ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

Hoàng Hà Đức

KHAI PHÁ DỮ LIỆU SỬ DỤNG GIẢI THUẬT DI TRUYỀN VÀ ỨNG DỤNG

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học
TS. Nguyễn Huy Đức

Thái Nguyên - 2016

LỜI CAM ĐOAN

Sau quá trình học tập tại **Trường Đại học công nghệ thông tin & truyền thông**, với những kiến thức lý thuyết và thực hành đã tích lũy được, với việc vận dụng các kiến thức vào thực tế, em đã tự nghiên cứu các tài liệu, các công trình nghiên cứu, đồng thời có sự phân tích, tổng hợp, đúc kết và phát triển để hoàn thành luận văn thạc sĩ của mình.

Em xin cam đoan luận văn này là công trình do bản thân em tự tìm hiểu, nghiên cứu và hoàn thành dưới sự hướng dẫn tận tình của thầy giáo **TS. Nguyễn Huy Đức**.

Thái Nguyên, tháng 6 năm 2016

Học viên

Hoàng Hà Đức

LỜI CÁM ƠN

Trong thời gian hai năm của chương trình đào tạo thạc sỹ, trong đó gần một nửa thời gian dành cho các môn học, thời gian còn lại dành cho việc lựa chọn đề tài, giáo viên hướng dẫn, tập trung vào nghiên cứu, viết, chỉnh sửa và hoàn thiện đề tài. Với quỹ thời gian như vậy và với vị trí công việc đang phải đảm nhận, không riêng bản thân em mà hầu hết các sinh viên cao học muốn hoàn thành tốt luận văn của mình trước hết đều phải có sự sắp xếp thời gian hợp lý, có sự tập trung học tập và nghiên cứu với tinh thần nghiêm túc, nỗ lực hết mình; tiếp đến cần có sự ủng hộ về tinh thần, sự giúp đỡ về chuyên môn một trong những điều kiện không thể thiếu quyết định đến việc thành công của đề tài.

Để hoàn thành được đề tài này trước tiên em xin gửi lời cảm ơn đến thầy giáo hướng dẫn **TS. Nguyễn Huy Đức**, người đã có những định hướng cho em về nội dung và hướng phát triển của đề tài, người đã có những đóng góp quý báu cho em về những vấn đề chuyên môn của đề tài, giúp em tháo gỡ kịp thời những vướng mắc trong quá trình làm luận văn.

Em cũng xin cảm ơn các Thầy Cô giáo Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên, cũng như bạn bè cùng lớp đã có những ý kiến đóng góp bổ sung cho đề tài luận văn của em. Xin cảm ơn gia đình, người thân cũng như đồng nghiệp luôn quan tâm, ủng hộ hỗ trợ về mặt tinh thần trong suốt thời gian từ khi nhận đề tài đến khi hoàn thiện đề tài này.

Trong nội dung của luận văn chắc chắn còn nhiều thiếu sót. Em rất mong các Thầy Cô cùng bạn bè đóng góp để bản luận văn của Em được hoàn thiện hơn.

Em xin trân trọng cảm ơn.

Thái Nguyên, tháng 6 năm 2016

Học viên

Hoàng Hà Đức

MỤC LỤC

DANH SÁCH HÌNH VẼ

DANH SÁCH TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa	
KPDL	Khai phá dữ liệu	
KPTT	Khai phá tri thức	
PCDL	Phân cụm dữ liệu	
CSDL	Cơ sở dữ liệu	
GA	Giải thuật di truyền	Genetic Algorithm
DE	Giải thuật tiến hóa vi phân	Differential Evolution
NST	Nhiễm sắc thể	
CDL	Cụm dữ liệu	
CNTT	Công nghệ thông tin	

MỞ ĐẦU

Phân cụm dữ liêu là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liêu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ liêu không đòi hỏi phải định nghĩa trước các mẫu dữ liêu huấn luyện. Vì thế, có thể coi phân cụm dữ liêu là một cách học không giám sát (unsupervised learning). Các Kỹ thuật phân cụm được ứng dụng rất nhiều trong các lĩnh vực tài chính ngân hành để phân loại các nhóm khách hàng khác nhau. Ngoài ra phân cụm dữ liêu còn có thể được sử dụng như một bước tiền xử lý cho các giải thuật khai phá dữ liệu khác như phân loại và mô tả đặc điểm, có tác dụng phát hiện ra các cụm.

Trong ngành khoa học máy tính, tìm kiếm lời giải tối ưu cho các bài toán là vấn đề được các nhà khoa học máy tính đặc biệt rất quan tâm. Mục đích chính của các thuật toán là tìm kiếm thuật giải chất lượng cao và sử dụng kỹ thuật trí tuệ nhân tạo đặc biệt rất cần thiết khi giải quyết các bài toán có không gian tìm kiếm lớn.

Giải thuật di truyền (Genetic Algorithm GA) là một trong những kỹ thuật tìm kiếm lời giải tối ưu đã đáp ứng được yêu cầu của nhiều bài toán và ứng dụng. Hiện nay, thuật toán di truyền được ứng dụng rất rộng rãi trong các lĩnh vực phức tạp. Thuật toán di truyền chứng tỏ được hiệu quả của nó trong các vấn đề khó có thể giải quyết bằng các phương pháp thông thường hay các phương pháp cổ điển, nhất là trong các bài toán cần có sự lượng giá, đánh giá sự tối ưu của kết quả thu được.

Chính vì vậy, trong phạm vi đề tài này, tôi chọn hướng phân cụm dữ liệu dựa trên giải thuật di truyền.

Luận văn gồm có 3 chương:

Chương I: Tổng quan về khai phá dữ liệu và phân cụm dữ liệu

Phần này giới thiệu một cách tổng quát về quá trình khám phá tri thức nói chung và khai phá dữ liệu nói riêng. Các phương pháp khai phá dữ liệu và phân cụm dữ liệu.

Chương II: Thuật toán phân cụm dữ liệu dựa trên giải thuật di truyền.

Trong chương này trình bày giải thuật di truyền, thuật toán phân cụm sử dụng giải thuật di truyền và so sánh hiệu quả của thuật toán Kmeans và thuật toán Kmeans sử dụng giải thuật di truyền.

Chương III: Thực nghiệm phân cụm dữ liệu về sinh viên của trường Cao đẳng Y tế Yên Bái.

Phần này mô tả bài toán, xây dựng chương trình. Cài đặt chương trình thử nghiệm ứng dụng kỹ thuật phân cụm trong công tác học sinh sinh viên của Trường Cao đẳng Y tế Yên Bái và một kết quả thu được.