ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN QUỲNH LAN

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ

Chuyên ngành: Khoa học máy tính

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên, 2013

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, Viện Khoa học và Công nghệ Việt Nam. Người đã tận tình dày công 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, trong suốt cả 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 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 bàn bè, người thân và gia đình luôn là người đồng hành, động viên, chia sẻ những khó khăn trong suốt thời gian hoàn thành luận văn.

Thái Nguyên, tháng 08 năm 2013

Nguyễn Quỳnh Lan

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này là sản phẩm tìm hiểu, nghiên cứu của mình. 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

Nguyễn Quỳnh Lan

MỤC LỤC

Danh mục các thuật ngữ

Thuật ngữ tiếng việt	**Thuật ngữ tiếng anh**
Tập thô	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
Xấp xỉ dưới	Lower Approximation
Xấp xỉ trên	Upper Lower 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

Bảng các ký hiệu, từ viết tắt

Ký hiệu, từ viết tắt	Diễn giải
IS = (U, A, V, f)	Hệ thông tin, hệ thông tin đầy đủ
IIS = (U, A, V, f)	Hệ thông tin không đầy đủ
DS = (U, C∪D, V, f)	Bảng quyết định, bảng quyết định đầy đủ
IDS = (U, C∪D, V, f)	Bảng quyết định không đầy đủ
|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
SIM(B)	Quan hệ dung sai trên tập thuộc tính B
$[u]_B$	Lớp tương đương chứa u của quan hệ IND(B)
$S_B(u)$	Lớp dung sai của đối tượng u trên quan hệ SIM(B)
U/B	Phân hoạch của U sinh bởi tập thuộc tính B
U/SIM(B)	Phủ của U sinh bởi tập thuộc tính B
COVER(U)	Họ tất cả các phủ của U
$\partial_B(u)$	Hàm quyết định suy rộng của đối tượng u đối với 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
PRED(C)	Họ tất cả các tập rút gọn Pawlak
SRED(C)	Họ tất cả các tập rút gọn sử dụng ma trận phân biệt
MRED(C)	Họ tất cả các tập rút gọn dựa trên metric
PCORE (C)	Tập lõi dựa trên miền dương
SCORE(C)	Tập lõi sử dụng ma trận phân biệt
MCORE(C)	Tập lõi dựa trên metric
H(P)	Entropy Shannon của tập thuộc tính P
H(Q/P)	Entropy Shannon có điều kiện của Q khi đã biết P
IE(P)	Entropy Liang mở rộng của tập thuộc tính P trong hệ thông tin không đầy đủ
IE(Q/P)	Entropy Liang mở rộng có điều kiện của Q khi đã biết P trong hệ thông tin không đầy đủ
K(P)	Trong hệ thông tin đầy đủ:là tri thức sinh bởi tập thuộc tính P. Trong hệ thông tin không đầy đủ là phủ sinh bởi tập thuộc tính P
$d_j(K(P), K(Q))$	Khoảng cánh giữa K(P) và K(Q) trong hệ thông tin đầy đủ dựa trên khoảng cách Jaccard giữa hai tập hợp
$d_E(K(P), K(Q))$	Khoảng cánh giữa K(P) và K(Q) trong hệ thông tin không đầy đủ dựa trên entropy Liang mở rộng
$SIG_B(b)$	Độ quan trọng của thuộc tính b đối với B

DANH SÁCH BẢNG

MỞ ĐẦU

Mười năm trở lại đây chúng ta đã chứng kiến sự phát triển mạnh mẽ và sôi động của lĩnh vực nghiên cứu về rút gọn thuộc tính sử dụng lý thuyết tập thô. Trong xu thế đó, nhiều nhóm nhà khoa học trên thế giới quan tâm nghiên cứu các phương pháp rút gọn thuộc tính trong bảng quyết định. Các phương pháp chính là: Phương pháp dựa trên miền dương, phương pháp sử dụng các phép toán trong đại số quan hệ, phương pháp sử dụng ma trận phân biệt, phương pháp sử dụng entropy thông tin, phương pháp sử dụng các độ đo trong tính toán hạt...

Mục tiêu của rút gọn thuộc tính là loại bỏ các thuộc tính dư thừa để tìm ra các thuộc tính cốt yếu và cần thiết trong cơ sở dữ liệu. Với bảng quyết định không đầy đủ rút gọn thuộc tính là tìm tập con nhỏ nhất của tập thuộc tính điều kiện bảo đảm thông tin phân lớp của bảng quyết định đó. Đối với một bảng quyết định không đầy đủ có thể có nhiều tập rút gọn khác nhau. Tuy nhiên, trong thực hành thường không đòi hỏi tìm tất cả các tập rút gọn mà chỉ cần tìm được một tập rút gọn theo một tiêu chuẩn đánh giá nào đó là đủ.

Các kết quả nghiên cứu cho thấy rút gọn thuộc tính làm giảm thiểu đáng kể khối lượng tính toán, nhờ đó có thể áp dụng đối với các bài toán có khối lượng dữ liệu lớn. Thuật toán khá đơn giản về mặt thực thi. Nên em quyết định lựa chọn đề tài luận văn: "**Nghiên cứu một số phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ**".

Mục tiêu của luận văn: Tập trung nghiên cứu rút gọn thuộc tính trong bảng quyết định đầy đủ từ đó làm cơ sở nghiên cứu tiếp phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ.

Đối tượng và phạm vi nghiên cứu: Các bảng quyết định đầy đủ, các bảng quyết định không đầy đủ với kích thước trung bình và lớn.

Phương pháp nghiên cứu

- Về nghiên cứu lý thuyết: Các Định lý, Mệnh đề…đã được chứng minh dựa vào các kiến thức cơ bản và các kết quả nghiên cứu đã công bố.

- Về nghiên cứu thực nghiệm: Cài đặt thuật toán, chạy thử nghiệm thuật toán.

Ý nghĩa khoa học của đề tài

-Đây là phương pháp được nhiều nhà khoa học nghiên cứu và đã có đóng góp trong thực tiễn.

-Có thể coi luận văn là một tài liệu tham khảo khá đầy đủ, rõ ràng về các kiến thức cơ bản trong bảng quyết định không đầy đủ.

Bố cục của luận văn: Gồm phần mở đầu và 3 chương nội dung, phần kết luận, danh mục tài liệu tham khảo và phụ lục.

Chương 1: Trình bày các khái niệm cơ bản về bảng quyết định đầy đủ, bảng quyết định không đầy đủ, mô hình tập thô truyền thống, mô hình tập thô dung sai, trình bày phương pháp xây dựng 1 metric trên họ các tri thức trong hệ thông tin đầy đủ dựa trên khoảng cách Jaccard giữa hai tập hợp hữu hạn, trình bày phương pháp rút gọn thuộc tính trong bảng quyết định đầy đủ.

Chương 2: Trình bày phương pháp xây dựng một metric trên họ các phủ trong hệ thông tin không đầy đủ dựa trên entropy Liang mở rộng, trình bày phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ.

Chương 3: Chương trình thử nghiệm trình bày các nội dung: mô tả dữ liệu, xây dựng chương trình, và kết quả thực nghiệm của thuật toán.

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