

**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 CẢM 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

| | |
|---|----|
| LỜI CAM ĐOAN | 1 |
| LỜI CẢM ƠN | 3 |
| CHƯƠNG 1 TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ PHÂN CỤM DỮ LIỆU | 11 |
| 1.1. Tổng quan về khám phá tri thức và khai phá dữ liệu..... | 11 |
| 1.1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu | 11 |
| 1.1.2. Quá trình khám phá tri thức. | 11 |
| 1.1.3. Quá trình khai phá dữ liệu..... | 13 |
| 1.2. Các phương pháp khai phá dữ liệu..... | 13 |
| 1.2.1. Phân lớp và dự đoán (<i>Classification & Prediction</i>)..... | 14 |
| 1.2.2. Luật kết hợp (<i>Association Rules</i>) | 14 |
| 1.2.3. Khai thác mẫu tuần tự (<i>Sequential / Temporal patterns</i>)..... | 14 |
| 1.2.4. Phân nhóm- đoạn (<i>Clustering / Segmentation</i>) | 15 |
| 1.2.5. Hồi quy (<i>Regression</i>) | 15 |
| 1.2.6. Tổng hợp hóa (<i>Summarization</i>) | 15 |
| 1.2.7. Mô hình hóa sự phụ thuộc (<i>dependency modeling</i>)..... | 16 |
| 1.2.8. Phát hiện sự biến đổi và độ lệch (<i>Change and deviation detection</i>).... | 16 |
| 1.3. Phân cụm dữ liệu..... | 16 |
| 1.3.1. Phân cụm dữ liệu là gì..... | 16 |
| 1.3.2. Các mục tiêu của phân cụm dữ liệu..... | 18 |
| 1.3.4. Các phương pháp phân cụm dữ liệu | 19 |
| 1.3.4.1. Phương pháp phân cụm phân cấp | 19 |
| 1.3.4.2. Phương pháp phân cụm dựa trên mật độ | 20 |
| 1.3.4.3. Phương pháp phân cụm phân hoạch | 21 |
| 1.3.4.4. Phương pháp phân cụm dựa trên lưới..... | 22 |
| 1.3.4.5. Phương pháp phân cụm dựa trên mô hình | 23 |
| 1.3.4.6. Phương pháp phân cụm có dữ liệu ràng buộc..... | 23 |
| CHƯƠNG 2: THUẬT TOÁN PHÂN CỤM DỮ LIỆU DỰA TRÊN GIẢI THUẬT DI TRUYỀN | 25 |
| 2.1. Giải thuật di truyền..... | 25 |

| | |
|---|----|
| 2.1.1. Lịch sử của giải thuật di truyền. | 25 |
| 2.1.2. Tóm tắt giải thuật di truyền..... | 25 |
| 2.1.3. Cách biểu diễn bài toán trong giải thuật di truyền (hay chọn cách biểu diễn cấu trúc dữ liệu cho bài toán) | 29 |
| 2.1.4. Mã hóa (encoding). | 35 |
| 2.1.5. Các phương pháp chọn(Selection)..... | 37 |
| 2.1.6. Chọn lọc Roulette (Roulette Wheel Selection). | 37 |
| 2.1.7. Các toán tử trong giải thuật di truyền | 41 |
| 2.1.8. Các tham số cần sử dụng trong giải thuật di truyền..... | 44 |
| 2.1.9. Điều kiện kết thúc thuật giải di truyền..... | 44 |
| 2.1.10. Nguyên lý hoạt động của giải thuật di truyền..... | 44 |
| 2.1.11. Ứng dụng của thuật giải di truyền. | 44 |
| 2.2. Thuật toán phân cụm sử dụng giải thuật di truyền..... | 44 |
| 2.2.1. Một số giải thuật cơ bản trong phân cụm dữ liệu..... | 44 |
| 2.2.2. Giải thuật phân cụm dựa trên giải thuật di truyền. | 57 |
| 2.3. So sánh hiệu quả của thuật toán Kmeans và thuật toán Kmeans sử dụng giải thuật di truyền..... | 61 |
| 2.3.1. Thuật Toán K-Means..... | 61 |
| 2.3.2. Thuật toán Kmean sử dụng giải thuật di truyền | 66 |
| 2.3.3. So sánh giữa k-means và k-means sử dụng giải thuật di truyền: | 69 |
| CHƯƠNG 3: THỰC NGHIỆM PHÂN CỤM DỮ LIỆU VỀ SINH VIÊN CỦA TRƯỜNG CAO ĐẲNG Y TẾ YÊN BÁI..... | 70 |
| 3.1. Mô tả bài toán..... | 70 |
| 3.1.1. Cơ sở dữ liệu. | 70 |
| 3.2. Xây dựng chương trình. | 71 |
| 3.2.2. Các chức năng của chương trình..... | 71 |
| 3.2.3. Giao diện chương trình..... | 71 |
| 3.2.3. Kết quả thực nghiệm. | 73 |
| KẾT LUẬN | 75 |
| TÀI LIỆU THAM KHẢO | 76 |
| PHẦN PHỤ LỤC..... | 78 |

DANH SÁCH HÌNH VẼ

| | |
|---|----|
| Hình 1.1: Quá trình khám phá tri thức | 12 |
| Hình 1.2: Quá trình khai phá dữ liệu..... | 13 |
| Hình 1.3: Ví dụ về phân cụm dữ liệu..... | 17 |
| Hình 1.4: Ví dụ phân cụm các ngôi nhà dựa trên khoảng cách..... | 18 |
| Hình 1.5: Ví dụ phân cụm các ngôi nhà dựa trên kích cỡ | 19 |
| Hình 1.6. Các chiến lược phân cụm phân cấp..... | 20 |
| Hình 1.7: Ví dụ về phân cụm theo mật độ (1)..... | 21 |
| Hình 1.8: Ví dụ về phân cụm theo mật độ (2)..... | 21 |
| Hình 1.9: Cấu trúc phân cụm dựa trên lưới..... | 22 |
| Hình 1.10: Ví dụ về phân cụm dựa trên mô hình..... | 23 |
| Hình 1.11: Các cách mà các cụm có thể đưa ra | 24 |
| Hình 2.1:Sơ đồ tổng quát của giải thuật di truyền | 28 |
| Hình 2.2: Nhiệm sắc thể bằng cây | 37 |
| Hình 2.2. Minh họa trường hợp tách dữ liệu thành 3 cụm..... | 45 |
| Hình 2.3. Khái quát giải thuật CURE | 48 |
| Hình 2.3. Các cụm dữ liệu được khám phá bởi CURE..... | 49 |
| Hình 2.4. Lân cận của P với ngưỡng Eps..... | 50 |
| Hình 2.5: Mật độ - đến được trực tiếp..... | 51 |
| Hình 2.6: Mật độ đến được | 51 |
| Hình 2.7: Mật độ liên thông | 51 |
| Hình 2.8: Cụm và nhiễu. | 52 |
| Hình 2.9: Hình dạng các cụm được khám phá bởi giải thuật DBSCAN | 53 |
| Hình 3.1. Cơ sở dữ liệu học sinh sinh viên | 71 |
| Hình 3.2. Giao diện chương trình | 71 |
| Hình 3.3. Màn hình khởi động | 73 |
| Hình 3.4. Màn hình phân cụm dữ liệu | 73 |

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 liệu là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu 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 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 không giám sát (unsupervised learning). Các Kỹ thuật phân cụm được ứng dụng rất nhiều trong các lĩnh vực tài chính ngân hàng để phân loại các nhóm khách hàng khác nhau. Ngoài ra phân cụm dữ liệu 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.