính	*Attribute Reduction*
Tập rút gọn	*Reduct*
Tập lõi	*Core*

Danh sách bảng

MỞ ĐẦU

Lý thuyết tập thô - do Zdzislaw Pawlak [10] đề xuất vào những năm đầu thập niên tám mươi của thế kỷ hai mươi - được xem là công cụ hữu hiệu để giải quyết các bài toán phân lớp, phát hiện luật…chứa dữ liệu không đầy đủ, không chắc chắn. Từ khi xuất hiện, lý thuyết tập thô đã được sử dụng hiệu quả trong các bước của quá trình khai phá dữ liệu và khám phá tri thức, bao gồm tiền xử lý số liệu, khai phá dữ liệu và đánh giá kết quả thu được. Rút gọn thuộc tính và trích lọc luật quyết định (luật phân lớp) là hai ứng dụng chính của lý thuyết tập thô trong khai phá dữ liệu. Rút gọn thuộc tính thuộc giai đoạn tiền xử lý dữ liệu còn trích lọc luật thuộc giai đoạn khai phá dữ liệu. Mục tiêu của rút gọn thuộc tính là loại bỏ các thuộc tính dư thừa nhằm tím tập con nhỏ nhất của tập thuộc tính điều kiện (tập rút gọn) mà bảo toàn thông tin phân lớp của bảng quyết định. Dựa trên tập rút gọn thu được, việc sinh luật và phân lớp đạt hiệu quả cao nhất.

Trong các bài toán thực tế, các bảng quyết định thường thiếu giá trị trên miền giá trị thuộc tính, gọi là các bảng quyết định không đầy đủ. Trên bảng quyết định không đầy đủ, Kryszkiewicz [5] đã mở rộng quan hệ tương đương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và đề xuất mô hình tập thô dung sai nhằm trích lọc luật trực tiếp không qua bước xử lý giá trị thiếu. Dựa trên mô hình tập thô dung sai, một số công trình công bố trong mấy năm gần đây đã đề xuất một số độ đo không chắc chắn nhằm giải quyết bài toán rút gọn thuộc tính và trích lọc luật, đáng chú ý là các công bố được liệt kê trong công trình [8].

Luận văn đặt ra hai mục tiêu chính:

1) Tổng hợp các công bố về các phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô dung sai, trên cơ

sở đó nghiên cứu phương pháp gia tăng rút gọn thuộc tính sử dụng hàm phân biệt mở rộng trong trường hợp bổ sung, loại bỏ tập thuộc tính. Bao gồm:

- Nghiên cứu phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng hàm phân biệt mở rộng, gồm các bước: xây dựng hàm phân biệt mở rộng; định nghĩa tập rút gọn và độ quan trọng của thuộc tính dựa trên hàm phân biệt mở rộng; xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng hàm phân biệt mở rộng; phân nhóm phương pháp sử dụng hàm phân biệt mở rộng.

- Nghiên cứu hướng tiếp cận gia tăng rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng hàm phân biệt mở rộng trong trường hợp bổ sung, loại bỏ tập thuộc tính.

2) Cài đặt thuật toán rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng hàm phân biệt mở rộng và các thuật toán gia tăng trong trường hợp bổ sung, loại bỏ tập thuộc tính. Thử nghiệm và đánh giá kết quả trên các bộ số liệu từ kho dữ liệu UCI.

Đối tượng nghiên cứu của luận văn là các *bảng quyết định không đầy đủ* khi bổ sung, loại bỏ tập thuộc tính.

Phạm vi nghiên cứu của luận văn *tập trung vào bài toán rút gọn thuộc tính* ở bước tiền xử lý số liệu trong quá trình khai phá dữ liệu.

Phương pháp nghiên cứu của luận văn là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Về nghiên cứu lý thuyết: tổng hợp và nắm bắt các kết quả nghiên cứu đã công bố. Về nghiên cứu thực nghiệm: luận văn thực hiện cài đặt các thuật toán, chạy thử nghiệm thuật toán với các bộ số liệu lấy từ kho dữ liệu UCI [13], so sánh và đánh giá nghiên cứu thực nghiệm với nghiên cứu lý thuyết.

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

Chương 1 trình bày các khái niệm cơ bản về lý thuyết tập thô của Pawlak [10] và mô hình tập thô mở rộng dựa trên quan hệ dung sai, gọi tắt là mô hình tập thô dung sai [5]. Trình bày tổng quan các kết quả nghiên cứu về các phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô dung sai.

Chương 2 trình bày hai nội dung chính:

- Thứ nhất là phương pháp rút gọn thuộc tính sử dụng hàm phân biệt mở rộng [14], bao gồm: xây dựng hàm phân biệt mở rộng; định nghĩa tập rút gọn và độ quan trọng của thuộc tính dựa trên hàm phân biệt mở rộng; xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng hàm phân biệt mở rộng; phân nhóm phương pháp sử dụng hàm phân biệt mở rộng.

- Thứ hai là xây dựng thuật toán theo hướng tiếp cận gia tăng tìm tập rút gọn của bảng quyết định không đầy đủ sử dụng hàm phân biệt mở rộng trong trường hợp bổ sung, loại bỏ tập thuộc tính [14].

Chương 3 trình bày kết quả thử nghiệm và đánh giá các thuật toán: bao gồm thuật toán tìm tập rút gọn sử dụng hàm phân biệt mở rộng và thuật toán gia tăng tìm tập rút gọn sử dụng hàm phân biệt mở rộng trong trường hợp bổ sung và loại bỏ tập thuộc tính. Thử nghiệm được thực hiện trên các bộ số liệu mẫu từ kho dữ liệu UCI [13].

Cuối cùng, phần kết luận nêu những đóng góp của luận văn và hướng phát triển tiếp theo.