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

TRẦN THỊ YẾN

PHÂN CỤM DỮ LIỆU TRỪ MỜ VÀ ỨNG DỤNG

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

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: **PGS. TS LÊ BÁ DŨNG**

Thái Nguyên - 2012

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời biết ơn sâu sắc đến PGS.TS Lê Bá Dũng, người đã tận tình hướng dẫn, chỉ bảo, giúp đỡ em trong suốt quá trình làm luận văn.

Em cũng xin được bày tỏ lòng biết ơn tới các thầy đã tham gia giảng dạy và chia sẻ những kinh nghiệm quý báu cho tập thể lớp nói chung và cá nhân em nói riêng.

Tôi xin gửi lời cảm ơn tới gia đình, bạn bè, đồng nghiệp đã luôn ủng hộ, động viên và giúp đỡ để tôi có thể hoàn thành tốt luận văn.

Tôi cũng xin gửi lời cảm ơn tới Ban giám hiệu trường Đại học Khoa học, Ban chủ nhiệm Khoa Toán-Tin đã tạo điều kiện thuận lợi cho tôi tham gia khóa học và hoàn thành luận văn.

Một lần nữa, xin chân thành cảm ơn.

Thái Nguyên, tháng 09 năm 2012

Học viên

Trần Thị Yến

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là kết quả của sự tìm hiểu, nghiên cứu các tài liệu một cách nghiêm túc dưới sự hướng dẫn của PGS. TS Lê Bá Dũng. Nội dung luận văn được phát triển từ ý tưởng, sự sáng tạo của bản thân và kết quả có được hoàn toàn trung thực.

Học viên

Trần Thị Yến

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

CURE	Clustering Using Representatives
DBSCAN	Density based Spatial Clutering of Application with Noise
DENCLUE	Clustering Based on Density Distribution Functions
EM	Expectation Maximization
FCM	Fuzzy C-Means
FSC	Fuzzy Subtractive Clustering
OPTICS	Ordering Points to Identify the Clustering Structure
SC	Subtractive Clustering

DANH MỤC CÁC BẢNG BIỂU, HÌNH ẢNH

MỞ ĐẦU

Ngày nay, khai phá dữ liệu (Datamining) đã trở thành một trong những xu hướng nghiên cứu phổ biến trong lĩnh vực học máy và công nghệ tri thức. Nhiều thành tựu nghiên cứu của Datamining đã được áp dụng trong thực tế. Datamining có nhiều hướng quan trọng và một trong các hướng đó là phân cụm dữ liệu (Data Clustering). Phân cụm dữ liệu là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" (similar) với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" (dissimilar) với nhau. Phân cụm dữ liệu là một phương pháp học không giám sát.

Hiện nay, các phương pháp phân cụm đã và đang được phát triển và áp dụng nhiều trong các lĩnh vực khác nhau, bao gồm: nhận dạng, phân tích dữ liệu, nghiên cứu thị trường, xử lý ảnh,… Các thuật toán phân cụm cũng rất đa dạng như K-means, Pam, C-means, C-means mờ, thuật toán phân cụm trừ,… Để tăng tính ổn định và chính xác của kết quả phân cụm, ngày càng có các tiếp cận mới. Một trong những cách tiếp cận đang được nghiên cứu đó là ứng dụng lý thuyết mờ vào bài toán phân cụm dữ liệu.

Luận văn này trình bày phân cụm dữ liệu, một cách tiếp cận mới về phân cụm dữ liệu là thuật toán phân cụm trừ mờ và ứng dụng vào bài toán cụ thể.

Luận văn bao gồm các nội dung chính sau:

Chương 1: Tổng quan về phân cụm dữ liệu

Chương 2: Phương pháp phân cụm trừ mờ

Chương 3: Ứng dụng phương pháp phân cụm trừ mờ

Chương 1.

TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU

1.1. Khái niệm và mục tiêu của phân cụm dữ liệu

Phân cụm dữ liệu là một kỹ thuật trong khai phá dữ liệu, là quá trình phân chia một tập dữ liệu ban đầu thành các cụm sao cho các phần tử trong một cụm "tương tự" với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định theo phương pháp phân cụm.

Trong học máy, phân cụm dữ liệu được xem là vấn đề học không có giám sát, vì nó phải giải quyết vấn đề tìm một cấu trúc trong tập hợp dữ liệu chưa biết trước các thông tin về cụm hay các thông tin về tập huấn luyện. Trong nhiều trường hợp, nếu phân lớp được xem là vấn đề học có giám sát thì phân cụm dữ liệu là một bước trong phân lớp dữ liệu, phân cụm dữ liệu sẽ khởi tạo các lớp cho phân lớp bằng cách xác định các nhãn cho các nhóm dữ liệu.

Phân cụm có ý nghĩa rất quan trọng trong hoạt động của con người. Ngay từ lúc bé, con người đã học cách làm thế nào để phân biệt giữa mèo và chó, giữa động vật và thực vật và liên tục đưa vào sơ đồ phân loại trong tiềm thức của mình. Phân cụm được sử dụng rộng rãi trong nhiều ứng dụng, bao gồm nhận dạng mẫu, phân tích dữ liệu, xử lý ảnh, nghiên cứu thị trường... Với tư cách là một chức năng khai phá dữ liệu, phân cụm có thể được sử dụng như một công cụ độc lập chuẩn để quan sát đặc trưng của mỗi cụm thu được bên trong sự phân bố của dữ liệu và tập trung vào một tập riêng biệt của các cụm để giúp cho việc phân tích đạt kết quả.

Một vấn đề thường gặp trong phân cụm là hầu hết các dữ liệu cần cho phân cụm đều có chứa dữ liệu nhiễu do quá trình thu thập thiếu chính xác hoặc thiếu đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lý dữ liệu nhằm khắc phục hoặc loại bỏ nhiễu trước khi chuyển sang giai đoạn phân tích cụm dữ liệu. Nhiễu ở đây được hiểu là các đối tượng dữ liệu không chính xác, không tường minh

hoặc là các đối tượng dữ liệu khuyết thiếu thông tin về một số thuộc tính... Một trong các kỹ thuật xử lý nhiễu phổ biến là việc thay thế giá trị các thuộc tính của đối tượng nhiễu bằng giá trị thuộc tính tương ứng. Ngoài ra, dò tìm phần tử ngoại lai cũng là một trong những hướng nghiên cứu quan trọng trong phân cụm, chức năng của nó là xác định một nhóm nhỏ các đối tượng dữ liệu khác thường so với các dữ liệu trong cơ sở dữ liệu, tức là các đối tượng dữ liệu không tuân theo các hành vi hoặc mô hình dữ liệu nhằm tránh sự ảnh hưởng của chúng tới quá trình và kết quả của phân cụm.

Tóm lại, phân cụm dữ liệu cần phải giải quyết các vấn đề cơ bản như sau:

- Biểu diễn dữ liệu,
- Xây dựng hàm tính độ tượng tự,
- Xây dựng các tiêu chuẩn phân cụm,
- Xây dựng mô hình cho cấu trúc cụm dữ liệu,
- Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo,
- Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm.

Theo các nghiên cứu cho thấy thì hiện nay chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc dữ liệu. Hơn nữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của các dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp. Vì vậy phân cụm dữ liệu vẫn đang là một vấn đề khó và mở, vì phải giải quyết nhiều vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt là đối với dữ liệu hỗn hợp đang ngày càng tăng trong các hệ quản trị dữ liệu và đây cũng là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu.

Mục tiêu của phân cụm là xác định được bản chất của các cụm dữ liệu trong tập dữ liệu chưa có nhãn, theo đó cho phép đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho việc ra quyết định. Tuy nhiên, không có tiêu chí nào được xem là tốt