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

TRỊNH ANH TUẤN

PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG PHÂN CỤM SINH HỌC TRƯỜNG TRUNG HỌC CƠ SỞ CHU VĂN AN

LUẬN VĂN THẠC SĨ
KHOA HỌC MÁY TÍNH

Người hướng dẫn: TS.Nguyễn Long Giang

THÁI NGUYÊN - 2016

LỜI CAM ĐOAN

Tác giả Trịnh Anh Tuấn xin cam kết rằng nội dung của Luận văn này chưa được nộp cho bất kỳ một chương trình cấp bằng cao học nào cũng như bất kỳ một chương trình đào tạo cấp bằng nào khác.

Ngoài ra, tác giả cũng xin cam kết Luận văn thạc sĩ này là nỗ lực riêng của cá nhân tác giả. Các kết quả, phân tích, kết luận trong Luận văn thạc sĩ này (ngoài các phần được trích dẫn) đều là kết quả làm việc của cá nhân tác giả.

Thái Nguyên, tháng 6 năm 2016

Tác Giả

Trịnh Anh Tuấn

LỜI CẢM ƠN

Để hoàn thành được luận văn này, trước hết tôi xin gửi lời cảm ơn sâu sắc nhất tới TS. Nguyễn Long Giang, Viện Công nghệ thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tận tình hướng dẫn, chỉ bảo, định hướng, đóng góp những ý kiến quý báu trong suốt quá trình thực hiện luận văn.

Tôi xin chân thành cảm ơn các thầy, cô giáo trong Bộ môn Khoa học máy tính, Khoa Công nghệ thông tin, Phòng Đào tạo Sau đại học - Nghiên cứu Khoa học, Trường Đại học Công nghệ thông tin và truyền thông Thái Nguyên đã tạo mọi điều kiện tốt nhất để tôi hoàn thành khóa học.

Xin cám ơn đồng nghiệp tại trường Trung học cơ sở Chu Văn An thành phố Thái Nguyên đã trợ giúp rất nhiều trong thời gian qua.

Trong quá trình thực hiện Luận văn, mặc dù đã cố gắng hết mình, song chắc chắn luận văn của em vẫn còn nhiều thiếu sót. Em rất mong nhận được sự chỉ bảo vào đóng góp tận tình của các thầy cô để luận văn của em được hoàn thiện hơn.

Thái Nguyên, tháng 6 năm 2016

Tác Giả

Trịnh Anh Tuấn

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1.1. Sự cần thiết lựa chọn đề tài

Sự phát triển của nhanh chóng các ứng dụng công nghệ thông tin và Internet vào nhiều lĩnh vực đời sống xã hội, quản lý kinh tế, khoa học kỹ thuật... trong mấy năm gần đây đã tạo ra nhiều cơ sở dữ liệu khổng lồ. Để khai thác hiệu quả nguồn thông tin từ các cơ sở dữ liệu khổng lồ đó nhằm mục đích dự báo, hỗ trợ ra quyết định, bên cạnh các phương pháp khai thác thông tin truyền thống, các nhà nghiên cứu đã phát triển các phương pháp, kỹ thuật và phần mềm mới hỗ trợ tiến trình khám phá, phân tích, tổng hợp thông tin, lĩnh vực này được gọi là khai phá dữ liệu và khám phá tri thức (Data mining and Knowledge discovery)

Khai phá dữ liệu và khám phá tri thức là một lĩnh vực quan trọng của ngành Công nghệ thông tin với mục tiêu là tìm kiếm các tri thức có ích, cần thiết, tiềm ẩn và chưa được biết trước trong cơ sở dữ liệu lớn. Đây là lĩnh vực đã và đang thu hút đông đảo các nhà khoa học trên thế giới và trong nước tham gia nghiên cứu. Khai phá dữ liệu có thể xem là nhiệm vụ quan trọng trong quá trình khám phá tri thức từ cơ sở dữ liệu, bao gồm ba bước chính: thu thập và tiền xử lý dữ liệu; lựa chọn các thuật toán khai phá dữ liệu; đánh giá kết quả và biểu diễn tri thức. Các bài toán quan trọng trong khai phá dữ liệu bao gồm: phân lớp (classification); hồi quy (regression); phân cụm (clustering); khai phá luật kết hợp (rule association)… Các kỹ thuật, công cụ sử dụng trong khai phá dữ liệu bao gồm: cây quyết định; mạng nơron nhân tạo; thuật toán di truyền; các kỹ thuật phân lớp, phân cụm; các phương pháp trong thống kê như phân tích tương quan, phân tích phương sai, hồi quy đơn biến, đa biến... Khai phá dữ liệu có nhiều ứng dụng trong các lĩnh vực khác nhau của đời sống như: phân tích, dự báo trong kinh tế, tài chính; chuẩn đoán bệnh trong y tế; tin sinh học; hỗ trợ quá trình sản xuất, kinh doanh...

Phân cụm (clustering) là bài toán có vai trò quan trọng trong khai phá dữ liệu và có nhiều ứng dụng trong thực tiễn. Mục tiêu của phương pháp phân cụm dữ liệu là quá trình nhóm các đối tượng tương tự nhau trong cơ sở dữ liệu vào các cụm sao cho các đối tượng trong cùng một cụm là tương đồng, còn các đối tượng thuộc các

cụm khác nhau sẽ không tương đồng. Điểm mạnh của phân cụm dữ liệu là đưa ra được những cấu trúc có ích hoặc những cụm các đối tượng tìm thấy trực tiếp từ dữ liệu mà không cần bất kì một tri thức cơ sở nào. Giống như cách tiếp cận học máy, phân cụm dữ liệu được hiểu như là phương pháp học không có thầy (unsupervised learning). Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát (learning by observation), trong khi phân lớp dữ liệu là học bằng ví dụ (learning by example). Trong phương pháp này sẽ không thể biết kết quả các cụm thu được sẽ như thế nào khi bắt đầu quá trình. Vì vậy, cần có một chuyên gia để đánh giá các cụm thu được. Phân cụm dữ liệu được sử dụng nhiều trong các ứng dụng về phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, phân loại trang Web, phân loại, đánh giá học sinh, sinh viên trong các trường học… Ngoài ra, phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lí cho các thuật toán khai phá dữ liệu khác.

Là một người công tác trong lĩnh vực giáo dục phổ thông, với mong muốn áp dụng các kiến thức đã học về các phương pháp phân cụm vào bài toán thực tiễn là phân cụm học sinh của trường Trung học cơ sở Chu Văn An, thành phố Thái Nguyên dựa vào kết quả học tập, tác giả luận văn chọn đề tài: "*Phân cụm dữ liệu và ứng dụng phân cụm học sinh trường Trung học cơ sở Chu Văn An*".

1.2. Mục tiêu đề tài

Nắm bắt được một cách tổng thể các phương pháp phân cụm trong khai phá dữ liệu. Trên cơ sở đó, áp dụng các kỹ thuật phân cụm vào giải quyết bài toán thực tiễn tại địa phương nơi tác giả làm việc là phân cụm kết quả học tập của học sinh trường Trung học cơ sở Chu Văn An, thành phố Thái Nguyên.

1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các phương pháp phân cụm dữ liệu trong khai phá dữ liệu và cơ sở dữ liệu về kết quả học tập của học sinh trường Trung học cơ sở Chu Văn An, thành phố Thái Nguyên.