

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

TỔNG ĐÌNH TIẾN

TỐI ƯU HÓA CÁC THÔNG SỐ HỆ MỜ SỬ DỤNG
PHÂN CỤM DỮ LIỆU TRỪ VÀ GIẢI THUẬT DI TRUYỀN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2013

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

TỔNG ĐÌNH TIẾN

TỐI ƯU HÓA CÁC THÔNG SỐ HỆ MỜ SỬ DỤNG
PHÂN CỤM DỮ LIỆU TRỪ VÀ GIẢI THUẬT DI TRUYỀN

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

PGS. TS LÊ BÁ DŨNG

THÁI NGUYÊN - 2014

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
Tống Đình Tiến

LỜI CẢM ƠN

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

Em cũng xin được gửi lời 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.

Em cũng xin gửi lời cảm ơn tới Ban giám hiệu trường Đại học Công Nghệ Thông Tin Và Truyền Thông, ban đào tạo sau đại học đã tạo điều kiện thuận lợi cho tôi tham gia khóa học và hoàn thành luận văn.

Cuối cùng 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.

Một lần nữa, em xin chân thành cảm ơn.

Thái Nguyên, 15 tháng 3 năm 2014

Học viên

Tống Đình Tiến

MỤC LỤC

LỜI CAM ĐOAN

LỜI CẢM ƠN

MỤC LỤC	i
DANH MỤC CÁC KÍ HIỆU, CHỮ CÁI VIẾT TẮT	iii
DANH MỤC CÁC BẢNG BIỂU	iv
DANH MỤC CÁC HÌNH.....	v
MỞ ĐẦU	1
CHƯƠNG 1 TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU VÀ CÁC THUẬT TOÁN TRONG PHÂN CỤM DỮ LIỆU	3
1.1. Khái niệm và mục tiêu của phân cụm dữ liệu	3
1.1.1. Khái niệm về phân cụm dữ liệu	3
1.1.2. Mục tiêu của phân cụm dữ liệu.....	4
1.2. Các yêu cầu của phân cụm dữ liệu	5
1.3. Các ứng dụng của phân cụm dữ liệu	6
1.4. Các kỹ thuật tiếp cận và một số thuật toán cơ bản trong phân cụm dữ liệu.....	7
1.4.1. Các phương pháp phân cụm phân hoạch - Partitioning Methods.....	7
1.4.2. Phương pháp phân cụm phân cấp - Hierarchical Methods.....	9
1.4.3. Phương pháp phân cụm dựa trên mật độ - Density-Based Methods	10
1.4.4. Phương pháp phân cụm dựa trên lưới - Grid-Based Methods.....	10
1.4.5. Phương pháp phân cụm dựa trên mô hình - Model-Based Clustering Methods	11
1.4.6. Phương pháp phân cụm có dữ liệu ràng buộc	12
1.5. Một số thuật toán cơ bản trong phân cụm dữ liệu.....	13
1.5.1. Thuật toán K-means	13
1.5.2. Thuật toán CURE.....	15
1.5.3. Thuật toán DBSCAN.....	17
1.5.4. Thuật toán STING	18
1.5.5. Thuật toán EM	19

CHƯƠNG 2 PHƯƠNG PHÁP TỐI ƯU HÓA CÁC THÔNG SỐ HỆ MỜ SỬ DỤNG PHÂN CỤM DỮ LIỆU TRỪ VÀ GIẢI THUẬT DI TRUYỀN.....	22
2.1. Phân cụm dữ liệu trừ	22
2.1.1. Khái niệm về phân cụm dữ liệu trừ	22
2.1.2. Thuật toán phân cụm dữ liệu trừ.....	23
2.1.2.1. Thuật toán phân cụm dữ liệu trừ (SC - Subtractive Clustering)	23
2.1.2.2. Thuật toán phân cụm dữ liệu trừ mờ (FSC - Fuzzy Subtractive Clustering).....	27
2.1.2.3. Thuật toán phân cụm trừ mờ loại hai khoảng	29
2.2. Giải thuật di truyền.....	31
2.2.1. Giải thuật di truyền và các phương pháp tối ưu truyền thống	31
2.2.2. Một giải thuật di truyền đơn giản	34
2.2.3. Giải thuật di truyền trong công việc-sự mô phỏng bằng tay	38
2.2.4. Lợi ích trong việc tìm kiếm những tương đồng quan trọng	41
2.2.4.1. Những khuôn mẫu giống nhau.....	42
2.2.4.2. Cái nào sẽ tồn tại và cái nào sẽ bị loại bỏ	44
CHƯƠNG 3 ỨNG DỤNG PHÂN CỤM DỮ LIỆU TRỪ VÀ GIẢI THUẬT DI TRUYỀN TRONG VIỆC ĐO VÀ ĐIỀU KHIỂN NHIỆT	51
3.1. Phát biểu bài toán	51
3.2. Một số ứng dụng của phân cụm dữ liệu trừ cho đo và điều khiển tự động ...	52
3.2.1. Ứng dụng thuật toán phân cụm trừ cho xây dựng hệ luật	52
3.2.2. Xây dựng hệ luật điều khiển mờ.....	53
3.2.3. Tối ưu các thông số cho luật điều khiển mờ.....	54
3.3 Thử nghiệm sử dụng các thuật toán phân cụm dữ liệu trừ, giải thuật di truyền để xây dựng chương trình đo và điều khiển nhiệt độ.	59
3.3.1. Các chức năng của chương trình.	59
3.3.2. Giao diện chương trình	59
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	62
TÀI LIỆU THAM KHẢO	63

DANH MỤC CÁC KÍ HIỆU, CHỮ CÁI VIẾT TẮT

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

DANH MỤC CÁC BẢNG BIỂU

Bảng 2.1. Ví dụ về các chuỗi và các độ thích nghi tương ứng	35
Bảng 2.2. Ví dụ về một quần thể cỡ 4 ban đầu	39
Bảng 2.3 Quần thể sau khi ghép chéo.....	40
Bảng 3.1. Hệ luật mờ của hệ thống điều khiển cho ban đầu.....	54
Bảng 3.2. Kết quả phân cụm trừ	55

DANH MỤC CÁC HÌNH

Hình 1.1. Các cách phân cụm phân cấp	09
Hình 1.2. Cấu trúc phân cụm dựa trên lưới điểm.....	11
Hình 1.3. Cấu trúc phân cụm dựa trên sự ràng buộc.....	12
Hình 1.4. Thuật toán K-Means.....	14
Hình 2.1. Hai nhóm dữ liệu của phân cụm trừ mờ.....	26
Hình 2.2. Kết quả phân cụm dữ liệu của SC theo Chiu	29
Hình 2.3. a-b Sự phụ thuộc của SC vào các tham số r_a và h	29
Hình 2.4. a-b. Sự phụ thuộc của SC vào tham số m	30
Hình 2.5. Sơ đồ thuật toán phân cụm trừ loại 2 khoảng	33
Hình 2.6 Sự sinh sản đơn giản phân bố các chuỗi con cháu nhờ sử dụng bánh xe Rulet với các khe hở tỉ lệ với bộ thích nghi. Bánh xe mẫu được dựa trên bảng 2.1 và bảng 2.2	35
Hình 2.7. Lược đồ của sự ghép chéo đơn giản chỉ ra sự sắp xếp hai chuỗi và trao đổi thông tin giữa hai chuỗi sd vị trí trao đổi một cách ngẫu nhiên.....	37
Hình 3.1. Luật được hình thành qua phép chiếu vào không gian đầu vào X	52
Hình 3.2. Phân cụm trừ cho bảng 3.1	55
Hình 3.3. Sơ đồ giải thuật di truyền	56
Hình 3.4. Sơ đồ hệ thống điều khiển AQM tổng quát	57
Hình 3.5. Kết quả mô phỏng cho hệ điều khiển trước và sau phân cụm trừ.....	58
Hình 3.6. Dữ liệu thu thập cho hệ điều khiển	59
Hình 3.7. Hệ luật được hình thành qua phân cụm	60
Hình 3.8. Biểu diễn hệ luật dwosi dạng đồ thị.....	60
Hình 3.9. Hàm thuộc và mặt suy diễn được tạo	61
Hình 3.10 a Tín hiệu ra tiệm cận với tín hiệu yêu cầu	61
Hình 3.10 b Tác động điều khiển	61

MỞ ĐẦU

Trong ngành khoa học máy tính, việc đi tìm kiếm lời giải tối ưu cho các bài toán là vấn đề được các nhà khoa học máy tính đặc biệt rất quan tâm. Mục đích chính của các thuật toán tìm kiếm lời giải là tìm ra lời giải tối ưu nhất cho bài toán trong thời gian nhỏ nhất. Các thuật toán như tìm kiếm không có thông tin, vét cạn (tìm kiếm trên danh sách, trên cây hoặc đồ thị) sử dụng phương pháp đơn giản nhất và trực quan nhất hoặc các thuật toán tìm kiếm có thông tin sử dụng Heuristics để áp dụng các tri thức về cấu trúc của không gian tìm kiếm nhằm giảm thời gian cần thiết cho việc tìm kiếm được sử dụng nhiều nhưng chỉ với không gian tìm kiếm nhỏ và không hiệu quả khi tìm kiếm trong không gian tìm kiếm lớn. Tuy nhiên, trong thực tiễn có rất nhiều bài toán tối ưu với không gian tìm kiếm rất lớn cần phải giải quyết. Vì vậy, việc đòi hỏi thuật giải chất lượng cao và sử dụng kỹ thuật trí tuệ nhân tạo đặc biệt rất cần thiết khi giải quyết các bài toán có không gian tìm kiếm lớn. Giải thuật di truyền (Genetic algorithm) là một trong những kỹ thuật tìm kiếm lời giải tối ưu đã đáp ứng được yêu cầu của nhiều bài toán và ứng dụng.

Thuật giải di truyền đã được phát minh ra để bắt chước quá trình phát triển tự nhiên trong điều kiện quy định sẵn của môi trường. Các đặc điểm của quá trình này đã thu hút sự chú ý của John Holland (ở đại học Michigan) ngay từ những năm 1970. Holland tin rằng sự gắn kết thích hợp trong thuật giải máy tính có thể tạo ra một kỹ thuật giúp giải quyết các vấn đề khó khăn giống như trong tự nhiên đã diễn ra-thông qua quá trình tiến hóa.

Trong thực tế cuộc sống, chúng ta bắt gặp nhiều bài toán như dự đoán thị trường chứng khoán, dự đoán lưu lượng nước, dự đoán lượng ga tiêu thụ, dự đoán năng lực sản xuất, định giá tài sản,... Đó là các bài toán thuộc lớp bài toán dự đoán và phân lớp, có thể xem là các bài toán cơ bản và có nhiều ứng dụng trong thực tiễn. Đã có nhiều phương pháp được đưa ra để giải quyết các lớp bài toán đó như phương pháp thống kê, cây quyết định, mạng nơron nhân tạo... Việc áp dụng các phương pháp của khai phá dữ liệu (đặc biệt là các phương pháp học máy mạng Nơron kết