

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

PHẠM THỊ THU

**CÁC THUẬT TOÁN PHÂN CỤM DỮ LIỆU
VÀ ỨNG DỤNG TRONG PHÂN LOẠI PROTEIN**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên – 2015

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

PHẠM THỊ THU

**CÁC THUẬT TOÁN PHÂN CỤM DỮ LIỆU VÀ
ỨNG DỤNG TRONG PHÂN LOẠI PROTEIN**

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01 01

Người hướng dẫn khoa học

PGS.TS. Đoàn Văn Ban

Thái Nguyên - 2015

LỜI CẢM ƠN

Để hoàn thành chương trình cao học và viết luận văn này, tôi đã nhận được sự hướng dẫn, giúp đỡ và góp ý nhiệt tình của quý thầy cô trường Đại học Công nghệ thông tin và Truyền thông. Đặc biệt là những thầy cô ở Viện công nghệ thông tin Hà Nội đã tận tình dạy bảo cho tôi suốt thời gian học tập tại trường.

Tôi xin gửi lời cảm ơn sâu sắc đến PGS.TS Đoàn Văn Ban đã dành nhiều thời gian và tâm huyết hướng dẫn tôi hoàn thành luận văn này.

Mặc dù tôi đã có nhiều cố gắng hoàn thiện luận văn bằng tất cả năng lực của mình, tuy nhiên không thể tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp quý báu của quý thầy cô và các bạn.

Tôi xin chân thành cảm ơn!

LỜI CAM ĐOAN

Tôi xin cam đoan tất cả các nội dung của luận văn này hoàn toàn được hình thành và phát triển từ quan điểm của chính cá nhân tôi, dưới sự hướng dẫn chỉ bảo của PGS.TS Đoàn Văn Ban. Các số liệu kết quả có được trong luận văn tốt nghiệp là hoàn toàn trung thực.

Học viên

Phạm Thị Thu

BẢNG KÝ HIỆU CÁC CHỮ VIẾT TẮT

Chữ viết tắt	Nghĩa tiếng anh	Nghĩa tiếng việt
KDD	Knowledge Discovery in Database	Khám phá tri thức trong cơ sở dữ liệu
CSDL	Data base	Cơ sở dữ liệu
KPDL		Khai phá dữ liệu
CURE	Clustering Using Representatives	Phân cụm dữ liệu sử dụng điểm đại diện
CLARA	Clustering Large Application	Thuật toán phân cụm ứng dụng lớn
SoT	Self-organizing Trees	Cây tự tổ chức
DNA	DesoxyriboNucleic Acid	Phân tử nucleic acid mang thông tin di truyền mã hóa cho hoạt động sinh trưởng và phát triển của các dạng sống
RNA	RiboNucleic Acid	Là một trong hai loại axit nucleic, là cơ sở di truyền ở cấp độ phân tử.
rRNA	ribosome RNA	Là ARN mã hóa và mang thông tin từ AND
tRNA	transfer RNA	Là RNA vận chuyển
mRNA	messenger RNA	RNA thông tin
SCOP	Structural Classification of Proteins	Phân loại cấu trúc các protein
CATH	Class Architecture Topology Homologous superfamily	Phân loại cấu trúc protein với CATH
DDD	Dali Domain Dictionary	Từ điển miền Dali
PDB	Protein Data Bank	Ngân hàng dữ liệu protein
FSSP	Families of Structurally Similar Proteins	Dòng họ protein với cấu trúc tương tự

DANH MỤC HÌNH VẼ

<i>Số hiệu hình & tên hình vẽ</i>	<i>Trang</i>
Hình 1.1. Ví dụ phân cụm của tập dữ liệu vay nợ thành 3 cụm	6
Hình 1.2. Các chiến lược phân cụm phân cấp	15
Hình 1.3. Một số hình dạng khám phá bởi phân cụm trên mật độ	16
Hình 1.4. Mô hình cấu trúc dữ liệu lưới	18
Hình 2.1. Các thiết lập để xác định danh giới các cụm ban đầu	25
Hình 2.2. Tính toán trọng tâm của các cụm mới	26
Hình 2.3. Minh họa trực quan quá trình phân cụm	28
Hình 2.4. Phân cụm Chameleon	31
Hình 2.5. Sự di chuyển về trung tâm cụm	34
Hình 2.6. Sự sáp nhập của các cụm	35
Hình 2.7. Cụm dữ liệu khai phá bởi thuật toán CURE	35
Hình 2.8. Nguyên lý chung của AntTree	37
Hình 2.9. Kiến trúc khác nhau giữa SOM và SoT	40
Hình 2.10. Phân việc từ cây $tree_{c..}$ cho $tree_c$	44
Hình 2.11. Tách $subtree_x$ khỏi cây $tree_{c..}$ và đưa vào list	44
Hình 2.12. Tái liên kết $subtree_x$ vào $tree_c$	45
Hình 3.1. Thuyết trung tâm của sinh học phân tử	47
Hình 3.2. Cấu trúc DNA	48
Hình 3.3. Sự phát triển của cấu trúc dữ liệu protein	51
Hình 3.4. Dữ liệu đầu vào của thuật toán	57
Hình 3.5. Giao diện chọn bộ dữ liệu	65

Hình 3.6. Thông tin về bộ dữ liệu	66
Hình 3.7. Kết quả phân cụm với số tâm cụm bằng 10	67
Hình 3.8. Kết quả phân cụm bằng SoT với số tâm cụm bằng 10	67
Hình 3.9. Giao diện hiển thị 10 phân cụm trong thuật toán SoT	68
Hình 3.10. Chi tiết phân cụm thứ tám trong thuật toán SoT	68
Hình 3.11. Tập tin kết quả phân cụm clara	69

DANH MỤC BẢNG

Bảng 3.1. Nguồn tài nguyên cho phân loại cấu trúc protein	52
Bảng 3.2. Các cấp độ chính của CATH	53

MỤC LỤC

LỜI CẢM ƠN	i
LỜI CAM ĐOAN	ii
BẢNG KÝ HIỆU CÁC CHỮ VIẾT TẮT	iii
DANH MỤC HÌNH VẼ.....	iv
MỞ ĐẦU.....	1
CHƯƠNG 1. KHAI PHÁ DỮ LIỆU	3
1.1. Khái niệm chung	3
1.2. Phân lớp dữ liệu	4
1.3. Phân cụm dữ liệu.....	5
1.3.1. Tổng quan về phân cụm dữ liệu	5
1.3.2. Các yêu cầu cơ bản đối với các kỹ thuật phân cụm dữ liệu.....	9
1.3.3. Các kiểu dữ liệu trong phân cụm dữ liệu	9
1.3.4. Độ đo trong phân cụm dữ liệu.....	11
1.3.5. Các kỹ thuật tiếp cận với bài toán phân cụm	13
1.4. Luật kết hợp	20
1.4.1. Một số khái niệm cơ sở	20
1.4.2. Các tính chất sau với tập mục phổ biến	21
1.4.3. Các tính chất với luật kết hợp	21
1.5. Một số ứng dụng của phân cụm dữ liệu.....	22
1.5.1. Ứng dụng trong tin sinh học	22
1.5.2. Ứng dụng trong phân loại đối tượng văn bản	23
1.5.3. Ứng dụng trong phân đoạn ảnh, nhận dạng	23
1.6. Kết luận chương 1	24
CHƯƠNG 2. CÁC THUẬT TOÁN PHÂN CỤM	25
2.1. Thuật toán K-means	25
2.2. Thuật toán CHAMELEON	29
2.3. Thuật toán CLARA	32
2.4. Thuật toán CURE	33
2.5. Thuật toán AntTree	37
2.6. Thuật toán cây tự tổ chức SoT	39
2.7. Kết luận chương 2	46
CHƯƠNG 3. CHƯƠNG TRÌNH THỬ NGHIỆM	47

3.1. Protein và các kỹ thuật phân loại Protein.....	47
3.1.1. Thuyết trung tâm của sinh học phân tử.....	47
3.1.2. Các kỹ thuật phân loại Protein	50
3.2. Cài đặt thử nghiệm thuật toán phân cụm dữ liệu trong phân loại Protein	55
3.2.1. Phát biểu bài toán	55
3.2.2. Mô tả dữ liệu	56
3.2.3. Chuẩn bị dữ liệu	57
3.2.4. Môi trường cài đặt và thử nghiệm.....	61
3.3. Nhận xét, đánh giá chương trình thử nghiệm.....	70
3.4. Kết luận chương 3	70
KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU	71
TÀI LIỆU THAM KHẢO.....	72

MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển vượt bậc của công nghệ thông tin, khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin không ngừng được nâng cao. Theo đó, lượng thông tin được lưu trữ trên các thiết bị nhớ không ngừng tăng lên.

Khai phá dữ liệu là quá trình khám phá các tri thức mới có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Quá trình khám phá tri thức là một chuỗi lặp gồm các bước: làm sạch dữ liệu, tích hợp dữ liệu, chọn lựa dữ liệu, đánh giá mẫu, biểu diễn tri thức. Khai phá dữ liệu liên quan đến nhiều lĩnh vực khác nhau như: công nghệ cơ sở dữ liệu, lý thuyết thống kê, học máy, khoa học thông tin, trực quan hóa,...

Vấn đề ứng dụng các kỹ thuật khai phá dữ liệu, phân cụm dữ liệu trong Tin sinh học, một lĩnh vực còn khá mới, đã ra đời, sử dụng các công nghệ của các ngành toán học ứng dụng, tin học, thống kê, khoa học máy tính, trí tuệ nhân tạo, hóa học, sinh học để giải quyết các vấn đề của sinh học. Việc tìm hiểu và nghiên cứu phân loại protein đã nổi lên như một hướng đi mới với những trải nghiệm hướng vào việc khám phá cấu trúc của các phân tử sinh học.

Nghiên cứu và ứng dụng một cách hiệu quả các phương pháp khai phá dữ liệu là vấn đề hấp dẫn, đã và đang thu hút sự quan tâm chẳng những của các nhà nghiên cứu, ứng dụng mà của cả các tổ chức, doanh nghiệp. Do đó, tôi đã chọn đề tài nghiên cứu “ **Các thuật toán phân cụm dữ liệu và ứng dụng trong phân loại Protein** ”