

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

Phan Thị Thu Nga

**NGHIÊN CỨU PHƯƠNG PHÁP PHÂN CỤM NỬA GIẢM
SÁT VÀ ỨNG DỤNG**

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên, tháng 9 năm 2015

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài với tiêu đề *Nghiên cứu phương pháp phân cụm nửa giám sát và ứng dụng* là công trình nghiên cứu được tôi thực hiện dưới sự hướng dẫn của giáo viên hướng dẫn khoa học.

Các kết quả nghiên cứu và kết quả thử nghiệm nêu trong luận văn là trung thực và chưa từng được công bố trong bất kỳ tài liệu nào khác. Trong phần kiến thức chung, nghiên cứu giải thuật áp dụng và một số thực nghiệm kết quả tương ứng tôi có tham khảo ở một số tài liệu và đã có trích dẫn đúng và đầy đủ.

Học viên

Phan Thị Thu Nga

MỤC LỤC

LỜI CAM ĐOAN	ii
LỜI MỞ ĐẦU	1
CHƯƠNG 1. TỔNG QUAN VỀ HỌC MÁY VÀ BÀI TOÁN PHÂN CỤM, PHÂN CỤM NỬA GIÁM SÁT.....	4
1.1 Khái niệm về trí tuệ nhân tạo và học máy	4
1.2 Các nghiên cứu sử dụng trong học máy.....	6
1.3 Ứng dụng của phương pháp học máy	6
1.4 Khái niệm về bài toán phân cụm.....	7
1.5 Các yêu cầu của bài toán phân cụm	8
1.6 Các chiến lược trong phương pháp phân cụm dữ liệu [2].	11
1.8 Đánh giá kết quả của thuật toán phân cụm	17
1.9 Phương pháp phân cụm nửa giám sát	17
CHƯƠNG 2. MỘT SỐ THUẬT TOÁN PHÂN CỤM NỬA GIÁM SÁT	19
2.1. Giới thiệu bài toán phân cụm nửa giám sát	19
2.2 Thuật toán phân cụm nửa giám sát với K-Means	21
2.2.1 Thuật toán K-Means.....	21
2.2.2. Thuật toán Seed K-Means	26
2.3 Thuật toán phân cụm nửa giám sát SSDBSCAN.....	27
2.3.1 Thuật toán DBSCAN	27
2.3.2 Thuật toán SSDBSCAN	29
2.3.3 Thuật toán ActSSDBSCAN.....	32
2.4. Thuật toán phân cụm nửa giám sát Fuzzy C-Means.....	37
2.4.1 Thuật toán Fuzzy C-Means	37
2.4.2 Thuật toán Seed Fuzzy C-means	40
CHƯƠNG 3. ỨNG DỤNG THUẬT TOÁN PHÂN CỤM TRONG LĨNH VỰC XỬ LÝ ẢNH.....	43
3.1 Giới thiệu tổng quan.....	43

3.2 Phân vùng ảnh (Image segmentation) sử dụng Fuzzy C-Means	43
3.2.1 Tóm lược về vấn đề xử lý ảnh số (<i>Digital Image Processing</i>)	43
3.2.2 Lập trình và thử nghiệm	45
3.3 Phân cụm ảnh với thuật toán SSDBSCAN	53
3.3.1 Dữ liệu	53
3.3.2 Kết quả	55
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	58
❖ Những kết quả đã đạt được	58
❖ Hướng phát triển của đề tài	58
PHỤ LỤC	60
TÀI LIỆU THAM KHẢO	63

DANH MỤC CÁC HÌNH

Số hiệu hình	Tên hình	Trang
Hình 1.1	Các mô hình của học máy	5
Hình 1.2	Ví dụ về dữ liệu của bài toán phân cụm: hình bên trái bao gồm 3 cụm rời nhau tuyến tính trong khi hình bên phải có các cụm hình dạng bất kỳ và không tuyến tính	8
Hình 1.3	Một số ví dụ về phân đoạn ảnh sử dụng clustering	14
Hình 1.4	Ứng dụng clustering trong việc phát hiện những vùng bị hỏng trên trái cây	15
Hình 1.5	Bài toán phân cụm	17
Hình 1.6	Bài toán phân cụm nửa giám sát	17
Hình 2.1	Minh họa thuật toán K-Means: Sự phụ thuộc vào trọng tâm tại bước khởi tạo dẫn đến các kết quả khác nhau sau mỗi lần chạy	22
Hình 2.2	Dữ liệu (với 9 cụm) mà K-Means không thể phát hiện chính xác các cụm	22
Hình 2.3	Minh họa các bước của thuật toán hierarchical clustering	24
Hình 2.4	Quá trình xây dựng cluster của DBSCAN: Các hình cầu bán kính ϵ và chứa ít nhất MinPts điểm sẽ được kết nối để tạo ra các cluster	26
Hình 2.5	Minh họa thuật toán Fuzzy C-Means với 4 cụm	30

Hình 2.6	(a) Các điểm tương ứng với tập dữ liệu đầu vào, các seed (các dữ liệu đã được gán nhãn) tương ứng là các điểm ký hiệu bởi các dấu cộng, dấu nhân, và dấu sao. (b) các ràng buộc (constraint) must-link và cannot-link được biểu diễn tương ứng bằng các đoạn thẳng nét liền và nét đứt.	31
Hình 2.7	Quá trình xây dựng cluster của SSDBSCAN	33
Bảng 2.1	Các tập dữ liệu sử dụng (n : số phần tử cần clustering, m : số thuộc tính, và k là số cluster)	38
Hình 2.8	Kết quả thực nghiệm của thuật toán ActSSDBSCAN và SSDBSCAN	40
Hình 2.9	Kết quả thực nghiệm của thuật toán SECM	43
Hình 3.1	Tổng quan về hệ thống xử lý ảnh trên máy tính [5]	46
Hình 3.2	Ảnh gốc (a) và các vùng (màu sáng) được phân tách bởi thuật toán Fuzzy C-means	48
Hình 3.3	Các cụm phân tách từ thuật toán Fuzzy C-means	50
Hình 3.4	Phân tách vùng với 2 cụm khác nhau	50
Hình 3.5	Ví dụ về phân vùng ảnh cây cối	52
Hình 3.6	Ví dụ về phân vùng ảnh động vật	55
Hình 3.7	Dữ liệu lấy từ trang UCI	56
Hình 3.8	Kết quả với tập dữ liệu image210	57

Hình 3.9	Kết quả với tập dữ liệu image300	58
Hình 3.10	Kết quả với tập dữ liệu image420	58
Hình 3.11	Kết quả với tập dữ liệu image560	59

LỜI MỞ ĐẦU

1. Lí do chọn đề tài

Công nghệ thông tin là một lĩnh vực quan trọng trong sự phát triển của xã hội trên toàn thế giới. Những ứng dụng của công nghệ thông tin (CNTT) có thể gặp bất kỳ nơi đâu, trong lĩnh vực gì. Từ những năm 50 của thế kỷ 20, Liên Xô đã phóng thành công tàu vũ trụ, để thực hiện được điều này thì rõ ràng cần đến các chương trình máy tính và phần mềm. Những năm 60 nước Mỹ đã có hệ thống mạng máy tính phục vụ cho quốc phòng và an ninh cũng như các lĩnh vực trọng yếu. Mạng Internet ra đời đã kết nối toàn thế giới và mạng lại hiệu quả vô cùng lớn cho nhân loại. Ngày nay sự phát triển về CNTT đã là toàn cầu hóa, tất cả các quốc gia đều coi việc phát triển lĩnh vực này là mũi nhọn là công cụ để thúc đẩy sự phát triển. Tất cả các hệ thống tinh vi hiện đại nhất như tàu vũ trụ, máy bay, tàu ngầm, ô tô, người máy đều có hạt nhân quan trọng là CNTT.

Trí tuệ nhân tạo là một hướng nghiên cứu của lĩnh vực CNTT và Khoa học máy tính nhằm phát triển các hệ thống thông minh giải quyết các bài toán trong thực tế giống như hoạt động của bộ não người. Trí tuệ nhân tạo được bắt đầu nghiên cứu từ những năm 50 của thế kỷ 20 và trong khoảng 30 năm trở lại đây đã được cộng đồng các nhà khoa học quan tâm mạnh mẽ. Rất nhiều các hội thảo lớn được tổ chức hàng năm tại Mỹ cũng như trên thế giới. Các ứng dụng tiêu biểu của trí tuệ nhân tạo vào đời sống xã hội bao gồm: người máy, robot, xử lý ngôn ngữ tự nhiên, nhận dạng, phát hiện dị thường, an ninh quốc phòng, tin sinh học, khoa học vũ trụ và trái đất...

Trong khuôn khổ luận văn Thạc sỹ của mình, qua việc được trang bị các môn học lý thuyết như thuật toán, xử lý ảnh, trí tuệ nhân tạo,... tôi đã lựa

chọn đề tài *Nghiên cứu phương pháp phân cụm nửa giám sát và ứng dụng*. Chủ đề phân cụm dữ liệu là một nhánh nhỏ nằm trong lĩnh vực học máy (machine learning) của trí tuệ nhân tạo nhằm nghiên cứu và ứng dụng bài toán phân cụm trong thực tế. Hơn nữa phân cụm có giám sát là một hướng đi tốt trong khoảng 10 năm trở lại đây và đã chứng minh tính ưu việt của nó.

2. Đối tượng và phạm vi nghiên cứu

2.1. Đối tượng nghiên cứu

Vấn đề phân cụm dữ liệu (clustering), phân cụm nửa giám sát (semi-supervised clustering).

2.2. Phạm vi nghiên cứu

- **Lý thuyết:** Nghiên cứu các thuật toán cơ bản về phân cụm dữ liệu và phân cụm nửa giám sát.
- **Thực nghiệm:** Lập trình trên ngôn ngữ C# cho ứng dụng phân vùng ảnh và phân cụm ảnh.

3. Phương pháp nghiên cứu.

- Phương pháp nghiên cứu khoa học và suy luận logic.
- Phương pháp nghiên cứu mô tả, giải thích, giải pháp.

4. Ý nghĩa khoa học và thực tiễn của đề tài.

- Về khoa học

Nghiên cứu các thuật toán phân cụm và phân cụm nửa giám sát, đánh giá ưu, nhược điểm và giải thích kết quả đạt được của từng phương pháp

- Về thực tiễn

Ứng dụng thuật toán phân cụm vào bài toán phân vùng ảnh và phân cụm ảnh từ đó rút ra các nhận xét, đánh giá vào thực tiễn.

5. Cấu trúc của luận văn

Chương 1: Giới thiệu tổng quan về học máy và bài toán phân cụm dữ liệu và phân cụm dữ liệu nửa giám sát.

Chương 2: Trình bày một số thuật toán phân cụm dữ liệu cơ bản, thuật toán phân cụm nửa giám sát cũng như đánh giá ưu nhược điểm của mỗi thuật toán.

Chương 3: Cài đặt thử nghiệm hai ứng dụng của bài toán phân cụm bao gồm bài toán phân vùng ảnh và bài toán phân cụm ảnh.

Tóm lược các kết quả đã thực hiện được và đề ra hướng phát triển tiếp theo của đề tài.