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

LÊ THỊ MAI HƯƠNG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN PHÂN CỤM DỮ LIỆU NỬA GIÁM SÁT VÀ ỨNG DỤNG PHÂN ĐOẠN ẢNH X-QUANG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2017

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

LÊ THỊ MAI HƯƠNG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN PHÂN CỤM DỮ LIỆU NỬA GIÁM SÁT VÀ ỨNG DỤNG PHÂN ĐOẠN ẢNH X-QUANG

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

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Giáo viên hướng dẫn: TS.Nguyễn Đình Dũng

THÁI NGUYÊN - 2017

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do chính tôi thực hiện, dưới sự hướng dẫn khoa học của TS. Nguyễn Đình Dũng, các kết quả lý thuyết được trình bày trong luận văn là sự tổng hợp từ các kết quả đã được công bố và có trích dẫn đầy đủ, kết quả của chương trình thực nghiệm trong luận văn này được tác giả thực hiện là hoàn toàn trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm.

Thái Nguyên, tháng 6 năm 2016

Học viên

Lê Thị Mai Hương

LỜI CẢM ƠN

Luận văn này được hoàn thành tại Trường Đại học Công nghệ Thông tin và Truyền thông dưới sự hướng dẫn của TS. Nguyễn Đình Dũng. Tác giả xin bày tỏ lòng biết ơn tới các thầy cô giáo thuộc Trường Đại học Công nghệ Thông tin và Truyền thông, các thầy cô giáo thuộc Viện Công nghệ Thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tạo điều kiện, giúp đỡ tác giả trong quá trình học tập và làm luận văn tại Trường, đặc biệt tác giả xin bày tỏ lòng biết ơn tới TS. Nguyễn Đình Dũng đã tận tình hướng dẫn và cung cấp nhiều tài liệu cần thiết để tác giả có thể hoàn thành luận văn đúng thời hạn.

Xin chân thành cảm ơn anh chị em học viên cao học và bạn bè đồng nghiệp đã trao đổi, khích lệ tác giả trong quá trình học tập và làm luận văn tại Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

Cuối cùng tác giả xin gửi lời cảm ơn đến gia đình, những người đã luôn bên cạnh, động viên và khuyến khích tôi trong quá trình thực hiện đề tài.

Thái Nguyên, ngày tháng năm 2017

Học viên

Lê Thị Mai Hương

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Từ viết đầy đủ
DB	Davies-Bouldin
eSFCM	Semi-supervised Entropy regularized Fuzzy Clustering
FCM	Fuzzy C-Mean
PBM	Pakhira, Bandyopadhyay and Maulik
SSFCM	Semi-Supervised Fuzzy C-Mean
SSSFC	Semi-Supervised Standard Fuzzy Clustering
SSWC	Simplified Silhouete Width Criterion
CSDL	Cơ sở dữ liệu
PCDL	Phân cụm dữ liệu

DANH MỤC HÌNH VẼ

LỜI MỞ ĐẦU

Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp đồng thời cũng tìm ra các mẫu tiềm ẩn trong dữ liệu đó. Hiện nay việc khai phá dữ liệu được nghiên cứu theo các hướng mô tả khái niệm, luật kết hợp, phân lớp và dự đoán, phân cụm (xem [1], [2], [7]) và có nhiều ứng dụng trong thực tế, trong đó phân đoạn ảnh X-Quang trong lĩnh vực y tế là một ứng dụng điển hình [13]. Ngày nay, việc xử lý các hình ảnh y tế có vai trò quan trọng trong việc tự động hóa phân tích, hỗ trợ chẩn đoán và điều trị các bệnh khác nhau. Trong đó, quá trình phân đoạn thường được yêu cầu như là giai đoạn sơ bộ. Tuy nhiên các phân vùng trong hình ảnh y tế rất phức tạp nên việc phân đoạn chính xác là rất quan trọng.

Trong các phương pháp phân đoạn ảnh hiện có, phân cụm là một phương pháp được sử dụng rộng rãi bởi tính đơn giản và hiệu quả mà nó mang lại (xem [8]-[12]). Phân cụm dữ liệu là lĩnh vực học máy không giám sát, nó có chức năng tổ chức một tập đối tượng dữ liệu thành các cụm sao cho những đối tượng trong cùng một cụm thì tương tự như nhau còn các đối tượng ở các cụm khác nhau thì kém tương tự nhau hơn. Nhược điểm chung của thuật toán phân cụm là chất lượng phân cụm phụ thuộc nhiều vào các tham số và thông tin khởi tạo. Để giảm thiểu các hạn chế này, gần đây đã có nhiều tác giả (xem [8]-[12]) giải quyết theo cách tiếp cận nửa giám sát, trong đó việc phân cụm được thực hiện dựa vào các thông tin bổ trợ đóng vai trò điều khiển quá trình phân cụm, nhờ đó mà chất lượng phân cụm được nâng lên đáng kể.

Mục tiêu của luận văn là nghiên cứu, tìm hiểu một số thuật toán phân cụm nửa giám sát và xây dựng được một ứng dụng thử nghiệm cho thuật toán phân đoạn ảnh X-quang hỗ trợ chuẩn đoán bệnh trong lĩnh vực nha khoa. Các kết quả đạt được trong luận văn này là kết quả trong quá trình học tập và nghiên cứu của tác giả tại Trường Đại học Công nghệ Thông tin và Truyền thông. Ngoài phần

mở đầu, kết luận và tài liệu tham khảo, nội dung luận văn được trình bày thành ba chương:

Chương 1 trình bày các khái niệm cơ bản về phân cụm dữ liệu và bài toán phân đoạn ảnh X-quang nha khoa.

Chương 2, tác giả tìm hiểu một số thuật toán phân cụm dữ liệu trong đó tập trung nghiên cứu thuật toán phân cụm dữ liệu nửa giám sát.

Chương 3 là kết quả thực nghiệm cho thuật toán phân cụm nửa giám sát đối với bài toán phân đoạn ảnh X-quang nha khoa.