

LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT VIỆT NAM

FAIR

KỶ YẾU HỘI NGHỊ KHOA HỌC CÔNG NGHỆ
QUỐC GIA LẦN THỨ XII

**NGHIÊN CỨU CƠ BẢN VÀ ỨNG DỤNG
CÔNG NGHỆ THÔNG TIN**

HUẾ, 07-08/6/2019

**Proceedings of the 12th National Conference on
Fundamental and Applied Information Technology Research
(FAIR'2019)**



MỤC LỤC

STT	TÊN BÀI BÁO	Trang
1.	A SENSE TAGGING ALGORITHM USING UNSUPERVISED METHOD <i>Quang Duc Huynh, Phuoc Tran, Huu Nguyen</i>	1
2.	ADVANCED INTELLIGENT IDENTIFICATION OF PMSM PARAMETER USING MODIFIED JAYA OPTIMIZATION ALGORITHM <i>Pham Quoc Khanh, Ho Pham Huy Anh</i>	8
3.	AN EMPIRICAL STUDY OF THE IMPACT OF THE MPOS SYSTEM ON THE PROCESS CHANGE OF RESTAURANTS <i>Nhu Hang Ha, Duc Man Nguyen, Chia An Liu, Thu Van Van, Anh Dao Nguyen, Quyet Thang Huynh</i>	17
4.	AN IMPROVED POSITIVE SELECTION ALGORITHM FOR FLOW-BASED INTRUSION DETECTION <i>Nguyen Van Truong, Nguyen Xuan Hoai</i>	23
5.	AUTOMATED PNEUMONIA DETECTION IN X-RAY IMAGES VIA DEPTHWISE SEPARABLE CONVOLUTION BASED LEARNING <i>Nghia Duong Trung, Tuyen Tran Ngoc, Hiep Xuan Huynh</i>	32
6.	BAG-SVM-SGD FOR DEALING WITH LARGE-SCALE MULTI-CLASS DATASETS <i>Thanh Nghi Do, Huu Hoa Nguyen, The Phi Pham</i>	41
7.	BLACK FRIDAY SALE PREDICTION VIA EXTREME GRADIENT BOOSTED TREES <i>Nghia Duong Trung, Tan Dang Thien, Tien Dao Luu, Hiep Xuan Huynh</i>	49
8.	CẢI THIỆN HIỆU NĂNG CỦA CƠ CHẾ QUẢN LÝ HÀNG ĐỘI ENRED DỰA TRÊN CHIỀU DÀI HÀNG ĐỘI TRUNG BÌNH <i>Mai Thành Trung, Võ Thanh Tú</i>	57
9.	CẢI TIẾN MỘT SỐ THUẬT TOÁN HEURISTIC GIẢI BÀI TOÁN CLIQUE LỚN NHẤT <i>Phan Tấn Quốc, Huỳnh Thị Châu Ái</i>	64
10.	CẢI TIẾN THUẬT TOÁN XỬ LÝ TRUY VẤN TRÊN CƠ SỞ DỮ LIỆU ĐÓ THỊ NEO4J PHÂN TÁN <i>Phạm Hữu Mão, Ngô Thanh Hùng</i>	73
11.	CẢI TIẾN TRỌNG SỐ KẾT HỢP KỸ THUẬT RÚT TRÍCH ĐA ĐẶC ĐIỂM TRONG VIỆC ĐÒ TÌM NHỮNG BÁO CÁO LỖI TRÙNG NHAU <i>Nhan Minh Phúc, Nguyễn Hoàng Duy Thiện, Dương Ngọc Vân Khanh</i>	78
12.	CHATBOT CHO SINH VIÊN CÔNG NGHỆ THÔNG TIN <i>Đỗ Thanh Nghị, Hoàng Tùng</i>	85
13.	DATA MINING IN HEALTHCARE SYSTEM ON PATIENTS CLINICAL SYMPTOMS DATASET <i>Trần Đình Toàn, Huỳnh Thị Châu Lan, Trần Văn Thọ, Lê Minh Hưng, Trần Văn Lăng</i>	92
14.	DESIGN OF A FUZZY MEDICAL EXPERT SYSTEM FOR CONSULTING PROSTATE DISEASES <i>Mai Ngọc Anh, Phạm Duy Dương</i>	102
15.	DỰ BÁO GIÁ BITCOIN BẰNG KẾT HỢP MÔ HÌNH ARIMA VÀ MẠNG NƠRON <i>Lê Hữu Vinh, Nguyễn Đình Thuận</i>	110

16.	DỰ BÁO MỨC NƯỚC SÔNG MEKONG SỬ DỤNG LSTM VÀ DỮ LIỆU QUAN TRÁC THƯỢNG NGUỒN	119
	<i>Trần Nguyễn Minh Thư, Nguyễn Hồng Hải, Phạm Trường An</i>	
17.	ĐÁNH GIÁ CÁC PHƯƠNG PHÁP DỰA TRÊN DEEP LEARNING CHO BÀI TOÁN PHÁT HIỆN LOGO	127
	<i>Nguyễn Nhật Duy, Đỗ Văn Tiến, Ngô Đức Thành, Huỳnh Ngọc Tín, Lê Đình Duy</i>	
18.	ĐÁNH GIÁ HIỆU NĂNG CỦA MỘT SỐ GIAO THỨC ĐỊNH TUYẾN THEO YÊU CẦU TRÊN MÔ HÌNH ĐIỂM NHÓM DI ĐỘNG	135
	<i>Lê Vũ, Nguyễn Tấn Khôi, Võ Thanh Tú</i>	
19.	ĐỀ XUẤT MÔ HÌNH KIẾN TRÚC HỆ THỐNG THÔNG TIN TỔNG THỂ TẠI CÁC TRƯỜNG ĐẠI HỌC SỬ PHẠM Ở VIỆT NAM	144
	<i>Nguyễn Duy Hải, Lê Văn Năm</i>	
20.	ĐỊNH VỊ NGUỒN PHÁT SÓNG VÔ TUYẾN BẰNG PHƯƠNG PHÁP DRSSI CẢI TIẾN	152
	<i>Lê Hải Toàn, Nguyễn Thanh Bình, Lương Vinh Quốc Danh và Nguyễn Thị Trâm</i>	
21.	FHURIM: THUẬT TOÁN KHAI PHÁ TẬP MỤC HỮU ÍCH CAO HIỂM	159
	<i>Huỳnh Triệu Vỹ, Lê Quốc Hải, Trương Ngọc Châu</i>	
22.	GIẢI PHÁP QUẢN LÝ TÀI SẢN NGĂN CHẶN BẰNG CÔNG NGHỆ BLOCKCHAIN	168
	<i>Trương Minh Tuyền, Nguyễn Hoàng Tùng, Huỳnh Phước Hải, Lê Hoàng Anh, Nguyễn Văn Hòa</i>	
23.	GIAO THỨC ĐỊNH TUYẾN AN NINH SỬ DỤNG CƠ CHẾ XÁC THỰC MẶT KHẨU SỬ DỤNG MỘT LẦN	176
	<i>Lê Đức Huy, Lương Thái Ngọc, Võ Thanh Tú</i>	
24.	GIẤU TIN THUẬN NGHỊCH BẰNG DỰ BÁO TRÊN NGŨ CẢNH ĐIỂM ẢNH KẾT HỢP EMD	183
	<i>Trâm Hoàng Nam, Huỳnh Văn Thanh, Võ Thành C, Dương Ngọc Vân Khanh, Nguyễn Thái Sơn</i>	
25.	GIẤU TIN THUẬN NGHỊCH CHO ẢNH STEREO DỰA TRÊN PHƯƠNG PHÁP DỊCH CHUYỂN HISTOGRAM VÀ EMD	191
	<i>Huỳnh Văn Thanh, Võ Phước Hưng, Nguyễn Thái Sơn, Trâm Hoàng Nam, Đỗ Thanh Nghị</i>	
26.	HỆ HỖ TRỢ QUYẾT ĐỊNH TRONG CHẨN ĐOÁN BỆNH: TIẾP CẬN TỪ HỆ MỜ PHỨC	199
	<i>Lương Thị Hồng Lan, Trần Thị Ngân, Hoàng Thị Minh Châu, Lê Bích Liên, Roãn Thị Ngân</i>	
27.	HỆ THỐNG PHÂN LOẠI ẢNH XUẤT HUYẾT NÃO THEO HƯỚNG TIẾP CẬN XỬ LÝ DỮ LIỆU LỚN	209
	<i>Phan Anh Cang, Phan Thượng Cang, Phạm Duy Khang, La Ngọc Nguyễn, Trần Hồ Đạt</i>	
28.	HỆ THỐNG THU THẬP CHỈ SỐ NƯỚC TIÊU THỤ TỰ ĐỘNG ỨNG DỤNG CÔNG NGHỆ TRUYỀN THÔNG LoRa	217
	<i>Lê Hoàng Văn, Lê Tuấn Anh, Lương Vinh Quốc Danh và Nguyễn Thị Trâm</i>	
29.	KẾT HỢP KỸ THUẬT GOM NHÓM VÀ PHẢN HỒI TƯƠNG ĐỒNG TRONG TÌM KIẾM ẢNH	225
	<i>Nguyễn Ti Hon, Hà Thị Phương Anh, Phạm Thế Phi</i>	
30.	KỸ THUẬT HỌC SÂU ĐỂ GIẢI QUYẾT BÀI TOÁN CHẨN ĐOÁN BỆNH LAO PHỔI	234
	<i>Đoàn Thiện Minh, Trần Văn Lăng, Văn Đình Vỹ Phương, Phan Mạnh Thương</i>	
31.	KỸ THUẬT PHÁT HIỆN NHANH VÀ CHẠM CỦA VẢI TRONG THỰC TẠI ẢO SỬ DỤNG PHƯƠNG PHÁP TÍNH TOÁN SONG SONG	239
	<i>Nghiêm Văn Hưng, Đặng Văn Đức, Trịnh Hiền Anh, Hoàng Việt Long, Nguyễn Văn Căn</i>	

32.	LEVERAGE THE BLOCKCHAIN TECHNOLOGY TO MANAGE SMART CONTRACT IN ASSET TRADING	247
	<i>Quoc Nhan Vo, Nhat Phuong Tran, Van Dat Ngo, Van Ha Truong, Quyet Thang Huynh, Nhu Hang Ha, Duc Man Nguyen</i>	
33.	MÔ HÌNH CHUỖI THỜI GIAN MỜ BẠC CAO HAI NHÂN TỐ KẾT HỢP VỚI TỐI ƯU BẦY ĐÀN CHO DỰ BÁO NHIỆT ĐỘ VÀ THỊ TRƯỜNG CHỨNG KHOÁN	257
	<i>Nghiêm Văn Tĩnh, Nguyễn Công Điều</i>	
34.	MỘT CÁI TIẾN VỀ ĐIỀU KHIỂN CHẤP NHẬN LẬP LỊCH DỰA TRÊN DỰ BÁO TỐC ĐỘ CHỤM ĐÈN KẾT HỢP ĐƯỜNG TRẺ FDL	268
	<i>Phạm Trung Đức, Võ Viết Minh Nhật, Đặng Thanh Chương</i>	
35.	MỘT KỸ THUẬT ĐỊNH VỊ ĐỐI TƯỢNG TRONG HỆ THỐNG CAMERA GIÁM SÁT PHỤC VỤ THEO DÕI TRỰC QUAN	277
	<i>Đỗ Năng Toàn, Hà Mạnh Toàn, Phạm Bá Máy, Ngô Đức Vĩnh</i>	
36.	MỘT MÔ HÌNH HỌC SÂU CHO PHÁT HIỆN CẢM XÚC KHUÔN MẶT	284
	<i>Nguyễn Thị Duyên, Trương Xuân Nam, Nguyễn Thanh Tùng</i>	
37.	MỘT PHƯƠNG PHÁP CHUYỂN ĐỔI MÔ HÌNH STER SANG OWL ONTOLOGY	290
	<i>Nguyễn Văn Toàn, Võ Hoàng Liên Minh, Nguyễn Thế Anh, Hoàng Quang</i>	
38.	MỘT PHƯƠNG PHÁP LỰA CHỌN THUỘC TÍNH GOM CỤM SỬ DỤNG LÝ THUYẾT THÔNG TIN	298
	<i>Phạm Công Xuyên, Nguyễn Thanh Tùng</i>	
39.	MỘT PHƯƠNG PHÁP TRA CỨU ẢNH HỌC BIỂU DIỄN VÀ HỌC ĐA TẬP CHO GIÁM CHIẾU VỚI THÔNG TIN TỪ NGƯỜI DÙNG	307
	<i>Cù Việt Dũng, Nguyễn Hữu Quỳnh, Ngô Quốc Tạo, Trần Thị Minh Thu</i>	
40.	MỘT PHƯƠNG PHÁP XÂY DỰNG NGỮ LIỆU SONG NGỮ ANH-VIỆT TỪ NGUỒN TÀI NGUYÊN INTERNET	315
	<i>Dương Minh Hùng, Lê Mạnh Thạnh, Lê Trung Hiếu</i>	
41.	MỘT THUẬT TOÁN THUYẾT VẤN ẢNH SỐ MẠNH DỰA TRÊN DWT, DCT, SVD VÀ ĐẶC TRƯNG SIFT	322
	<i>Võ Thành C, Võ Phước Hưng, Trầm Hoàng Nam, Nguyễn Thái Sơn, Đỗ Thanh Nghị</i>	
42.	NÂNG CAO CHẤT LƯỢNG ẢNH VIÊN THÂM DỰA TRÊN PHÂN CỤM BẢN GIÁM SÁT MỜ	330
	<i>Nguyễn Tu Trung, Trần Mạnh Tuấn, Đặng Thị Thu Hiền, Nguyễn Huy Đức, Kiều Tuấn Dũng, Nguyễn Văn Nam, Đỗ Oanh Cường</i>	
43.	NGHIÊN CỨU CƠ CHẾ TRUYỀN LẠI CHỨM CÓ ĐIỀU KHIỂN TRÊN MẠNG TCP/OBS	377
	<i>Dương Phước Đạt, Lê Mạnh Thạnh, Võ Viết Minh Nhật</i>	
44.	NGHIÊN CỨU ĐỀ XUẤT MÔ HÌNH MẠNG ĐỘNG CHO BÀI TOÁN LẬP LỊCH TÀI NGUYÊN TRONG MẠNG LONG TERM EVOLUTION (LTE)	345
	<i>Lê Minh Tuấn, Lê Hoàng Sơn, Phạm Thị Minh Phương, Vũ Như Lân, Đặng Thanh Hải, Đinh Thu Khánh</i>	
45.	NHẬN DẠNG CÁC BỘ PHẬN TRÊN ĐỐI TƯỢNG 3D DỰA VÀO KỸ THUẬT HỌC SÂU MASK R-CNN	353
	<i>Lê Tiến Mẫu, Nguyễn Tấn Khởi, Romain Raffin</i>	
46.	NHẬN DẠNG TRANG WEB CÓ NỘI DUNG KHIÊU DÂM DỰA TRÊN TEXT VÀ IMAGE	361
	<i>Phan Đình Duy, Nguyễn Văn Thanh, Vũ Đức Lung</i>	

47.	PHÁT HIỆN DGA BOTNET SỬ DỤNG KẾT HỢP NHIỀU NHÓM ĐẶC TRUNG PHẦN LOẠI TÊN MIỀN	369
	<i>Vũ Xuân Hạnh, Hoàng Xuân Dậu</i>	
48.	PHƯƠNG PHÁP GIA TĂNG MA TRẬN ĐỘ HỖ TRỢ TRÊN KHỐI DỮ LIỆU VÀ TRÊN LÁT CẮT KHI TẬP ĐỐI TƯỢNG THAY ĐỔI	379
	<i>Trịnh Đình Thắng, Đỗ Thị Lan Anh, Trần Minh Tuyền, Cao Hồng Huệ</i>	
49.	PHƯƠNG PHÁP LẬP GIẢI BÀI TOÁN BIÊN TAM ĐIỀU HÒA PHI TUYẾN	388
	<i>Nguyễn Quốc Hưng, Đặng Quang Á, Vũ Vinh Quang</i>	
50.	QUẢN LÝ QUY TRÌNH XỬ LÝ HỒ SƠ LIÊN TÓ CHỨC TRONG HỆ THỐNG THÔNG TIN DỊCH VỤ CÔNG CỦA CHÍNH PHỦ ĐIỆN TỬ	393
	<i>Tạ Tuấn Anh</i>	
51.	SECOND ORDER MUTATION TESTING FOR LUSTRE PROGRAMS	399
	<i>Le Van Phoi, Nguyen Thanh Binh, Le Thi Thanh Binh</i>	
52.	SEMANTIC EXTRACTION FROM HTML DATA TO OWL ONTOLOGY	406
	<i>Pham Thi Thu Thuy</i>	
53.	SỐ SÁNH CÁC ĐỘ ĐO TRONG PHÂN CỤM VĂN BẢN TIẾNG VIỆT	414
	<i>Tô Khánh Toàn, Võ Hải Đăng, Trần Thị Cẩm Tú, Trương Quốc Định, Huỳnh Xuân Hiệp</i>	
54.	SỬ DỤNG BÀI TOÁN THỎA MÃN RÀNG BUỘC ĐỂ SO KHỚP ONTOLOGY MỞ	423
	<i>Quách Xuân Hưng</i>	
55.	SỬ DỤNG MÔ HÌNH HỌC MÁY TRONG HỖ TRỢ ĐIỂN ĐOÁN THỦY LỰC, THỦY VĂN TRÊN HỆ THỐNG BẮC HUNG HẢI	430
	<i>Nguyễn Văn Nam, Trần Mạnh Tuấn, Đặng Thị Thu Hiền, Nguyễn Huy Đức, Kiều Tuấn Dũng, Đỗ Oanh Cường, Nguyễn Tu Trung</i>	
56.	THÍCH ỨNG MIỀN TRONG DỊCH MÁY NƠ RON CHO CẤP NGÔN NGỮ ANH - VIỆT	436
	<i>Phạm Nghĩa Luân, Nguyễn Văn Vinh, Nguyễn Huy Hoàng</i>	
57.	THIẾT KẾ MÔ HÌNH DỮ LIỆU ANCHOR TỪ MÔ HÌNH THỰC THỂ - MỐI QUAN HỆ CÓ YẾU TỐ THỜI GIAN	443
	<i>Nguyễn Thị Lan Anh, Trần Việt Khoa, Nguyễn Việt Liên, Hoàng Quang</i>	
58.	THUẬT TOÁN HIỆU QUẢ KHAI THÁC TẬP TƯƠNG QUAN HIẾM CÓ TRỌNG SỐ KẾT HỢP ĐỘ ĐO ALL-CONFIDENCE	450
	<i>Phan Thành Huân, Lê Hoài Bắc</i>	
59.	THỬ NGHIỆM ỨNG DỤNG KỸ THUẬT MÃ HÓA NÉN TÍN HIỆU ÂM THANH TẠI ĐÀI TIẾNG NÓI VIỆT NAM	460
	<i>Nguyễn Thanh Phong, Hoàng Lê Uyên Thực</i>	
60.	THỰC NGHIỆM TÓM TẮT RÚT TRÍCH VĂN BẢN TIẾNG VIỆT	468
	<i>Lâm Nhựt Khang, Phan Chí Khang, Trần Bảo Ngọc</i>	
61.	TÌM KIẾM ẢNH THEO NGỮ NGHĨA DỰA TRÊN ĐỒ THỊ CỤM	476
	<i>Nguyễn Văn Thịnh, Nguyễn Thị Định, Văn Thế Thành</i>	
62.	TÌM KIẾM TƯƠNG ĐỒNG TRÊN MẠNG DỮ LIỆU KHÔNG ĐỒNG NHẤT	487
	<i>Nguyễn Văn Gia, Đỗ Phúc</i>	
63.	TOWARDS MACHINE LEARNING APPROACHES TO IDENTIFY SHRIMP DISEASES BASED ON DESCRIPTION	494

64.	TRA CỨU ẢNH THEO NGỮ NGHĨA DỰA TRÊN CÂY PHÂN CỤM PHÂN CẤP	502
	<i>Nguyễn Minh Hải, Lê Thị Vĩnh Thanh, Văn Thế Thành, Trần Văn Lăng</i>	
65.	TRÍCH CHỌN ĐẶC TRƯNG VÀ PHÂN TÍCH ẢNH X-QUANG NHA KHOA	512
	<i>Quang Vinh Huỳnh, Trần Đình Khang, Nguyễn Đức Vương, Lê Khả Hải</i>	
66.	ỨNG DỤNG BLOCKCHAIN ĐỂ TĂNG CƯỜNG TÍNH TOÀN VỆ VÀ BẢO MẬT TRONG QUẢN LÝ LƯU TRỮ VÀ CHIA SẼ DỮ LIỆU IOT	520
	<i>Lê Trung Kiên, Phạm Thị Ngọc Mỹ, Nguyễn Hoài Quốc Trung, Phạm Hoàng Anh</i>	
67.	ỨNG DỤNG BRADLEY-TERRY MINORIZATION-MAXIMIZATION ĐỂ HỌC CÁC ĐẶC TRƯNG TRÊN CỞ CỐ ĐỘ PHẢN NHẢNH CAO	527
	<i>Nguyễn Quốc Huy, Đặng Công Quốc</i>	
68.	VỀ MỘT VẤN ĐỀ TƯƠNG ĐƯƠNG LIÊN QUAN ĐẾN TẬP RÚT GỌN TRONG BẢNG QUYẾT ĐỊNH	534
	<i>Vũ Đức Thi</i>	
69.	WIFI SENSOR MOTE - LARGE DATA EXCHANGE'S SOLLUTION FOR IOT PLATFORM	539
	<i>Nguyen Minh Son, Vu Duc Lung</i>	
70.	XÁC ĐỊNH BIÊN U GAN TRONG ẢNH CỘNG HƯỞNG TỬ Ồ BỤNG BA CHIỀU SỬ DỤNG THUẬT TOÁN HỌC NHANH THÔNG TIN CỤC BỘ	546
	<i>Lê Trọng Ngọc, Hồ Đắc Quán, Phạm Thế Bảo, Huỳnh Trung Hiếu</i>	
71.	XÁC ĐỊNH TƯƠNG ĐỒNG XUYÊN NGỮ ANH - VIỆT SỬ DỤNG MÔ HÌNH ĐỒ THỊ	552
	<i>Lê Thành Nguyên, Trần Gia Trọng Nhân, Trần Công Hậu, Đinh Điền</i>	
72.	XÂY DỰNG NGÂN HÀNG CÂU HỎI DỰA TRÊN LÝ THUYẾT TRẮC NGHIỆM HIỆN ĐẠI IRT VÀ ỨNG DỤNG	561
	<i>Nguyễn Tích Lăng</i>	
73.	XÂY DỰNG TỰ ĐỘNG TỬ ĐIỆN VIỆT-ANH VÀ ỨNG DỤNG TRONG LĨNH VỰC DU LỊCH	568
	<i>Nguyễn Tiên Hà, Nguyễn Thị Minh Huyền</i>	
74.	XỬ LÝ CÁC MỆNH ĐỀ VỀ DỮ LIỆU CHIA SẼ CỦA OPENMP TRÊN CÁC HỆ THỐNG SỬ DỤNG BỘ NHỚ PHÂN TÁN	577
	<i>Đỗ Xuân Huyền, Hà Việt Hải, Trần Văn Long</i>	

MỘT PHƯƠNG PHÁP LỰA CHỌN THUỘC TÍNH GOM CỤM SỬ DỤNG LÝ THUYẾT THÔNG TIN

Phạm Công Xuyên, Nguyễn Thanh Tùng

Lac Hong University

pcxuyen@lhu.edu.vn, nttung@lhu.edu.vn

TÓM TẮT: Bài toán gom cụm dữ liệu xuất hiện trong nhiều lĩnh vực khác nhau. Mục tiêu cơ bản của gom cụm là nhóm đối tượng thành các cụm sao cho các đối tượng trong cùng một cụm thì tương tự như nhau hơn là các đối tượng từ các cụm khác nhau. Gần đây, nhiều nhà nghiên cứu quan tâm đến vấn đề gom cụm dữ liệu phạm trù (categorical), trong đó các đối tượng dữ liệu được mô tả bởi các thuộc tính không phải thuộc tính số. Đặc biệt, phương pháp gom cụm phân cấp dữ liệu phạm trù sử dụng lý thuyết tập thô đã thu hút nhiều sự chú ý. Chìa khóa của các phương pháp này là làm thế nào để chọn được một thuộc tính gom cụm tốt nhất tại mỗi thời điểm trong số nhiều thuộc tính ứng viên.

Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất một thuật toán mới MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), cho việc lựa chọn thuộc tính phân cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng độ đo SU (Symmetric Uncertainty), một độ đo lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính; và tìm cách xác định thuộc tính gom cụm sao cho độ tương quan trung bình của nó với các thuộc tính khác đạt giá trị lớn nhất. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp dữ liệu phạm trù.

Từ khóa: Gom cụm, Dữ liệu phân loại, Gom cụm phân cấp, Lý thuyết tập thô, Lựa chọn thuộc tính gom cụm, Độ không chắc chắn đối xứng.

1. GIỚI THIỆU

Gom cụm là một trong những nhiệm vụ chính của khai thác dữ liệu. Nó có thể được định nghĩa như sau. Cho tập gồm n đối tượng $D = \{x_1, x_2, \dots, x_n\}$, trong đó mỗi đối tượng x_i được mô tả bằng một véc tơ M chiều trong không gian đặc trưng đã cho. Gom cụm là việc tìm ra các cụm/nhóm các đối tượng sao cho các đối tượng trong cùng một cụm thì có độ tương tự cao, còn các đối tượng trong các cụm khác nhau thì có độ tương tự thấp [6].

Bài toán gom cụm xuất hiện trong nhiều lĩnh vực khác nhau như Nhận dạng mẫu, Thị giác máy tính, Sinh học, Y học, Truy tìm thông tin,.... Hiện nay, có nhiều thuật toán gom cụm đã được xây dựng và giới thiệu trong các tài liệu. Các thuật toán này có thể được phân đại thể thành hai loại: phân hoạch và phân cấp. Các phương pháp phân hoạch tạo ra một phân hoạch trên tập dữ liệu, tối ưu hóa một hàm tiêu chuẩn. Các phương pháp phân cấp sinh ra một dãy các phân hoạch lồng nhau của tập dữ liệu.

Hầu hết các công trình về gom cụm đều tập trung vào dữ liệu số, nơi mà các tính chất hình học sẵn có có thể được khai thác để định nghĩa một cách tự nhiên khoảng cách giữa các điểm dữ liệu. Tuy nhiên, các ứng dụng khai thác dữ liệu thực tiễn thường gặp phải những tập dữ liệu, trong đó các thuộc tính là những thuộc tính phạm trù (categorical), mà với chúng ta không thể định nghĩa hàm khoảng cách một cách tự nhiên. Gần đây, gom cụm dữ liệu phạm trù đã thu hút sự quan tâm lớn của cộng đồng nghiên cứu khai thác dữ liệu [1, 4, 8, 10, 11, 12, 13]. Một trong các kỹ thuật gom cụm phân cấp dữ liệu phạm trù là sử dụng một dãy các thuộc tính gom cụm. Đầu tiên, coi tất cả các đối tượng là một cụm. Tại mỗi bước tiếp theo một thuộc tính được lựa chọn để phân chia một nút "lá". Quá trình lặp lại cho đến khi đạt được số cụm yêu cầu. Để áp dụng được kỹ thuật này, một vấn đề đặt ra là tại mỗi bước ta phải lựa chọn được, trong số nhiều thuộc tính ứng viên, một thuộc tính tốt nhất theo một tiêu chuẩn xác định để phân cụm các đối tượng.

Gần đây, xuất hiện một số công trình áp dụng lý thuyết tập thô, một công cụ xử lý sự không chắc chắn, trong quá trình lựa chọn thuộc tính gom cụm [8, 10, 11, 12, 13]. Mazlack và cộng sự [12] đề xuất kỹ thuật sử dụng Độ thô toàn phần (Total Roughness - TR) trong lý thuyết tập thô. Theo đó, độ thô toàn phần càng lớn thì chất lượng gom cụm của thuộc tính được chọn sẽ càng cao. Parmar và cộng sự [13] đề xuất MMR (Min-Min-Roughness), một thuật toán thuần túy dựa vào lý thuyết tập thô. Thuật toán MMR xác định thuộc tính gom cụm dựa vào tiêu chuẩn Độ thô cực tiểu (Min-Roughness, MR). Thế nhưng, Herawan và cộng sự [8] đã chỉ ra rằng, MMR chỉ là sự bổ sung của TR, hơn nữa kỹ thuật này có độ phức tạp tính toán cao. Để khắc phục vấn đề này, Herawan và cộng sự [8] đã đề xuất một kỹ thuật khác gọi là MDA (Maximum Dependency Attributes). MDA sử dụng độ phụ thuộc giữa các thuộc tính trong lý thuyết tập thô. Theo Herawan và cộng sự [8], kỹ thuật MDA có hiệu năng tốt hơn TR và MMR. Tuy nhiên, giữa TR, MMR và MDA có sự giống nhau về bản chất. Sự giống nhau này nằm ở chỗ tiêu chuẩn lựa chọn thuộc tính gom cụm của cả ba kỹ thuật đều được xác định chủ yếu thông qua số phần tử có trong xấp xỉ dưới của một thuộc tính đối với các thuộc tính khác.

Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), một thuật toán mới cho việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng Độ không chắc chắn đối xứng (Symmetric Uncertainty - SU), một độ đo của Lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính. Thuộc tính gom cụm được chọn là thuộc tính có độ tương quan trung bình với các thuộc tính khác lớn nhất. Ưu điểm của SU so với Độ lợi thông tin (Information Gain - IG), là SU không thiên vị các thuộc tính có nhiều giá trị. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp.

II. CÁC KHÁI NIỆM CƠ BẢN

Mục này sẽ trình bày một cách vắn tắt một số khái niệm liên quan đến Lý thuyết tập thô [14] và Lý thuyết thông tin [16].

Một tập dữ liệu gom cụm có thể được biểu diễn dưới dạng một bảng, trong đó mỗi hàng biểu diễn một đối tượng, một trường hợp hay một sự kiện, mỗi cột biểu diễn một thuộc tính, một tính chất hay một số đo có thể đo được trên mỗi đối tượng. Trong lý thuyết tập thô, một bảng dữ liệu như vậy được gọi là một hệ thống tin. Một cách hình thức, người ta định nghĩa hệ thống tin như sau.

Định nghĩa 1. Hệ thống tin là một bộ tứ $S = (U, A, V, f)$, trong đó U là một tập hữu hạn, không rỗng các đối tượng, A là một tập hữu hạn, không rỗng các thuộc tính, $V = \bigcup_{a \in A} V_a$ với V_a là tập tất cả các giá trị của thuộc tính a , và $f: U \times A \rightarrow V$ là hàm thông tin, gán giá trị $f(u, a) \in V_a$ cho mỗi cặp $(u, a) \in U \times A$.

Định nghĩa 2. Cho $S = (U, A, V, f)$ là một hệ thống tin, $B \subseteq A$. Hai phần tử $x, y \in U$ được gọi là B -không phân biệt được trong S nếu và chỉ nếu $f(x, a) = f(y, a)$ với mọi $a \in B$.

Ta ký hiệu quan hệ không phân biệt sinh bởi tập thuộc tính B bởi $IND(B)$. Dễ thấy, $IND(B)$ là một quan hệ tương đương và nó sinh ra một phân hoạch (một phép gom cụm) trên U . Phân hoạch trên U sinh bởi $IND(B)$ trong $S = (U, A, V, f)$ được ký hiệu là π_B và lớp tương đương trong π_B chứa $x \in U$, được ký hiệu là $[x]_B$.

Định nghĩa 3. Cho $S = (U, A, V, f)$ là một hệ thống tin, $B \subseteq A$ và $X \subseteq U$. B -xấp xỉ dưới của X , ký hiệu là $\underline{B}(X)$, và B -xấp xỉ trên của X , ký hiệu là $\overline{B}(X)$, được định nghĩa tương ứng như sau:

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\} \quad (1)$$

Định nghĩa trên nói rằng nếu đối tượng $x \in \underline{B}(X)$ thì nó chắc chắn thuộc vào tập X , còn khi $x \in \overline{B}(X)$ thì nó có thể thuộc vào tập X . Hiển nhiên, ta có $\underline{B}(X) \subseteq X \subseteq \overline{B}(X)$. X được gọi là định nghĩa được nếu $\underline{B}(X) = \overline{B}(X)$. trường hợp ngược lại, X được gọi là tập thô với B -biên $BN_B(X) = \overline{B}(X) - \underline{B}(X)$.

Định nghĩa 4. Cho $S = (U, A, V, f)$ là một hệ thống tin, $B \subseteq A$ và $X \subseteq U$. Độ chính xác của xấp xỉ X thông qua B được định nghĩa bởi

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (2)$$

Trong suốt bài báo này, $|X|$ ký hiệu số phần tử của tập X .

Hiển nhiên, $0 \leq \alpha_B(X) \leq 1$. Nếu $\alpha_B(X) = 1$, thì $\underline{B}(X) = \overline{B}(X)$, B -biên của X là tập rỗng, và X là tập rõ đối với B . Nếu $\alpha_B(X) < 1$, thì $\underline{B}(X) \subset \overline{B}(X)$, B -biên của X là khác rỗng, và X là tập thô đối với B .

Định nghĩa 5. Cho $S = (U, A, V, f)$ là một hệ thống tin, $B \subseteq A$ và $X \subseteq U$. Độ thô (roughness) của X đối với B được định nghĩa là

$$\rho_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (3)$$

Định nghĩa 6. Cho $S = (U, A, V, f)$ là một hệ thống tin. Với $P, Q \subseteq A$, ta nói Q phụ thuộc vào P với mức k ($0 \leq k \leq 1$), ký hiệu $P \Rightarrow_k Q$, nếu

$$k = \frac{\sum_{X \in \pi_Q} |P(X)|}{|U|} \quad (4)$$

Định nghĩa 7. Cho $S = (U, A, V, f)$ là một hệ thống tin, $P \subseteq A$. Giả sử P sinh ra phân hoạch $\pi_P = \{P_1, P_2, \dots, P_m\}$ trên U . Khi đó, entropy của P được xác định bởi

$$E(P) = - \sum_{i=1}^m Pr(P_i) \log_2 Pr(P_i) \quad (5)$$

trong đó $Pr(P_i) = |P_i|/|U|$, $i = 1, 2, \dots, m$ và quy định $0 \log_2 0 = 0$.

Định nghĩa 8. Cho $S = (U, A, V, f)$ là một hệ thống tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Entropy có điều kiện của Q đối với P được định nghĩa bởi

$$E(Q|P) = - \sum_{j=1}^n Pr(P_i) \sum_{i=1}^m Pr(Q_j|P_i) \log_2 Pr(Q_j|P_i) \quad (6)$$

trong đó $Pr(Q_j|P_i) = |X_i \cap Y_j|/|Y_j|$, $i = 1, 2, \dots, m$ và $j = 1, 2, \dots, n$.

Định nghĩa 9. Cho $S = (U, A, V, f)$ là một hệ thống tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Entropy đồng thời của P và Q được định nghĩa bởi

$$E(P, Q) = - \sum_{i=1}^m \sum_{j=1}^n Pr(P_i, Q_j) \log_2 Pr(P_i, Q_j) \quad (7)$$

trong đó $Pr(P_i, Q_j) = |X_i \cap Y_j|/|U|$, $i = 1, 2, \dots, m$ và $j = 1, 2, \dots, n$.

Từ các công thức (5), (6) và (7) ta có

$$E(Q|P) = E(P, Q) - E(P) \quad (8)$$

Định nghĩa 10. Cho $S = (U, A, V, f)$ là một hệ thống tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Thông tin tương hỗ (Mutual information) giữa P và Q được định nghĩa bởi

$$I(P; Q) = E(Q) - E(Q|P) = E(P) - E(P|Q) = E(P) + E(Q) - E(P, Q) \quad (9)$$

Định nghĩa 11. [16] Cho $S = (U, A, V, f)$ là một hệ thống tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Độ không chắc chắn đối xứng (Symmetrical uncertainty - SU) giữa P và Q được định nghĩa bởi

$$SU(P, Q) = 2 \frac{I(P; Q)}{E(P) + E(Q)} \quad (10)$$

SU là một độ đo chuẩn hóa, cho phép lượng hóa sự phụ thuộc lẫn nhau của hai thuộc tính P và Q . SU cho phép phản ánh mức độ tương quan ngay cả khi sự tương quan giữa P và Q không phải là tuyến tính [16]. Giá trị của độ đo này nằm trên đoạn $[0, 1]$. Nếu SU bằng 1, thì giá trị của thuộc tính này sẽ hoàn toàn dự đoán được nếu biết giá trị của thuộc tính kia; trường hợp SU bằng 0, hai thuộc tính P và Q là độc lập nhau.

III. BA KỸ THUẬT DỰA TRÊN LÝ THUYẾT TẬP THỎ

Cho $S = (U, A, V, f)$ là một hệ thống tin, $a_i \in A$, $V(a_i)$ là tập các giá trị của thuộc tính a_i ; $X(a_i = \alpha)$ là tập các đối tượng có cùng giá trị thuộc tính a_i là α , nghĩa là $X(a_i = \alpha)$ là một lớp tương đương trong U sinh bởi quan hệ không phân biệt $IND(a_i)$; $\underline{X}_{a_j}(a_i = \alpha)$ là xấp xỉ dưới, và $\overline{X}_{a_j}(a_i = \alpha)$ là xấp xỉ trên của $X(a_i = \alpha)$ đối với a_j .

A. Kỹ thuật TR (Total Roughness) [12]

Input: Tập dữ liệu gom cụm (Hệ thống tin) S .

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ thô trung bình $Rough_{a_j}(a_i)$ của nó đối với mỗi thuộc tính a_j với $j \neq i$, theo công thức

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X | a_i = \alpha_k)}{|V(a_i)|} \quad (11)$$

$$R_{a_j}(X | a_i = \alpha) = \frac{|X_{a_j}(a_i = \alpha)|}{|X_{a_j}(a_i = \alpha)|} \quad (12)$$

Bước 3: Với mỗi $a_i \in A$ tính **độ thô toàn phần** của nó đối với mọi $a_j, i \neq j$, theo công thức

$$TR(a_i) = \frac{\sum_{(j=1) \wedge (i \neq j)}^{|A|} Rough_{a_j}(a_i)}{|A| - 1} \quad (13)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị TR lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \operatorname{argmax}_{a_j \in A} \{TR(a_j)\} \quad (14)$$

End

2. Kỹ thuật MMR (Min-Min-Roughness) [13]

Input: Tập dữ liệu gom cụm (Hệ thống tin) S .

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ thô trung bình $Rough_{a_j}(a_i)$ của nó đối với mỗi thuộc tính a_j với $j \neq i$, theo công thức

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X | a_i = \alpha_k)}{|V(a_i)|} \quad (15)$$

trong đó

$$R_{a_j}(X | a_i = \alpha) = 1 - \frac{|X_{a_j}(a_i = \alpha)|}{|X_{a_j}(a_i = \alpha)|} \quad (16)$$

Bước 3: Với mỗi $a_i \in A$ tính **độ thô nhỏ nhất** của nó $MR(a_i)$ theo công thức

$$MR(a_i) = \min_{(a_j \in A) \wedge (j \neq i)} (Rough_{a_j}(a_i)) \quad (17)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MR nhỏ nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \operatorname{argmin}_{a_j \in A} \{MR(a_j)\} \quad (18)$$

End

C. Kỹ thuật MDA (Maximum degree of Dependency of Attributes) [8]

Input: Tập dữ liệu gom cụm (Hệ thống tin) S .

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ phụ thuộc của a_i vào mỗi a_j với $j \neq i$, theo công thức

$$\gamma_{a_j}(a_i) = \frac{\sum_{x \in U/a_i} |a_j X|}{|U|} \quad (19)$$

Bước 3. Chọn độ phụ thuộc lớn nhất $MD(a_i)$ của mỗi thuộc tính $a_i (a_i \in A)$ như sau

$$MD(a_i) = \max_{(a_j \in A) \wedge (j \neq i)} (\gamma_{a_j}(a_i)) \quad (20)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MD lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \operatorname{argmax}_{a_j \in A} \{MD(a_j)\} \quad (21)$$

IV. KỸ THUẬT MAX-MEAN-SU

Mục này đề xuất MAX-MEAN-SU, một kỹ thuật dựa vào độ không chắc chắn đối xứng trung bình lớn nhất (Maximum Mean of Symmetric Uncertainties) để lựa chọn thuộc tính gom cụm.

Định nghĩa 12. Cho $S = (U, A, V, f)$ là một hệ thống tin, và $a_i \in A$ là một thuộc tính. Độ không chắc chắn đối xứng trung bình giữa a_i với mỗi $a_j \in A, j \neq i$, được xác định bởi

$$MSU(a_i) = \frac{\sum_{j=1, j \neq i}^{|A|} SU(a_i, a_j)}{|A| - 1} \quad (22)$$

Rõ ràng, nếu giá trị MSU của một thuộc tính nào đó càng cao thì tính đại diện của nó cho các thuộc tính còn lại càng lớn. Do đó, nếu sử dụng nó để gom cụm thì chất lượng gom cụm thu được sẽ càng cao. Dựa vào định nghĩa trên, chúng tôi đề xuất thuật toán gom cụm MAX-MEAN-SU như sau.

Input: Tập dữ liệu gom cụm (Hệ thống tin) S .

Output: Thuộc tính gom cụm.

Begin

Bước 1. Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2. Với mỗi thuộc tính a_i xác định độ không chắc chắn đối xứng $SU(a_i, a_j)$ giữa a_i và mỗi $a_j, j \neq i$ theo công thức

$$SU(a_i, a_j) = 2 \times \frac{I(a_i; a_j)}{E(a_i) + E(a_j)} \quad (23)$$

Bước 3. Tính độ không chắc chắn đối xứng trung bình $MSU(a_i)$ của mỗi thuộc tính $a_i \in A$ như sau

$$MSU(a_i) = \frac{\sum_{j=1, j \neq i}^{|A|} SU(a_i, a_j)}{|A| - 1} \quad (24)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MSU lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \operatorname{argmax}_{a_j \in A} \{MSU(a_j)\}$$

End

Ta hãy minh họa thuật toán Max-Mean-SU bằng một ví dụ.

Ví dụ. Bảng 1 mô tả tập dữ liệu Animal world lấy từ [8]. Tập dữ liệu này gồm 9 đối tượng với 10 thuộc tính: Animal, Hair, Teeth, Eye, Feather, Feet, Eat, Milk, Fly và swim.

Bảng 1. Tập dữ liệu "Animal world"

Animal	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	pointed	forward	N	claw	meat	Y	N	Y
Cheetah	Y	pointed	forward	N	claw	meat	Y	N	Y
Giraffe	Y	blunt	side	N	hoof	grass	Y	N	N
Zebra	Y	blunt	side	N	hoof	grass	Y	N	N
Ostrich	N	N	side	Y	claw	grain	N	N	N
Penguin	N	N	side	Y	web	fish	N	N	Y
Albatross	N	N	side	Y	craw	grain	N	Y	Y
Eagle	N	N	forward	Y	craw	meat	N	Y	N
Viper	N	pointed	forward	N	N	meat	N	N	N

Trước tiên, ta làm việc với thuộc tính Hair. Phân hoạch của U sinh bởi Hair là

$$\pi_{\text{Hair}} = \{ \{ \text{Tiger, Cheetah, Giraffe, Zebra} \}, \{ \text{Ostrich, Penguin, Albatross, Eagle, Viper} \} \}.$$

Ta có

$$E(\text{Hair}) = -\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} = 0,991.$$

Phân hoạch của U sinh bởi Teeth là

$$\pi_{\text{Teeth}} = \{ \{ \text{Tiger, Cheetah, Viper} \}, \{ \text{Giraffe, Zebra} \}, \{ \text{Ostrich, Penguin, Albatross, Eagle} \} \}.$$

$$E(\text{Teeth}) = -\frac{3}{9} \log_2 \frac{3}{9} - \frac{2}{9} \log_2 \frac{2}{9} - \frac{4}{9} \log_2 \frac{4}{9} = 1,53.$$

Phân hoạch của U sinh bởi $\{\text{Hair}, \text{Teeth}\}$ là

$$\pi_{(\text{Hair}, \text{Teeth})} = \{\{\text{Tiger}, \text{Cheetah}\}, \{\text{Giraffe}, \text{Zebra}\}, \{\text{Ostrich}, \text{Penguin}, \text{Albatross}, \text{Eagle}\}, \{\text{Viper}\}\}$$

$$E(\text{Hair}, \text{Teeth}) = -\frac{2}{9} \log_2 \frac{2}{9} - \frac{2}{9} \log_2 \frac{2}{9} - \frac{4}{9} \log_2 \frac{4}{9} - \frac{1}{9} \log_2 \frac{1}{9} = 1,837.$$

Áp dụng công thức (9), ta có thông tin tương hỗ giữa Hair và Teeth như sau

$$I(\text{Hair}; \text{Teeth}) = E(\text{Hair}) + E(\text{Teeth}) - E(\text{Hair}, \text{Teeth}) = 0,991 + 1,53 - 1,837 = 0,684.$$

Áp dụng công thức (10), ta có độ không chắc chắn đối xứng giữa Hair và Teeth như sau

$$SU(\text{Hair}, \text{Teeth}) = 2 \frac{I(\text{Hair}; \text{Teeth})}{E(\text{Hair}) + E(\text{Teeth})} = 2 \frac{0,684}{0,991 + 1,53} = 0,543.$$

Bằng cách tương tự, ta có thể thu được độ không chắc chắn đối xứng giữa Hair và các thuộc tính khác:

$$SU(\text{Hair}, \text{Eye}) = 0,007, SU(\text{Hair}, \text{Feather}) = 0,595, SU(\text{Hair}, \text{Feet}) = 0,341,$$

$$SU(\text{Hair}, \text{Eat}) = 0,387, SU(\text{Hair}, \text{Milk}) = 1, SU(\text{Hair}, \text{Fly}) = 0,256, SU(\text{Hair}, \text{Swim}) = 0,007.$$

Độ không chắc chắn đối xứng trung bình MSU của thuộc tính Hair được tính theo công thức (22) như sau

$$MSU(\text{Hair}) = \frac{0,543 + 0,007 + 0,595 + 0,341 + 0,387 + 1 + 0,256 + 0,007}{8} = 0,392.$$

Bảng quá trình tương tự như đối với Hair, ta có thể tính toán với các thuộc tính khác. Giá trị MSU của tất cả các thuộc tính được trình bày trong Bảng 2.

Bảng 2. Độ không chắc chắn đối xứng và độ không chắc chắn trung bình giữa các thuộc trong Bảng 1

Thuộc tính	SU									MSU
	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim	
Hair	1	0,543	0,007	0,595	0,341	0,387	1	0,257	0,007	0,392
Teeth	0,543	1	0,5	0,786	0,622	0,695	0,543	0,279	0,191	0,520
Eye	0,007	0,5	1	0,092	0,341	0,701	0,007	0,003	0,007	0,207
Feather	0,595	0,786	0,092	1	0,341	0,446	0,595	0,364	0,007	0,403
Feet	0,341	0,622	0,341	0,341	1	0,743	0,341	0,186	0,341	0,407
Eat	0,387	0,695	0,701	0,446	0,743	1	0,387	0,14	0,23	0,466
Milk	1	0,543	0,007	0,595	0,341	0,387	1	0,257	0,007	0,392
Fly	0,257	0,279	0,003	0,364	0,186	0,14	0,257	1	0,003	0,186
Swim	0,007	0,191	0,007	0,007	0,341	0,23	0,007	0,003	1	0,099

Từ Bảng 2, có thể thấy thuộc tính Teeth có giá trị MSU lớn nhất, do đó Teeth được chọn làm thuộc tính gom cụm bởi thuật toán MAX-MEAN-SU.

V. ĐÁNH GIÁ THỰC NGHIỆM

A. Tiêu chuẩn đánh giá

Các kỹ thuật TR, MMR, MDA và MAX-MEAN-SU sử dụng các phương pháp khác nhau trong việc lựa chọn thuộc tính gom cụm. Để đo đạc được một cách chính xác chất lượng gom cụm của các thuộc tính được chọn bởi các kỹ thuật khác nhau là công việc không đơn giản. Vì mục tiêu của phân tích gom cụm là nhóm các đối tượng có cùng các đặc trưng, chúng tôi sử dụng *độ tương tự trung bình trong mỗi cụm* (average intra-class similarity) để đánh giá chất lượng gom cụm của các thuộc tính được chọn.

Định nghĩa 13. Cho hệ thống tin $S = (U, A, V, f)$ và giả sử tất cả các thuộc tính trong A đều là những thuộc tính phạm trù. Khi đó độ tương tự của hai đối tượng x_i và x_j trong U được định nghĩa như sau:

$$s(x_i, x_j) = \frac{\left| \{a_k \in A \mid f(x_i, a_k) = f(x_j, a_k)\} \right|}{|A|} \quad (25)$$

Định nghĩa 14. Cho hệ thống tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Nếu $|X_i| > 1$ thì độ tương tự trung bình (AS) của đối tượng x_{ij} với các đối tượng khác trong X_i được định nghĩa như sau:

$$AS(x_{ij}) = \frac{\sum_{k=1, k \neq j}^{|X_i|} s(x_{ij}, x_{ik})}{|X_i| - 1} \quad (26)$$

Định nghĩa 15. Cho hệ thống tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Khi đó, độ tương tự bên trong (CS) của cụm X_i được định nghĩa bởi:

$$CS(X_i) = \begin{cases} 1 & \text{nếu } |X_i| = 1, \\ \frac{\sum_{j=1}^{|X_i|} AS(x_{ij})}{|X_i|} & \text{trong các trường hợp khác} \end{cases} \quad (27)$$

Định nghĩa 16. Cho hệ thống tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm tạo ra bởi a_j được định nghĩa bởi:

$$ACS(a_j) = \frac{\sum_{i=1}^m CS(X_i)}{m} \quad (28)$$

Độ tương tự trung bình bên trong cụm (ACS) càng cao thì chất lượng gom cụm của thuộc tính gom cụm được chọn càng lớn.

B. Dữ liệu và kết quả tính toán đánh giá

Để đánh giá và so sánh MAX-MEAN-SU với các kỹ thuật TR, MMR, MDA, chúng tôi đã sử dụng bốn tập dữ liệu:

Animal world lấy từ tài liệu [8], được trình bày ở Bảng 1.

Zoo lấy từ UCI database, có 101 đối tượng, trình bày thông tin về các loài thú thông qua 18 thuộc tính.

Soybean lấy từ UCI database, có 47 đối tượng, 35 thuộc tính và một thuộc tính phân lớp.

Statlog (Heart) lấy từ UCI database, có 270 đối tượng, 13 thuộc tính và một thuộc tính phân lớp.

Chúng tôi đã lập trình cả bốn kỹ thuật TR, MMR, MDA và MAX-MEAN-SU bằng ngôn ngữ R với sự hỗ trợ của gói chương trình "RoughSets".

Bảng 3 thống kê các thuộc tính gom cụm được lựa chọn bởi các kỹ thuật trong mỗi tập dữ liệu.

Bảng 3. Các thuộc tính gom cụm được chọn bởi các kỹ thuật trong mỗi tập dữ liệu

	Datasets			
	Animal world	Zoo	Soybean	Statlog (Heart)
TR	Hair	Feathers	Plant-growth	Serum cholestoral
MMR	Hair	Feathers	Plant-growth	Serum cholestoral
MDA	Hair	Feathers	Plant-growth	Serum cholestoral
MAX-MEAN-SU	Teeth	Type	Class	Class

Từ Bảng 3, có thể thấy trong cả bốn tập dữ liệu xem xét, ba thuật kỹ thuật TR, MMR and MDA đều chọn cùng một thuộc tính. Kỹ thuật MAX-MEAN-SU của chúng tôi chọn thuộc tính khác trong cả bốn tập dữ liệu.

Bây giờ, chúng ta hãy đánh giá chất lượng gom cụm của các thuộc tính khác nhau được chọn bởi TR, MMR, MDA và MAX-MEAN-SU trong bốn tập dữ liệu, thông qua tính toán độ tương tự trung bình bên trong cụm (ACS).

Để minh họa công việc tính toán độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm tạo ra bởi một thuộc tính, chúng ta lấy thuộc tính Hair trong tập dữ liệu "Animal" làm ví dụ.

Phân hoạch của tập động vật U sinh bởi thuộc tính Hair gồm hai lớp tương đương:

$$X_1 = X(\text{Hair} = Y) = \{\text{Tiger, Cheetah, Giraffe, Zebra}\},$$

$$X_2 = X(\text{Hair} = N) = \{\text{Ostrich, Penguin, Albatross, Eagle, Viper}\}.$$

Áp dụng công thức (25), ta có

$$s(\text{Tiger, Cheetah}) = 1, \quad s(\text{Tiger, Giraffe}) = 0.444, \quad s(\text{Tiger, Zebra}) = 0.444.$$

Áp dụng công thức (26), ta có độ tương tự trung bình của Tiger với các động vật khác trong X_1 :

$$AS(\text{Tiger}) = \frac{1 + 0.444 + 0.444}{3} = 0.630$$

Bảng thực hiện quá trình tính toán tương tự, ta thu được độ tương tự trung bình của các động vật khác trong X_1 . Các kết quả tính toán cho trong Bảng 4.

Bảng 4. Độ tương tự, AS và CS của các động vật trong X_1 sinh ra bởi Hair

Animal	Tiger	Cheetah	Giraffe	Zebra	AS	CS
Tiger	-	1.000	0.444	0.444	0.630	0.630
Cheetah	1.000	-	0.444	0.444	0.630	
Giraffe	0.444	0.444	-	1.000	0.630	
Zebra	0.444	0.444	1.000	-	0.630	

Áp dụng công thức (27), độ tương tự bên trong của X_1 được xác định như sau.

$$CS(X_1) = \frac{0.630 + 0.630 + 0.630 + 0.630}{4} = 0.630$$

Bằng cách tương tự, ta thu được $CS(X_2) = 0.544$.

Cuối cùng, sử dụng công thức (28), ta thu được độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm sinh bởi Hair là:

$$ACS(\text{Hair}) = \frac{0.630 + 0.544}{2} = 0.587.$$

Bằng cách tương tự, chúng tôi đã tính toán được độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm sinh bởi Teeth trong "Animal world", bởi Feathers và bởi Type trong Zoo, bởi Plant-growth và bởi Class trong Soybean, bởi Serum cholestoral và bởi Class trong Statlog (Heart). Kết quả tính toán thu được cho trong Bảng 5.

Bảng 5. Độ tương tự trung bình bên trong cụm (ACS) của các thuộc tính

	Thuộc tính được chọn và giá trị ACS của nó			
	trên Animal world	trên Zoo	trên Soybean	trên Statlog (Heart)
TR	Hair 0.587	Feathers 0.740	Plant-growth 0.681	Serum cholestoral 0.553
MMR	Hair 0.587	Feathers 0.740	Plant-growth 0.681	Serum cholestoral 0.553
MDA	Hair 0.587	Feathers 0.740	Plant-growth 0.681	Serum cholestoral 0.553
MAX-MEAN-SU	Teeth 0.784	Type 0.866	Class 0.853	Class 0.606

Các kết quả tính toán trong Bảng 5 cho thấy chất lượng gom cụm của các thuộc tính được chọn bởi MAX-MEAN-SU là tốt hơn chất lượng gom cụm của các thuộc tính được chọn bởi TR, MMR and MDA techniques.

VI. KẾT LUẬN

Gần đây, một số công trình áp dụng lý thuyết tập thô vào việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp đã được đề xuất bởi một số tác giả. Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), một thuật toán mới cho việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng Độ không chắc chắn đối xứng (Symmetric Uncertainty - SU), một độ đo của Lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính. Thuộc tính gom cụm được chọn là thuộc tính có độ tương quan trung bình với các thuộc tính khác lớn nhất. Ưu điểm của SU so với Độ lợi thông tin (Information Gain - IG), là SU không thiên vị các thuộc tính có nhiều giá trị. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp.

TÀI LIỆU THAM KHẢO

- [1] Barbara, D., Li, Y., Couto, J.: COOLCAT: an entropy-based algorithm for categorical clustering. In: *Proc. of CIKM 2002*, 582-589, 2002.

- [2] Cao F. Y., Liang J. Y., Li D. Y., Bai L., A new initialization method for categorical data clustering. *Expert Syst. Appl.* 36, 10223-10228, 2009.
- [3] Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS - clustering categorical data using summaries. In: *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 73-83, 1999.
- [4] Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. *Information Systems*, 25(5), 345-366, 2000.
- [5] Huang, Z.: Extensions to the k-averages algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283-304, 1998.
- [6] Han J., and Kamber M., *Data Mining: Concepts and Techniques*, 2nd Morgan Kaufmann Publishers, 2006.
- [7] Hassanein W. A., clustering algorithms for categorical data using concepts of significance and dependence of attributes. *European Scientific Journal*, Vol. 10, No 3, 381-400, 2014.
- [8] Herawan, T., Deris, M. M., Abawajy, J. H.: A rough set approach for selecting clustering attribute. *Knowledge-Based Systems* 23, 220-231, 2010.
- [9] Jain A. K., Data clustering: 50 years beyond k-averages, *Pattern Recogn. Lett.*, 31(8), 651-666, 2010.
- [10] Jyoti Dr., Clustering categorical data using rough sets: a review. *International Journal of Advanced Research in IT and Engineering*, Vol. 2, No. 12, December, 30-37, 2013.
- [11] Khandelwal G., and Sharma R., "A Simple Yet Fast Clustering Approach for Categorical Data", *International Journal of Computer Applications* (0975 - 8887), Volume 120 - No.17, 25-30, June 2015.
- [12] Mazlack, L. J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing clustering attributes. In: *Proceedings of the ISCA 13th International Conference (CAINE 2000)*, 1-6, 2000.
- [13] Parmar, D., Wu, T., Blackhurst, J., MMR: an algorithm for clustering categorical data using rough set theory. *Data and Knowledge Engineering*, 63, 879-893, 2007.
- [14] Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science*, 11, 341-356, 1982.
- [15] Suchita S. Mesakar, M. S. Chaudhari, Review Paper On Data Clustering Of Categorical Data. *International Journal of Engineering Research & Technology*, Vol. 1 Issue 10, December, 2012.
- [16] Yu L., Liu H.: Feature Selection for High-Dimensional Data: A Fast Correlation Based Filter Solution. *ICML* 856-863, 2003.

AN APPROACH FOR SELECTING CLUSTERING ATTRIBUTE USING INFORMATION THEORY

Pham Cong Xuyen, Nguyen Thanh Tung

ABSTRACT: Clustering problem appears in many different areas. The basic objective of clustering is to group objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. Recently, many researchers have contributed to categorical data clustering, where data objects are made up of categorical attributes. In particular, hierarchical clustering categorical data using the rough set theory has attracted much attention. The key to these approaches is how to select one attribute that is the best to cluster the objects at each time from many candidates of attributes.

In this paper, we review three rough set based techniques: TR (Total Roughness), MMR (Min-Min Roughness) and MDA (Maximum Dependency Attribute), and propose MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), an alternative algorithm for hierarchical clustering attribute selection. MAX-MEAN-SU uses SU (Symmetric Uncertainty), a measure of information theory that allows to quantify the degree of mutual correlation between two attributes, to determine the clustering attribute so that its average correlation with other attributes reaches its maximum value. To evaluate and compare MAX-MEAN-SU with three rough set based techniques, we use the concept of average intra-class similarity to measure the clustering quality of each attribute which is chosen by each method. The experiment results show that the clustering quality of the attribute selected by our method is higher than that of attributes selected by TR, MMR and MDA methods. Therefore, MAX-MEAN-SU can be used as an effective technique to select attributes in hierarchical clustering of categorical data.

Keywords: Clustering, Categorical Data, Hierarchical Clustering, Rough Set Theory, Clustering Attribute-Selection, Symmetric Uncertainty.

**Proceedings of the 12th National Conference on
Fundamental and Applied Information
Technology Research (FAIR'2019)**

ISBN: 978-604-913-915-4



9 786049 139154