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

NGUYỄN THỊ HUỆ

MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG

Chuyên ngành: Khoa học máy tính

Thái Nguyên - 2014

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành nhất đến ***PGS.TS Bùi Thế Hồng,*** người đã tận tình hướng dẫn, giúp đỡ em trong suốt thời gian thực hiện luận văn này.

Em cảm ơn các thầy trong Viện Công Nghệ Thông Tin Hà Nội cùng các thầy cô trong trường Đại học Công nghệ thông tin và truyền thông – ĐH Thái Nguyên đã giảng dạy em, giúp em có những kiến thức quý báu trong những năm học qua.

Mặc dù đã cố gắng hết sức cùng với sự tận tâm của thầy giáo hướng dẫn song do trình độ còn hạn chế nên luận văn của em khó tránh khỏi những thiếu sót. Em rất mong nhận được sự thông cảm và góp ý của thầy cô và các bạn.

Thái Nguyên, tháng 06 năm 2014

Học viên

Nguyễn Thị Huệ

LỜI CAM ĐOAN

Em xin cam đoan toàn bộ nội dung bản luận văn này là do em tự sưu tầm, tra cứu và sắp xếp cho phù hợp với nội dung yêu cầu của đề tài.

Tất cả các thử nghiệm của luận văn đều do em tự thiết kế và xây dựng, thuật toán phân cụm Hierarchical clustering được viết bằng MATLAB và kết quả thử nghiệm của thuật toán này được so sánh với kết quả thử nghiệm trên cùng bộ dữ liệu được phân tích bằng thuật toán chuẩn của phần mềm phân tích thống kê dữ liệu SPSS 20.0. Bảng dữ liệu về *Tỉ suất chết của trẻ em dưới 1 tuổi, tỉ suất sinh thô và tổng tỉ suất sinh năm 2007* của các nước trên thế giới là do em sưu tầm từ những nguồn tin cậy của một số tổ chức của liên hợp quốc (Worldbank, UNFPA, UNDP) và từ đĩa DVD Microsoft Student with Encara Prenium 2009. Nếu sai em xin hoàn toàn chịu trách nhiệm.

Thái Nguyên, tháng 06 năm 2014

Nguyễn Thị Huệ

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Tên viết tắt	Tên tiếng Anh	Định nghĩa
1	IMR	Infant Mortality Rate	Tỉ suất chết của trẻ em dưới 1 tuổi (‰)
2	BR	Crude Birth Rate	Tỉ suất sinh thô (‰)
3	TFR	Total Fertility Rate	Số con trung bình sinh ra sống của một người phụ nữ trong suốt thời gian sinh sản

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay, chúng ta thường phải xử lý những tập dữ liệu lớn bao gồm rất nhiều các quan sát, các đối tượng. Để hiểu rõ về cấu trúc của các tập dữ liệu này, người ta thường tiến hành hai kiểu phân tích. Kiểu thứ nhất là *phân lớp* các đối tượng dữ liệu theo một thuộc tính phân lớp nào đó. Kỹ thuật này bao gồm hai bước. Bước thứ nhất là xây dựng mô hình dựa vào một tập dữ liệu mẫu được phân chia theo một thuộc tính lớp. Bước thứ hai là phân lớp các đối tượng dữ liệu theo mô hình đã xây dựng ở bước một. Kiểu này được gọi là học có giám sát tức là phải có mẫu trước. Kiểu thứ hai là *phân cụm.* Phân cụm là kỹ thuật phân chia một tập lớn các đối tượng thành các cụm khác nhau theo một số thuộc tính nào đó sao cho các đối tượng trong cùng một cụm là tương đồng với nhau theo các thuộc tính này và các cụm khác nhau là hoàn toàn khác biệt với nhau cùng trên các thuộc tính đã cho. Nói cách khác, mục tiêu của phân cụm là phân chia các quan sát thành các nhóm đồng nhất và khác biệt.

Không giống như phân loại dữ liệu, phân cụm không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì vậy, thông thường cần có một chuyên gia về lĩnh vực đó để đánh giá các cụm thu được. Phân cụm dữ liệu được sử dụng nhiều trong các ứng dụng về phân cụm các quốc gia, các vùng lãnh thổ theo một số tiêu chí về nhân khẩu học, về phát triển kinh tế và xã hội, hoặc phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, …Cho đến hiện nay, phân tích cụm đã được sử dụng nhiều trong phân tích thống kê và đang được áp dụng rộng rãi trong khai phá dữ liệu. Những nghiên cứu tiếp theo về kỹ thuật này là rất cần thiết và hứa hẹn nhiều triển vọng.

Do đặc thù của kỹ thuật phân cụm và do khả năng ứng dụng rất phong phú của kỹ thuật này nên em đã chọn nghiên cứu đề tài ***Một số kỹ thuật phân cụm dữ liệu và ứng dụng*** làm luận văn tốt nghiệp cao học.

2. Mục tiêu của đề tài

Nghiên cứu các kỹ thuật phân cụm dữ liệu trong phân tích thống kê dữ liệu cũng như trong khai phá dữ liệu và thử nghiệm phân tích cụm trong nhân khẩu học.

3. Đối tượng và phạm vi nghiên cứu

- Nghiên cứu một số kỹ thuật phân cụm trong phân tích thống kê dữ liệu và trong khai phá dữ liệu.
- Phân tích thống kê dữ liệu.
- Khai phá dữ liệu.
- Điều tra nhân khẩu học

4. Phương pháp nghiên cứu

- Tìm hiểu, thu thập các tài liệu có liên quan.

- Nghiên cứu các phương pháp phân cụm trong phân tích thống kê dữ liệu, trong khai phá dữ liệu và cài đặt thuật toán phân cụm Hieararchical Clustering.

5. Ý nghĩa khoa học của đề tài.

- Phân tích cụm là một kỹ thuật có phạm vi ứng dụng rất rộng, đặc biệt là trong lĩnh vực phân tích điều tra xã hội học và khai phá dữ liệu. Phân tích và đánh giá các kỹ thuật phân cụm khác nhau là một vấn đề cần thiết trong việc chọn lựa một kỹ thuật thích hợp với mỗi kiểu ứng dụng.

- Đề tài của luận văn nhằm mục đích nghiên cứu đánh giá so sánh kỹ thuật phân cụm đã được cài đặt trong bộ chương trình phân tích thống kê SPSS và kỹ thuật phân cụm áp dụng trong khai phá dữ liệu. Qua đó có thể sẽ rút ra được những kết luận về hiệu quả của hai kiểu phân tích cụm này.