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

NGUYỄN MINH TÚ

TÌM HIỂU CÁC PHƯƠNG PHÁP CỤM DỮ LIỆU ỨNG DỤNG XÂY DỰNG BẢN ĐỒ PHÂN BỐ BỆNH TRÊN ĐỊA BÀN TỈNH THÁI NGUYÊN

Mã số: 60480101

Người hướng dẫn khoa học: TS. NGUYỄN MINH HẢI

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

LỜI CẢM ƠN

Em xin chân thành cảm ơn Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên đã tạo điều kiện cho em thực hiện luận văn này.

Em xin gửi lời cảm ơn sâu sắc tới thầy giáo TS Nguyễn Hải Minh, trưởng khoa Công nghệ thông tin – Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên đã trực tiếp hướng dẫn em trong quá trình thực hiện luận văn.

Em cũng xin gửi lời cảm ơn tới các thầy, cô đã có những ý kiến đóng góp bổ ích và đã tạo mọi điều kiện tốt nhất cho em trong suốt thời gian thực hiện luận văn. Xin cảm ơn các bạn học đồng khóa đã thường xuyên động viên, giúp đỡ tôi trong quá trình học tập.

Cuối cùng, em xin gửi lời cảm ơn đến gia đình và đồng nghiệp vì sự ủng hộ và động viên đã dành cho em trong suốt quá trình học tập cũng như thực hiện luận văn này.

Thái Nguyên, tháng 11 năm 2015

Học viên

Nguyễn Minh Tú

LỜI CAM ĐOAN

Em xin cam đoan về nội dung đồ án tốt nghiệp với tên đề tài "**Tìm hiểu các phương pháp phân cụm dữ liệu ứng dụng xây dựng bản đồ phân bố bệnh trên địa bàn tỉnh Thái Nguyên**" không sao chép nội dung từ các luận văn khác, hay các sản phẩm tương tự mà không phải do em làm ra. Sản phẩm luận văn là do chính bản thân em tìm hiểu và xây dựng nên.

Nếu có gì sai em xin chịu mọi hình thức kỷ luật của Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

Thái Nguyên, tháng 11 năm 2015

Học viên

Nguyễn Minh Tú

MỤC LỤC

LỜI CẢM ƠN i
LỜI CAM ĐOAN iii
MỤC LỤC iv
DANH MỤC BẢNG vi
DANH MỤC CÁC HÌNH VẼ vii
MỞ ĐẦU 1
MỘT SỐ KẾT QUẢ NGHIÊN CỨU KHÁC 2
CHƯƠNG I. KHAI PHÁ DỮ LIỆU 3
1.1. Tổng quan khai phá dữ liệu 3
1.2. Quá trình khám phá tri thức và khai phá dữ liệu 3
1.2.1. Quá trình khám phá tri thức 3
1.2.2. Quá trình khai phá dữ liệu 6
1.3. Các kỹ thuật trong khai phá dữ liệu 7
1.4. Một số thách thức trong khai phá dữ liệu 10
CHƯƠNG II. PHÂN CỤM DỮ LIỆU VÀ CÁC THUẬT TOÁNPHÂN CỤM 12
2.1. Khái niệm phân cụm dữ liệu 13
2.1.1. Một số định nghĩa 13
2.1.2. Dữ liệu và độ đo 13
2.2. Các yêu cầu đối với phương pháp phân cụm dữ liệu 14
2.3. Các kỹ thuật phân cụm 16
2.3.1. Các kỹ thuật phân cụm cơ bản 16
2.3.2. Các kỹ thuật phân cụm khác 17
2.3.3. Một số tiêu chuẩn đánh giá hiệu quả phân cụm 23
2.4. Một số thuật toán trong phân cụm dữ liệu 24
2.4.1. Các thuật toán phân cụm phân hoạch 24
2.4.2. Các thuật toán phân cụm phân cấp 32
2.4.3. Các thuật toán phân cụm dựa trên mật độ 38

2.4.3. Các thuật toán phân cụm dựa vào lưới 42
2.4.4. Các thuật toán phân cụm dựa trên mô hình 44
CHƯƠNG 3: XÂY DỰNG BẢN ĐỒ PHÂN BỐ BỆNH 46
3.1. Bài toán phân cụm dữ liệu hồ sơ bệnh án 46
3.2. Dữ liệu và tiêu chí xác định 46
3.2.1. Một số đặc điểm tập dữ liệu gốc 46
3.2.2. Tiền xử lý dữ liệu gốc 48
3.3. Lựa chọn phương pháp phân cụm 54
3.4. Kết quả phân cụm dữ liệu mẫu 55
3.5. Biểu điễn kết quả phân bố bệnh trên bản đồ 61
KẾT LUẬN 63
TÀI LIỆU THAM KHẢO 65

DANH MỤC BẢNG

Bảng 3.1: Các loại bệnh và ký hiệu 55
Bảng 3.2: Các khu vực hành chính và ký hiệu.....56
Bảng 3.3: Dữ liệu đầu vào cho phân cụm phân cấp.....56
Bảng 3.4: Biểu diễn kết quả phân cụm chi tiết 59
Bảng 3.5: Biểu diễn kết quả phân cụm theo tiêu chí bệnh.....61

DANH MỤC CÁC HÌNH VẼ

Hình 1.1: Các giai đoạn trong quá trình khám phá tri thức 4
Hình 1.2: Quá trình khai phá dữ liệu.....7
Hình 2.1: Ví dụ về phân cụm theo mật độ[4].....18
Hình 2.2: Cấu trúc phân cụm dựa trên lưới[4].....19
Hình 2.3: Ví dụ về phân cụm dựa trên mô hình[4].....20
Hình 2.4: Cách mà các cụm có thể đưa ra.....22
Hình 2.5: Thuật toán k-means.....25
Hình 2.6: Sự thay đổi tâm cụm trong k-means khi có phần tử ngoại lai 28
Hình 2.7: Phân cụm phân cấp tập theo phương pháp "dưới lên" [4].....33
Hình 2.8: Single link 33
Hình 2.9: Complete link.....33
Hình 2.10: Các bước cơ bản của AGNES[4] 35
Hình 2.11: Các bước cơ bản của DIANA[4] 36
Hình 2.12: Cấu trúc cây CF.....37
Hình 2.13: Hình dạng các cụm được khám phá bởi thuật toán DBSCAN 40
Hình 2.14: Sắp xếp cụm trong OPTICS phụ thuộc vào ε[4].....41
Hình 3.1: Sơ đồ khối giải quyết bài toán 46
Hình 3.2: Phân tích dữ liệu gốc, thuộc tính "HO TEN" 49
Hình 3.3: Dữ liệu gốc sau khi loại bỏ thuộc tính thừa và dữ liệu trùng lặp.....50
Hình 3.4: Phân tích dữ liệu gốc, thuộc tính "QUAN HUYEN" 51
Hình 3.5: Loại bỏ một số giá trị của thuộc tính "QUAN HUYEN" 52
Hình 3.6: Dữ liệu trước và sau khi lọc thuộc tính "CHUAN DOAN DAU RA" 53
Hình 3.7: Cấu và phân bố dữ liệu mẫu.....58
Hình 3.8: Thiết lập tham số thuật toán K-means 58
Hình 3.9: Kết quả phân cụm chi tiết 59
Hình 3.10: Kết quả phân cụm dựa trên loại bệnh 60

Hình 3.11: Bản đồ phân bố bệnh các khu vực 62

MỞ ĐẦU

Đề tài tìm hiểu các phương pháp phân cụm dữ liệu, đánh giá ưu nhược điểm của mỗi phương pháp để tìm ra phương pháp phù hợp áp dụng trên tập dữ liệu mẫu. Kết quả sẽ được dùng để xây dựng bản đồ phân bố bệnh trên địa bàn tỉnh Thái Nguyên nhằm hỗ trợ công tác lên kế hoạch dự trù cơ sở vật chất, thuốc và các trang thiết bị khác cho các trung tâm y tế của Tỉnh.

Thái Nguyên là một tỉnh trung du miền núi thuộc vùng Đông Bắc của Việt Nam với diện tích hơn 3500 km^2 và dân số khoảng hơn một triệu người; bao gồm 9 đơn vị hành chính: Thành phố Thái Nguyên; Thị xã Sông Công và 7 huyện: Phổ Yên, Phú Bình, Đồng Hỷ, Võ Nhai, Định Hóa, Đại Từ, Phú Lương. Trong đó, tổng số gồm 180 xã, trong đó có 125 xã vùng cao và miền núi, còn lại là các xã đồng bằng và trung du.Tỉnh Thái Nguyên có nhiều dân tộc anh em sinh sống. Tuy nhiên, dân cư phân bố không đều, vùng cao và vùng núi dân cư rất thưa thớt, trong khi đó ở thành thị và đồng bằng dân cư lại dày đặc. Mật độ dân số thấp nhất là huyện Võ Nhai 72 người/ km^2, cao nhất là Thành phố Thái Nguyên với mật độ 1.260 người/ km^2.

Do sự khác biệt lớn trong cơ cấu dân số, lối sống, trình độ dân trí nên có những sự khác biệt đáng kể trong các hình thức bệnh trong các khu vực hành chính khác nhau. Nếu các thông tin về hình thức bệnh và các vấn đề sức khỏe trong mỗi khu vực hành chính được thu thập đầy đủ, nó sẽ có thể sẽ giúp việc phân bổ nguồn lực hiệu quả để phát triển các chính sách y tế công cộng cho các khu vực khác nhau.

Luận văn sử dụng các kỹ thuật khai thác dữ liệu để phân tích dữ liệu y tế thuộc Đại học Y Dược Thái Nguyên trongbốn tháng đầu năm 2015. Hy vọng rằng việc sử dụng các công cụ này một cách hiệu quả có thể phân tích và điều tra hình thức bệnh trong khu vực hành chính khác nhau của Thái Nguyên để tiếp tục xây dựng một bản đồ y tế cho tỉnh Thái Nguyên.

MỘT SỐ KẾT QUẢ NGHIÊN CỨU KHÁC

Ching-Kuo Wei et al. [2] Nghiên cứu này sử dụng các kỹ thuật khai phá dữ liệu điều tra các loại bệnh trong các khu vực hành chính khác nhau và phân tích sự khác nhau giữa các khu vực hành chính để tiếp tục xây dựng một bản đồ phân bố bệnh.

Nghiên cứu hy vọng sẽ giúp xây dựng trong tương lai các chiến lược y tế và phân bố các nguồn lực một cách thích hợp.

Lavrac [4] đề xuất một số kỹ thuật khai thác dữ liệu có thể được áp dụng trong y học, và đặc biệt là một số kỹ thuật máy học bao gồm các cơ chế mà làm cho chúng phù hợp hơn cho việc phân tích cơ sở dữ liệu y tế (nguồn gốc của các quy tắc mang tính biểu tượng, sử dụng các kiến thức nền, độ nhạy và độ đặc hiệu của giới thiệu gây ra). Tầm quan trọng của thông dịch các kết quả phân tích dữ liệu là thảo luận và minh họa trên các ứng dụng y tế đã chọn.

Lavrac et al. [5] đề xuất một phương pháp khai thác dữ liệu và công nghệ trực quan được sử dụng để hỗ trợ việc ra quyết định liên quan đến sức khỏe cộng đồng tại Slovenia.Mục đích nhằm khai thác cơ sở dữ liệu y tế công cộng để xác định khả năng đáp ứng của các dịch vụ y tế công cộng đối với các khu vực. Các kết quả có thể sử dụng để phát triển các chính sách chăm sóc sức khỏe cơ quan y tế.