

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

□ □ □

VŨ NGỌC THANH

PHÂN CỤM DỮ LIỆU DỰA TRÊN MẬT ĐỘ
VÀ ỨNG DỤNG

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYỄN – 2016
ĐẠI HỌC THÁI NGUYỄN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN
THÔNG

□ □ □

VŨ NGỌC THANH

**PHÂN CỤM DỮ LIỆU DỰA TRÊN MẬT ĐỘ
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 NGUYỄN – 2016

LỜI CẢM ƠN

Để hoàn thành được luận văn này, trước hết em xin gửi lời cảm ơn sâu sắc nhất tới TS. Nguyễn Huy Đức, đã tận tình hướng dẫn, chỉ bảo, định hướng, đóng góp những ý kiến quý báu trong suốt quá trình em thực hiện luận văn.

Em xin chân thành cảm ơn các thầy, cô giáo trong trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên đã tạo mọi điều kiện tốt nhất để em hoàn thành khóa học này. Đồng thời, em cũng xin cảm ơn gia đình, bạn bè, những người luôn khuyến khích và giúp đỡ tôi trong mọi hoàn cảnh khó khăn. Tôi xin cảm ơn cơ quan và các đồng nghiệp đã hết sức tạo điều kiện cho tôi trong suốt quá trình học tập và làm luận văn này.

Thái Nguyên, ngày 17 tháng 09 năm 2016.

Học viên

Vũ Ngọc Thanh

MỤC LỤC

MỤC LỤC.....	1
DANH MỤC HÌNH ẢNH	7
DANH MỤC TỪ VIẾT TẮT.....	8
MỞ ĐẦU	9
CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ PHÂN CỤM DỮ LIỆU	11
1.1. Tổng quan về khai phá dữ liệu.....	11
1.1.1. Khái niệm	11
1.1.2. Tiến trình khai phá dữ liệu.....	12
1.1.3. Các mô hình khai phá dữ liệu	14
1.1.4. Các hướng tiếp cận và kỹ thuật sử dụng trong khai phá dữ liệu...	15
1.1.5. Các dạng dữ liệu có thể khai phá	16
1.1.6. Các ứng dụng của khai phá dữ liệu.....	17
1.2. Tổng quan về phân cụm dữ liệu.....	19
1.2.1. Khái niệm	19
1.2.2. Các mục tiêu của phân cụm dữ liệu	20
1.2.3. Các ứng dụng của phân cụm dữ liệu.....	22
1.2.4. Các yêu cầu của phân cụm dữ liệu.....	23
1.2.5. Những vấn đề còn tồn tại trong phân cụm dữ liệu.....	26
1.2.6. Một số khái niệm cần thiết khi tiếp cận phân cụm dữ liệu	26
1.2.7. Những kỹ thuật tiếp cận trong phân cụm dữ liệu.....	31
CHƯƠNG 2: PHÂN CỤM DỮ LIỆU DỰA TRÊN MẬT ĐỘ	37
2.1. Giới thiệu.....	37
2.2. Thuật toán DBSCAN	38
2.3. Thuật toán DBRS	49
2.4. Thuật toán OPTICS.....	55
2.5. Thuật toán DENCLUDE	56
CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH THỰC NGHIỆM.....	60

3.1. Ý tưởng bài toán.....	60
3.2. Nguồn dữ liệu đầu vào	60
3.3. Phương pháp giải quyết bài toán.....	60
3.4. Kết quả thực nghiệm	61
KẾT LUẬN	Error! Bookmark not defined.
TÀI LIỆU THAM KHẢO.....	66
PHỤ LỤC	67

DANH MỤC HÌNH ẢNH

Hình 1.1: Tiến trình khám phá tri thức từ cơ sở dữ liệu

Hình 1.2: Kiến trúc điển hình của một hệ khai phá dữ liệu

Hình 1.3: Ví dụ về phân cụm dữ liệu

Hình 1.4: Ví dụ phân cụm các ngôi nhà dựa trên khoảng cách

Hình 1.5: Ví dụ phân cụm các ngôi nhà dựa trên kích cỡ

Hình 2.1: Ví dụ về đối tượng nòng cốt, đối tượng biên và đối tượng nhiễu

Hình 2.2: Ví dụ về mật độ đạt được trực tiếp

Hình 2.3: Ví dụ về mật độ đạt được

Hình 2.4: Ví dụ về mật độ liên thông

Hình 2.5: Minh họa đồ thị khoảng cách 4-dist đã được sắp xếp của một CSDL

Hình 2.6: Kết quả thực nghiệm đánh giá thời gian thực hiện thuật toán (tính theo giây) trên 2 thuật toán của nhóm tác giả

Hình 2.7: Các cụm phát hiện được bởi CLARANS (a) và DBSCAN (b)

Hình 2.8: Các cụm được phát hiện bởi DBRS(a), DBSCAN(b), K-Means(c), CLARANS(d)

Hình 2.9: Sắp xếp cụm trong OPTICS phụ thuộc vào ϵ

Hình 2.10: DENCLUE với hàm phân phối Gaussian

Hình 3.1: Kết quả sau khi phân cụm của chương trình thực nghiệm

DANH MỤC TỪ VIẾT TẮT

Từ hoặc cụm từ	Từ viết tắt	Từ tiếng Anh
Cơ sở dữ liệu	CSDL	Database
Khai phá tri thức trong cơ sở dữ liệu	KDD	Knowledge Discovery in Databases
Khai phá tri thức	KPTT	Knowledge Discovery
Khai phá dữ liệu	KPDL	Data Mining
Phân cụm dữ liệu	PCDL	Data Clustering

MỞ ĐẦU

Trong vài thập niên gần đây, cùng với sự thay đổi và phát triển không ngừng của ngành công nghệ thông tin nói chung và trong các ngành công nghệ phần cứng, phần mềm, truyền thông và hệ thống các dữ liệu phục vụ trong các lãnh vực kinh tế - xã hội nói riêng. Thì việc thu thập thông tin cũng như nhu cầu lưu trữ thông tin càng ngày càng lớn. Bên cạnh đó việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Hàng triệu cơ sở dữ liệu đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lý ... trong đó có nhiều cơ sở dữ liệu cực lớn cỡ Gigabyte, thậm chí là Terabyte. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền công nghệ thông tin thế giới hiện nay. Một vấn đề được đặt ra là phải làm sao trích chọn được những thông tin có ý nghĩa từ tập dữ liệu lớn để từ đó có thể giải quyết được các yêu cầu của thực tế như trợ giúp ra quyết định, dự đoán,... và khai phá dữ liệu (Data mining) đã ra đời nhằm giải quyết các yêu cầu đó.

Ngay từ những ngày đầu khi xuất hiện, Data mining đã trở thành một trong những xu hướng nghiên cứu phổ biến trong lĩnh vực học máy tính và công nghệ tri thức. Nhiều thành tựu nghiên cứu của Data mining đã được áp dụng trong thực tế. Data mining 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 tìm kiếm để phân ra các cụm dữ liệu, các mẫu dữ liệu từ tập Cơ sở dữ liệu lớn. Phân cụm dữ liệu là một phương pháp học không giám sát.

Phân cụm dữ liệu là một trong những kỹ thuật để khai thác dữ liệu có hiệu quả. Phân cụm dữ liệu đã được ứng dụng trong nhiều lĩnh vực khác nhau: kinh tế, bảo hiểm, quy hoạch đô thị, nghiên cứu về địa chấn v.v... Có rất nhiều kỹ thuật tiếp cận trong phân cụm dữ liệu, tùy thuộc vào bài toán thực tế mà chúng ta có thể chọn những phương pháp cho phù hợp. Trong luận văn này em xin được trình bày những nghiên cứu của bản thân về phương pháp **“Phân cụm dữ liệu dựa trên mật độ và ứng dụng”**.

Bố cục luận văn như sau:

Ngoài các phần mở đầu, mục lục, danh mục hình ảnh, kết luận, tài liệu tham khảo, phụ lục. Luận văn được chia là 3 phần chính:

- Phần 1: *Tổng quan về khai phá dữ liệu và phân cụm dữ liệu*
Phần này giới thiệu các khái niệm cơ bản về khai phá dữ liệu và phân cụm dữ liệu. Các phương pháp, lĩnh vực và các hướng tiếp cận trong phân cụm dữ liệu.
- Phần 2: *Phương pháp phân cụm dữ liệu dựa trên mật độ*
Phần này trình bày chi tiết phương pháp phân cụm dữ liệu dựa trên mật độ và các thuật toán tiêu biểu của phương pháp này.
- Phần 3: *Xây dựng chương trình thực nghiệm*
Xây dựng chương trình thực nghiệm phân cụm dữ liệu dựa trên mật độ với giải thuật DBSCAN.