LIÊN HIỆP CÁC HỘI KHOA HỢC VÀ KỸ THUẬT VIỆT NAM TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT ĐẠI HỌC ĐÀ NẰNG

FAIR

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

NGHIÊN CƯU CƠ BẢN VÀ ƯNG DỤNG CÔNG NGHỆ THÔNG TIN

Proceedings of the 16th National Conference on Fundamental and Applied Information Technology Research

(FAIR'2023)

ĐÀ NẮNG 28-29/9/2023





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

Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin

Proceedings of the 16th National Conference on Fundamental and Applied Information Technology Research (FAIR'2023)

> TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT ĐẠI HỌC ĐÀ NẮNG 28-29/9/2023

NHÀ XUẤT BẢN KHOA HỌC TỰ NHIÊN VÀ CÔNG NGHỆ

BAN BIÊN TẬP

- GS.TS. Vũ Đức Thi Đại học Quốc gia Hà Nội
- PGS.TS. Trần Văn Lăng Viện Hàn lâm Khoa học và Công nghệ Việt Nam
- GS.TS. Từ Minh Phương Học viện Công nghệ Bưu chính Viễn thông
- PGS.TS. Lê Mạnh Thạnh Đại học Huế
- PGS.TS. Trần Đình Khang Đại học Bách khoa Hà Nội
- PGS.TS. Võ Trung Hùng Đại học Đà Nẵng
- TS. Lê Quang Minh Đại học Quốc gia Hà Nội
- ThS. Phan Thị Quế Anh Nhà xuất bản Khoa học tự nhiên và Công nghệ

LỜI NÓI ĐẦU

Hội nghị Khoa học công nghệ quốc gia về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (gọi tắt là FAIR) lần thứ I được tổ chức vào tháng 10 năm 2003 tại Khoa Công nghệ - Đại học Quốc gia Hà Nội (tiền thân của Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội ngày nay). Hội nghị tổ chức nhằm góp phần thúc đẩy nghiên cứu cơ bản và ứng dụng về Công nghệ thông tin tại Việt Nam dưới sự chủ trì của Liên hiệp các Hội Khoa học và Kỹ thuật Việt Nam, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Năm 2023 Ban chỉ đạo Hội nghị đã phối hợp cùng Trường Đại học Sư phạm Kỹ thuật, Đại học Đà Nẵng và các cơ quan khoa học, các nhà khoa học từ các viện nghiên cứu, các trường đại học đã tổ chức Hội nghị Khoa học quốc gia lần thứ XVI (FAIR'2023) với chủ đề "Khoa học dữ liệu trong chuyển đổi số" vào hai ngày 28 - 29/9/2023 tại Đà Nẵng.

FAIR'2023 lần này đã tập hợp khá đông đủ cộng đồng nghiên cứu cơ bản và nghiên cứu ứng dụng về CNTT từ mọi miền đất nước. Mặc dù có rất nhiều hội nghị chuyên ngành diễn ra trong cả nước, nhưng FAIR vẫn là một trong những hội nghị quốc gia được cộng đồng nghiên cứu khoa học trong nước quan tâm.

Hội nghị đã nhận được 154 báo cáo khoa học đăng ký tham dự về tất cả các vấn đề thời sự của Công nghệ thông tin và truyền thông. Ban Chương trình đã tiến hành công việc phản biện và xét duyệt chặt chẽ và chấp nhận 96 bài được trình bày tại Hội nghị, trong số này có 95 bài được lựa chọn đăng trong Kỷ yếu. Các bài báo cáo được trình bày trong 6 tiểu ban diễn ra song song:

- Trí tuệ nhân tạo và ứng dụng
- Khoa hoc dữ liêu
- Hệ thống thông tin
- Xử lý ảnh và Thị giác máy tính
- Xử lý ngôn ngữ tư nhiên
- Mang và Kỹ thuật máy tính

Thay mặt Ban Tổ chức và Ban Chương trình, chúng tôi xin cảm ơn các tác giả đã gửi bài tham gia Hội nghị, các nhà khoa học đã tham gia phản biện và có ý kiến xác đáng, khách quan về nội dung các bài gửi đăng. Chúng tôi xin bày tỏ lòng biết ơn sâu sắc tới Liên hiệp các Hội Khoa học và Kỹ thuật Việt Nam, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 3 đơn vị bảo trợ về mặt chuyên môn cho Hội nghị FAIR và Trường Đại học Sư phạm Kỹ thuật, Đại học Đà Nẵng - đơn vị đăng cai Hội nghị FAIR'2023 đã dành nhiều công sức và thời gian tổ chức Hội nghị này. Đồng thời cũng xin được cảm ơn các đơn vị tài trợ đã giúp đỡ nhiều mặt và tài trợ kinh phí góp phần làm cho Hội nghị FAIR'2023 thành công tốt đẹp. Cuối cùng, chúng tôi xin đặc biệt cảm ơn Nhà xuất bản Khoa học tự nhiên và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã hỗ trợ và giúp đỡ xuất bản cuốn Kỷ yếu này.

BAN BIÊN TÂP

-		

BAN CHỈ ĐAO

Trưởng ban:

GS.VS. Đặng Vũ Minh Liên hiệp các hội KHKT VN

Đồng trưởng ban:

TSKH. Phan Xuân Dũng Liên hiệp các hội KHKT VN

Thành viên:

GS.TS. Đăng Quang Á Viên HLKH&CN VN PGS.TS. Phan Cao Tho Trường ĐHSPKT, ĐH ĐN PGS.TS. Trần Văn Lăng Viên HLKH&CN VN GS.TSKH. Pham Thế Long Hoc viên KTOS GS.TSKH. Nguyễn Khoa Sơn Viện HLKH&CN VN

GS.TS. Vũ Đức Thi ĐHOG-HN GS.TS. Nguyễn Thanh Thủy ĐHQG-HN

BAN TỔ CHỨC

Trưởng ban:

GS.TS. Vũ Đức Thi **ĐHQG-HN**

Đồng trưởng ban:

PGS.TS. Phan Cao Tho Đại học Đà Nẵng

Phó trưởng ban:

GS.TS. Đặng Quang Á Viên HL KH&CN VN

Thành viên:

PGS.TS. Trần Đình Khang Đai học BK Hà Nôi PGS.TS. Trần Văn Lăng Viên HL KH&CN VN TS. Lê Quang Minh Viên CNTT, ĐHOG-HN GS.TS. Từ Minh Phương Hoc viên CNBCVT PGS.TS. Lê Manh Thanh Đại học Huế

Viện CNTT, ĐHQG-HN PGS.TS. Trần Xuân Tú

BAN TỔ CHỨC ĐỊA PHƯƠNG

BAN THƯ KÝ ĐỊA PHƯƠNG

Trưởng ban: Trưởng ban:

PGS.TS. Phan Cao Tho TS. Nguyễn Thị Hải Vân

Phó trưởng ban:

Phó trưởng ban: TS. Đoàn Lê Anh PGS.TS. Võ Trung Hùng

Thành viên: Thành viên:

PGS.TS. Phan Quý Trà TS. Hoàng Thị Mỹ Lệ TS. Nguyễn Thị Hải Vân TS. Pham Tuấn TS. Trần Hoàng Vũ TS. Nguyễn Tấn Thuận TS. Hoàng Thị Mỹ Lệ ThS. Nguyễn Thị Thanh Nga

ThS. Lê Vũ

ThS. Ngô Tấn Thống ThS. Phạm Minh Tuấn

BAN KỸ THUẬT VÀ XUẤT BẢN

Trưởng ban: TS. Lê Quang Minh

Thành viên:

ThS. Phan Thị Quế Anh CN. Lê Thi Thiên Hương TS. Trang Hồng Sơn ThS. Phan Manh Thường

BAN CHƯƠNG TRÌNH

Trưởng ban:

PGS.TS. Trần Văn Lăng Viện HL KH&CN VN

Phó trưởng ban:

PGS.TS. Trần Đình Khang Đại học BK Hà Nội GS.TS. Từ Minh Phương Học viện CNBCVT PGS.TS. Lê Manh Thanh Đai học Huế

PGS.TS. Võ Trung Hùng Trường ĐHSPKT, ĐHĐN

Thành viên:

GS.TS. Đặng Quang Á Viện HLKH&CN VN PGS.TS. Nguyễn Việt Anh Viện CNTT

PGS.TS. Phạm Thế Anh
Trường ĐH Hồng Đức
PGS.TS. Ngô Xuân Bách
Học viện CNBCVT

PGS.TS. Nguyễn Thanh Bình Trường ĐH CNTT-TT Việt Hàn TS. Phan Anh Cang Trường ĐHSPKT Vĩnh Long

PGS.TS. Phạm Văn Cường

TS. Nguyễn Ngọc Cương

PGS.TS. Lê Bá Dũng

PGS.TS. Lương Thế Dũng

Học viện CNTT

Học viện KT Mật mã

PGS.TS. Cao Tuấn Dũng ĐHBK Hà Nội TS. Vũ Thị Đào Học viện KT Mật mã

PGS.TS. Đinh Điền Trường ĐHKHTN, ĐHQG-HCM

PGS.TS. Trần Cao Đê

Trường ĐH Cần Thơ

PGS.TS. Tran Cao Đẹ

PGS.TS. Đặng Văn Đức

Viện CNTT

PGS.TS. Trần Quang Đức

TS. Nguyễn Huy Đức

PGS.TS. Nguyễn Long Giang

Trường ĐH Can Thơ

Viện CNTT

Viện CNTT

TS. Nguyễn Hoàng Hà Trường ĐHKH, ĐH Huế PGS.TS. Trần Mạnh Hà Trường ĐH HUFLIT

TS. Nguyễn Công Hào

Trung tâm CNTT, ĐH Huế

PGS.TS. Nguyễn Mậu Hân

Trường ĐHKH, ĐH Huế

TS. Nguyễn Hữu Hoà

Trường ĐH Cần Thơ

PGS.TS. Huỳnh Xuân Hiệp
Trường ĐH Cần Thơ
TS. Đặng Thị Thu Hiền
Trường ĐH Thủy lợi
TS. Lâm Thành Hiển
Trường ĐH Lạc Hồng
TS. Võ Đình Hiếu
Trường ĐH Công nghệ
PGS.TS. Huỳnh Trung Hiếu
Trường ĐHCN TP. HCM

PGS.TS. Huynn Trung Hieu

PGS.TS. Nguyễn Ngọc Hóa

Trường ĐH Công nghệ

PGS.TS. Trần Văn Hoài

Trường ĐHBK TP. HCM

TS. Huỳnh Hữu Hưng

Trường ĐHBK, ĐHĐN

TS. Lê Minh Hưng Trường ĐH CNTT

TS. Nguyễn Thị Minh Huyền Trường ĐHKHTN, ĐHQG-HN PGS.TS. Phạm Nguyên Khang Trường ĐH Cần Thơ PGS.TS. Nguyễn Tấn Khôi Trường ĐHBK, ĐHĐN PGS.TS. Bùi Thu Lâm Học viện KT Mật mã TS. Vũ Như Lân Trường ĐH Thăng Long TS. Nguyễn Đình Lầu Trường ĐHSP, ĐHĐN TS. Hoàng Thị Mỹ Lệ Trường ĐHSPKT, ĐHĐN PGS.TS. Ngô Thành Long Học viện KT Quân sự

PGS.TS. Vũ Đức Lung Trường ĐH CNTT TS. Lê Quang Minh Viện CNTT, ĐHQG-HN TS. Hoàng Trọng Minh

PGS.TS. Nguyễn Hiếu Minh

PGS.TS. Nguyễn Hà Nam

PGS.TS. Nguyễn Hà Nam

PGS.TS. Nguyễn Thái Nghe

PGS.TS. Đỗ Thanh Nghị

PGS.TS. Phùng Trung Nghĩa

Trường ĐH Cần Thơ

Trường ĐH Cần Thơ

Trường ĐH CNTT&TT

TS. Trần Đức Nghĩa

Viện CNTT

PGS.TS. Lý Quốc Ngọc Trường ĐHKHTN, ĐHQG-HCM PGS.TS. Võ Viết Minh Nhât Trường ĐHKH, ĐH Huế

PGS.TS. Huỳnh Công Pháp Trường ĐH CNTT-TT Việt Hàn

GS.TS. Đỗ Phúc Trường ĐH CNTT
TS. Đặng Hoài Phương Trường ĐHBK, ĐHĐN
PGS.TS. Lê Hồng Phương Trường ĐHKHTN, ĐHQG-HN
PGS.TS. Habra Outra

PGS.TS. Hoàng Quang
Trường ĐHKH, ĐH Huế
PGS.TS. Nguyễn Hữu Quỳnh
PGS.TS. Lê Hoàng Sơn
Viện CNTT, ĐHQG-HN
PGS.TS. Nguyễn Thái Sơn
Trường ĐH Trà Vinh

PGS.TS. Ngô Quốc Tạo Viện CNTT

TS. Nguyễn Văn Tảo Trường ĐH CNTT&TT

TS. Trần Minh Tân Bộ TTTT

TS. Vũ Đức Thái Trường ĐH CNTT&TT
TS. Ngô Đức Thành Trường ĐH CNTT
PGS.TS. Trịnh Đình Thắng Trường ĐHSP 2 Hà Nội

GS.TS. Vũ Đức Thi ĐHQG-HN

PGS.TS. Nguyễn Đình Thuân Trường ĐH CNTT TS. Phạm Thị Thu Thúy Trường ĐH Nha Trang

PGS.TS. Đỗ Năng Toàn Viện CNTT

TS. Nguyễn Anh Tuấn Trường ĐH HUFLIT TS. Phạm Minh Tuấn Trường ĐHBK, ĐHĐN

PGS.TS. Nguyễn Thanh Tùng Trường ĐH CMC

PGS.TS. Trần Xuân Tú Viện CNTT, ĐHQG-HN PGS.TS. Võ Thanh Tú Trường ĐHKH, ĐH Huế PGS.TS. Trương Công Tuấn Trường ĐHKH, ĐH Huế TS. Nguyễn Trần Quốc Vinh Trường ĐHSP, ĐH-ĐN PGS.TS. Lê Sỹ Vinh Trường ĐH Công nghê

PGS.TS. Vũ Việt Vũ Viện CNTT, ĐHQG-HN

PHỤ TRÁCH CÁC TIỂU BAN

PGS.TS. Đỗ Thanh Nghị Tiểu ban Trí tuệ nhân tạo và ứng dụng

GS.TS. Đỗ Phúc Tiểu ban Khoa học dữ liệu PGS.TS. Trương Công Tuấn Tiểu ban Hệ thống thông tin

PGS.TS. Trương Công Tuấn

PGS.TS. Đỗ Năng Toàn

Tiểu ban Hệ thống thông tin

Tiểu ban Xử lý ảnh và Thị giác máy tính

PGS.TS. Đinh Điền

Tiểu ban Xử lý ngôn ngữ tự nhiên

PGS.TS. Võ Thanh Tú Tiểu ban Mạng và Kỹ thuật máy tính

CÁC ĐƠN VỊ ĐỒNG TỔ CHỨC







LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT VIỆT NAM

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT -ĐẠI HỌC ĐÀ NẪNG

CÁC ĐƠN VỊ TÀI TRỢ CHÍNH





Công ty cổ phần Netnam Corporation Công ty TNHH Phần mềm FPT Miền Trung





Trường Đại học Công nghệ Thông tin và Truyền thông Công ty cổ phần V.B.P.O

MỤC LỤC

1.	PHAT HIỆN VA PHAN VUNG KHỔI U NÀO TRÊN ANH 3D MRI VỚI KỲ THUẬT 3D U-NET Phan Anh Cang, Nguyễn Khắc Tường, Phan Thượng Cang, Trần Văn Lăng	1
2.	VIPRIME: ỨNG DỤNG NHẬN DẠNG NHẠC CỔ TRUYỀN VIỆT NAM VỚI MẠNG HỌC SÂU TÍCH CHẬP	12
	Trần Thanh Huy, Trần Minh Đạt, Huỳnh Gia Khương, Mã Trường Thành, Phạm Nguyên Khang, Đỗ Thanh Nghị	
3.	BRAIN TUMOR SEGMENTATION BASED ON DEEP SUPERVISION AND CONTEXT FEATURE FUSION	20
	Tai Do Nhu, Son Vo Thanh Hoang, Hai Tran Minh, Tram Tran Nguyen Quynh, Huy Nguyen Thanh, Thanh Nguyen Thi Ngoc, Huy Nguyen Quoc, Soo-Hyung Kim	
4.	BREASTHISTOLOGYGEN: MỘT SỐ PHƯƠNG PHÁP SINH ẢNH MÔ UNG THƯ VÚ KHÔNG ĐIỀU KIỆN DỰA VÀO ƯỚC LƯỢNG MẬT ĐỘ TIỀM ẨN VÀ KHÔNG TIỀM ẨN	27
	Trần Đình Toàn, Nguyễn Đức Toàn, Lê Minh Hưng, Hoàng Lê Uyên Thục	
5.	CHUẨN ĐOÁN CÁC BỆNH VỀ DA BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY	37
	Nguyễn Minh Hải, Văn Thế Thành, Trần Văn Lăng	
6.	CHẨN ĐOÁN UNG THƯ VÚ TRÊN NHŨ ẢNH SỬ DỤNG CÁC MÔ HÌNH HỌC SÂU VÀ MẠNG VISION TRANSFORMER	45
	Phan Anh Cang, Võ Đình Nghĩa, Trần Hồ Đạt, Phan Thượng Cang	
7.	DỰ ĐOÁN HOẠT TÍNH CỦA HỢP CHẤT HOÁ HỌC DỰA TRÊN TIẾP CẬN HỌC MÁY	53
	Đỗ Hoàng Tú, Trần Văn Lăng, Phạm Công Xuyên, Lê Mậu Long	
8.	MÔ HÌNH PHÂN LOẠI ẢNH MÔ IHC SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU KẾT HỢP VỚI HÀM WEIGHTED DOUBLE SOFTMAX CROSS-ENTROPY ĐỂ XÁC ĐỊNH GIAI ĐOẠN UNG THƯ VÚ	63
	Trần Đình Toàn, Nguyễn Đức Toàn, Lê Minh Hưng, Hoàng Lê Uyên Thục	
9.	MỘT KỸ THUẬT GIA TĂNG DỮ LIỆU SỬ DỤNG HỌC SÂU CẢI THIỆN CHẨN ĐOÁN SỚM UNG THƯ DA Ở VIỆT NAM	70
	Vũ Văn Hiệu	
10.	NHẬN DẠNG CỬ CHỈ TAY ĐIỀU KHIỂN CHUỘT MÁY TÍNH TỪ XA SỬ DỤNG CẢM BIẾN QUÁN TÍNH VÀ HỌC SÂU	80
	Đặng Huỳnh Khánh Dương, Nguyễn Nho Song Hoàng, Lê Phúc Khương, Lê Nguyễn Ngọc Lâm, Ninh Khánh Duy	
11.	PHÂN TÍCH HIỆU QUẢ HOẠT ĐỘNG CỦA MỘT SỐ KIẾN TRÚC HỌC SÂU CHO BÀI TOÁN NHẬN DẠNG CẢM XÚC DỰA TRÊN TÍN HIỆU ĐIỆN NÃO	88
	Dương Thị Mai Thương, Phùng Trung Nghĩa	
12.	PHÁT HIỆN VÀ PHÂN LOẠI PROTEIN HẠT NHÂN TẾ BÀO UNG THƯ VÚ TRÊN ẢNH MÔ HỌC DÙNG KỸ THUẬT DEEP LEARNING	95
	Phan Anh Cang, Le Thi My Tien, Tran Thai Bao	

13.	SU DỤNG HỆ THONG TRI TUỆ NHAN TẬO PHAN LOẠI TRANG SUC THUY TINH VAN HOÁ ÓC EO VÀ SA HUỲNH	105
	Ngô Hồ Anh Khôi, Bùi Hoàng Bắc, Võ Khương Duy	
14.	ỨNG DỤNG PHƯƠNG PHÁP HỌC LIÊN KẾT VỚI THUẬT TOÁN FEDBN TRONG PHÁT HIỆN UNG THƯ VÚ	113
	Ngô Văn Tuấn Huy, Đào Duy Tuấn, Trần Thị Minh Hạnh	
15.	XÂY DỰNG MỘT HỆ THỐNG THÔNG MINH HỖ TRỢ GIÁM SÁT BẤT THƯỜNG VỚI CÂY ATISO TẠI ĐÀ LẠT	122
	Hải Đặng Thanh, Toàn Trần Tuấn, Thi Mai Hà, Châu Hoàng Thị Minh, Anh Lê Tuấn, Vinh Nguyễn Trần Quốc, Hiệp Nguyễn Minh	
16.	A NEUTROSOPHIC-THREE PARALLEL PATHS CONVOLUTIONAL NEURAL NETWORK CLASSIFICATION MODEL	129
	Roan Thi Ngan, Nguyen Dinh Quy, Le Hoang Son, Vu Anh Tuan, Tran Thi Ngan, Lan Luong Hong, Ta Tuan Anh, Nguyen Thi Van Anh, Hoang The Anh	
17.	APPLY PROBABILISTIC APPROACH WITH UNSUPERVISED LEARNING FOR PATIENT MATCHING IN HEALTH INFORMATION EXCHANGE	137
	Pham Thanh Dat, Nguyen Van Duong, Nguyen Tuan Cuong, Ha Thi Hong Van	
18.	DỰ BÁO CHỈ SỐ CHẤT LƯỢNG KHÔNG KHÍ SỬ DỤNG KẾT HỢP MÔ HÌNH CNN, LSTM VÀ DEEP Q-NETWORK	147
	Trần Duy Phương, Nguyễn Đình Thuân	
19.	DỰ ĐOÁN SỰ BIẾN ĐỔI CỦA ẢNH MÂY VỆ TINH SỬ DỤNG MÔ HÌNH SUY DIỄN MỜ PHỨC KẾT HỢP PHƯƠNG PHÁP MAP REDUCE	154
	Giang Lê Trường, Hoàng Nguyễn Xuân, Lương Nguyễn Văn, Chung Phạm Bá Tuấn, Hoàng Lê Minh, Thông Phạm Huy, Lan Lương Thị Hồng, Thanh Nguyễn Quang, Cành Hoàng Thị	
20.	DỰ ĐOÁN CODE-SMELL DỰA TRÊN PHÂN LOẠI ĐA NHÃN	163
	Nguyễn Thanh Bình, Nguyễn Hữu Nhật Minh, Lê Thị Mỹ Hạnh, Nguyễn Thanh Bình	
21.	NÂNG CAO CHẤT LƯỢNG MẠNG NƠRON MIN-MAX MỜ DỰA TRÊN ĐÁNH GIÁ SIÊU HỘP VÀ ỨNG DỤNG	171
	Đình Minh Vũ, Thanh Sơn Nguyễn, Bá Dũng Lê, Quang Huy Hoàng, Đức Lưu Nguyễn	
22.	NEW DATASETS AND DEEP LEARNING APPROACH FOR MILITARY TRAINING EXERCISES	181
	Phan Hai Hong, Phan Xuan Hoang, Dang Thi Thuong	
23.	TRANSFORMING ENROLLMENT ADVISING IN EDUCATION WITH DEEP LEARNING MODELS AND CHATBOTS	188
	Nguyen Nang Hung Van, Ho Le Minh Nhat, Vo Duc Hoang, Do Phuc Hao	
24.	HỆ THỐNG KHUYẾN NGHỊ TUẦN TỰ DỰA TRÊN SỞ THÍCH NGẮN HẠN VÀ DÀI HẠN CỦA NGƯỜI DÙNG	196
	Nguyễn Thanh Hùng, Nguyễn Trung Hiếu, Nguyễn Tiến Thịnh, Đỗ Hồng Quân, Phùng Thế Huân, Lê Minh Tuấn, Hồ Ngọc Tú	
25.	PHÂN CỤM CÁC ĐỊA ĐIỂM DU LỊCH TẠI VIỆT NAM SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM MÒ	204
	Phùng Thế Huân, Vũ Đức Thái, Đỗ Đình Cường, Nguyễn Duy Minh, Quách Xuân Trưởng, Nguyễn Trần Quốc Vinh	

26.	PHÂN TÍCH Ý KIẾN ĐÁNH GIÁ DỊCH VỤ NGÂN HÀNG VIỆT NAM DỰA TRÊN CÁC KĨ THUẬT HỌC SÂU	210
	Khưu Thuỳ Loan, Vũ Nguyễn Năng Khánh, Dang Ngoc Hoang Thanh	
27.	PHÁT HIỆN GIAN LẬN TRONG THỂ TÍN DỤNG BẰNG CÁCH SỬ DỤNG HỌC MÁY VÀ LOGIC MỜ	216
	Duc Thuan Tran, Dinh Thuan Nguyen	
28.	XÁC ĐỊNH ĐỘ TƯƠNG TỰ CỦA CÁC ẢNH CHỮ KÝ DỰA TRÊN TIẾP CẬN XẾP HẠNG ĐA TẠP	225
	Huy Trần Văn, Hùng Trần Công, Huy Ngô Hoàng, Tuyết Đào Văn, Phạm Thị Kim Dzung, Quyền Nguyễn Văn, Nghiệp Lê Đình	
29.	XÂY DỰNG HỆ THỐNG NHẬN DẠNG DỮ LIỆU ĐA PHƯƠNG THỨC CỦA TÍN HIỆU ĐIỆN NÃO ĐỔ DỰA TRÊN ĐẠI SỐ TENSOR	233
	Dương Thanh Linh, Lương Duy Đức, Nguyễn Trần Quốc Vinh, Nguyễn Thị Ngọc Anh	
30.	MỘT GIẢI PHÁP HỌC SÂU TÍCH HỢP ĐỂ DỰ BÁO LƯỢNG MƯA QUA ẢNH RADAR Sơn Hà Gia, Tuấn Trần Mạnh, Tân Nguyễn Hồng	241
31.	HƯỚNG TIẾP CẬN HIỆU QUẢ CHO VIỆC PHÂN LỚP DỰA TRÊN LUẬT KẾT HỢP	250
32.	A COMBINATION OF FEATURE SELECTION AND DATA SAMPLING TECHNIQUES FOR SOFTWARE FAULT PREDICTION	258
	Thi Minh Phuong Ha, Thanh Long Nguyen, Thanh Binh Nguyen	
33.	DKD-TREE: PHƯƠNG PHÁP LẬP CHỈ MỤC PHÂN TẦN TẬP VECTOR NHIỀU CHIỀU QUI MÔ LỚN	266
	Phan Hồng Trung, Đỗ Phúc	
34.	MỘT SỐ TÍNH CHẤT VỀ BIỂU DIỄN KHÓA QUA HỮU HẠN PHÉP DỊCH CHUYỂN LƯỢC ĐỔ KHỔI	274
	Trịnh Đình Thắng, Trần Minh Tuyến, Trịnh Ngọc Trúc	
35.	MỘT TIẾP CẬN TỐM TẮT NỘI DUNG TRONG VIDEO TIN TỨC	280
	Nguyễn Thanh Hải, Lê Việt Khoa, Đỗ Khánh Toàn, Nguyễn Thái Nghe	
36.	MỘT SỐ PHƯƠNG PHÁP XỬ LÝ TRUY VẤN MỜ HIỆU QUẢ DỰA VÀO ĐỘ ĐO TƯƠNG TỰ	288
	Nguyễn Tấn Thuận, Trần Thị Thúy Trinh, Trương Ngọc Châu, Đoàn Văn Ban	
37.	MÁY TÍNH HỖ TRỢ NGHIÊN CỨU SỰ TỒN TẠI NGHIỆM CỦA MỘT SỐ BÀI TOÁN BIÊN PHI TUYẾN	296
	Đặng Quang Á	
38.	MỘT PHÁT TRIỂN TRONG PHÂN CỤM BÁN GIÁM SÁT MỜ TÍCH CỰC DỰA VÀO VÙNG BIÊN	303
	Dũng Dương Tiến, Nam Hà Hải, Giang Nguyễn Long, Sơn Lê Hoàng, Tuấn Trần Mạnh	
39.	MỘT PHƯƠNG PHÁP HIỆU QUẢ ĐỂ KHAI THÁC CÁC TẬP MỤC TỐI TIỀU CÓ LỢI ÍCH TRUNG BÌNH CAO PHỔ BIẾN	312
	Dương Văn Hải, Trương Chí Tín, Hoàng Minh Tiến, Trần Ngọc Anh	

40.	NGHIÊN CỨU XÂY DỰNG CẦU TRÚC TÔPÔ THEO TIẾP CẬN TẬP THÔ, ỨNG DỤNG CHO BÀI TOÁN RÚT GỌN THUỘC TÍNH	321
	Trần Thanh Đại, Nguyễn Long Giang, Vũ Đức Thi, Đinh Thu Khánh, Triệu Thu Hương, Trần Thị Huệ, Trịnh Văn Hà, Kiều Tuấn Dũng, Cù Kim Long	
41.	HỆ THỐNG AIOT QUAN TRẮC VÀ DỰ BÁO CHẤT LƯỢNG KHÔNG KHÍ THEO THỜI GIAN THỰC	329
	Lê Nguyễn Bình Nguyên, Đặng Nguyễn Nam Anh, Nguyễn Đức Đăng Khôi, Lê Duy Tân	
42.	HỆ THỐNG NHẬN DẠNG QUẢ MĂNG CỤT CHÍN DỰA TRÊN MÔ HÌNH FASTER R-CNN CẢI TIẾN	337
	Trịnh Trung Hải, Hồ Phan Hiếu, Nguyễn Hà Huy Cường, Ninh Khánh Duy	
43.	PHƯƠNG PHÁP XÂY DỰNG MẠCH LƯỢNG TỬ CHO SBOX CÓ THỂ XÁC ĐỊNH ĐƯỢC MA TRẬN UNITA BIẾN ĐỔI	345
	Nguyễn Văn Nghị, Vũ Minh Thắng, Nguyễn Hải Quân, Đỗ Quang Trung, Nguyễn Văn Duẩn	
44.	PHƯƠNG PHÁP XÂY DỰNG ỨNG DỤNG TRỢ LÝ Y TẾ THÔNG MINH DỰA TRÊN KỸ THUẬT TỰ TẠO HƯỚNG DẪN ĐỂ TINH CHỈNH LLM	354
	Đặng Anh Toàn, Huỳnh Thị Mỹ Trang, Trần Văn Lăng	
45.	PORTFOLIO MAXIMIZATION FOCUSING ON DIVIDEND-PAYING STOCKS AND BANKING INTERESTS	364
	Phuc Tan Huynh, Trang Hong Son, Nguyen Thi Thu Du, Nguyen Van Huy, Nguyen Quang Ky, Khoa Dang Vo, Nguyen Huynh Tuong	
46.	STUDENT SELECTION TO UNIVERSITY ADMISSION: STABLE MATCHING THEORY AND GALE-SHAPLEY ALGORITHM	372
	Trinh Bao Ngoc, Do Thi Phuong Thao, Dinh Thi Minh Nguyet, Bui Quoc Khanh, Nguyen Xuan Thang	
47.	YOLO-FLOW: MÔ HÌNH NHANH VÀ CHÍNH XÁC CHO PHÁT HIỆN RÁC THẢI NHỰA TRÊN SÔNG	381
	Trang Thanh Trí, Huỳnh Ngọc Thái Anh, Phạm Thế Phi, Đỗ Thanh Nghị	
48.	A HYBRID QUANTUM-CLASSICAL NEURAL NETWORK UTILIZING THE LION OPTIMIZER FOR COVID-19 IMAGE CLASSIFICATION	388
	Oanh Cuong Do, Chi Mai Luong, Giang Son Tran	
49.	CHÚ THÍCH ẢNH TỰ ĐỘNG DỰA TRÊN PHÁT HIỆN ĐỐI TƯỢNG VÀ CƠ CHẾ CHÚ Ý	395
	Nguyễn Văn Thịnh, Trần Văn Lăng, Văn Thế Thành	
50.	MỘT KỸ THUẬT SÀNG LỌC TRỂ TỰ KỶ DỰA TRÊN ĐẶC TRƯNG KHUÔN MẶT	405
	Thành Trần Văn, Hiến Lâm Thành, Toàn Đỗ Năng, Tú Huỳnh Tuấn, Truyền Nguyễn Trọng	
51.	NGHIÊN CỨU SỬ DỤNG MÔ HÌNH DETR VỚI BÀI TOÁN PHÁT HIỆN ĐỐI TƯỢNG TRONG KHÔNG ẢNH	413
	Nguyễn Tấn Trần Minh Khang, Ngô Hương Giang, Nguyễn Thị Thanh Trúc	
52.	PERFORMANCE EVALUATION OF MEDIAPIPE AND OPENPOSE FOR SKELETON DATA EXTRACTION	420
	Khac Anh Phu, Van Dung Hoang, Van Tuong Lan Le	
53.	PHÁT HIỆN ĐỐI TƯỢNG NHỎ TRONG ẢNH TỪ TRÊN KHÔNG SỬ DỤNG PHƯƠNG PHÁP ORIENTED REPPOINTS	428
	Nguyễn Tấn Trần Minh Khang, Nguyễn Xuân Quang, Tạ Việt Phương	

54.	PHÁT HIỆN PHƯƠNG TIỆN GIAO THÔNG TRONG KHÔNG ẢNH BẰNG MÔ HÌNH YOLO V5 NHỎ GỌN	435
	Nguyễn Hữu Lợi, Nguyễn Tấn Trần Minh Khang	
55.	PHÁT HIỆN SỎI THẬN SỬ DỤNG KỸ THUẬT VISION TRANSFORMER	442
	Phan Thượng Cang, Nguyễn Ngọc Hoàng Quyên, Phan Anh Cang	
56.	PHÁT TRIỂN MÔ HÌNH TÌM KIẾM ẢNH SỬ DỤNG CẦU TRÚC KD-TREE VÀ TÚI TỪ THỊ GIÁC	451
	Nguyễn Thị Định, Nguyễn Phương Nam, Văn Thế Thành, Lê Mạnh Thạnh	
57.	IMPACT OF COLOR SPACES ON MEDICAL IMAGE FUSION	460
	Quoc Viet Kieu, Vinh Nam Huynh, Giang Son Tran	
58.	HỌC BIỂU DIỄN ẢNH VỚI MẠNG NƠRON TÍCH CHẬP ĐỒ THỊ CHO TRA CỨU ẢNH	468
	Nguyễn Văn Thanh, Nguyễn Hữu Quỳnh, Phạm Huy Hoàng, Đào Thị Thúy Quỳnh, Cù Việt Dũng	
59.	HỌC ĐỘ ĐO KHOẢNG CÁCH GIỮA CÁC BỘ ĐẶC TRƯNG ẢNH DỰA TRÊN THÔNG TIN PHẢN HỒI LIÊN QUAN TRONG TRUY VẤN ẢNH DỰA TRÊN NỘI DUNG	477
	Hoàng Xuân Trung, Trần Công Hùng, Ngô Hoàng Huy, Đào Văn Tuyết,	
	Đoàn Văn Hòa, Trần Văn Huy, Ngô Nguyên Khôi	
60.	TỔNG HỢP ĐẶC TRƯNG THÔNG TIN VỀ NỘI DUNG VÀ PHONG CÁCH CHO BÀI TOÁN CHÓNG GIẢ MẠO KHUÔN MẶT	485
	Bùi Quốc Bảo, Trần Anh Đạt, Nguyễn Duy Quang, Nguyễn Trọng Khánh	
61.	TRUY VẤN HÌNH ẢNH SỬ DỤNG PHƯƠNG PHÁP KẾT HỢP VISUALRANK VÀ VISION TRANSFORMER	492
	Phan Anh Cang, Đỗ Thị Ngọc Hiền, Trần Hồ Đạt	
62.	KẾT HỢP K-MEANS VỚI RS-TREE CHO BÀI TOÁN TÌM KIẾM ẢNH THEO NGỮ NGHĨA	499
	Lê Thị Vĩnh Thanh, Lê Mạnh Thạnh, Văn Thế Thành, Nguyễn Thị Uyên Nhi	
63.	FALL AND CHEST-HOLDING POSTURE DETECTION FROM SURVEILLANCE CAMERA WITH YOLOV7	511
	Nguyen Tung Lam, Tran Giang Son	
64.	AN IMPROVEMENT IN MEASURING THE SIMILARITY OF VIETNAMESE DOCUMENTS Pham Thi Thu Thuy	519
65.	CAAS: Hỗ TRỢ TƯ VẤN TUYỂN SINH VỚI CHAT-VOICE SỬ DỤNG CHATGPT VÀ MÁY HỌC	526
00.	Mã Trường Thành, Châu Thế Khanh, Nguyễn Thiên Phúc, Huỳnh Gia Khương, Nguyễn Tí Hon, Trần Việt Châu, Đỗ Thanh Nghị	320
66.	DỊCH NGHĨA TỰ ĐỘNG THƠ VĂN CHỮ HÁN CỦA VIỆT NAM SANG TIẾNG VIỆT ĐƯƠNG ĐẠI SỬ DỤNG DỊCH MÁY THỐNG KÊ	534
	Thái Hoàng Lâm, Đinh Điền	
67.	APPLICATION OF LORA IN DATA TRANSFER IN AGRICULTURE Tran Vinh Phuc	541
68.	NGHIÊN CÚU ĐỀ XUẤT HỆ THỐNG GỢI Ý SOẠN THẢO VĂN BẢN HÀNH CHÍNH	550
	Phùng Thế Huân, Lê Minh Tuấn, Hoàng Thị Cành, Phạm Huy Thông, Nguyễn Thị Hồng Hạnh, Đỗ Hồng Quân, Đỗ Huy Khôi, Nguyễn Vạn Nhã	

70.	NÔM TEXT RECOGNITION USING AN END-TO-END TRANSFORMER-BASED ARCHITECTURE WITH PRE-TRAINED MODELS	566
	Nguyen Xuan Quang, Nguyen Quang Tan, Le Thi Thuy Hang, Dinh Dien	
71.	HỌC ÍT MẪU VỚI SỰ HỖ TRỢ CỦA TỪ ĐIỂN CHO NHẬN DIỆN NÔNG SẢN	573
	Trần Anh Đạt, Đào Hoàng Ngân, Bùi Quốc Bảo, Nguyễn Văn Nam, Nguyễn Thị Phương Thảo, Nguyễn Hữu Quỳnh	
72.	MỘT CẢI TIẾN KHẢ NĂNG THÍCH ỨNG CHO CHATBOT ĐA NGÔN NGỮ	580
	Ngô Văn Sơn, Thái Thị Phương, Võ Viết Minh Nhật	
73.	NHẬN DIỆN BẢNG CHỮ CÁI NGÔN NGỮ KÝ HIỆU TIẾNG VIỆT SỬ DỤNG MÔ HÌNH HỌC SÂU	588
	Võ Đức Hoàng	
74.	TRÍCH XUẤT THÔNG TIN THỰC THỂ VÀ QUAN HỆ TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH ĐỒ THỊ ĐỘNG	594
	Phạm Lương Hào, Nguyễn Thanh Phước, Phạm Công Thiện,	
	Tô Thành Nhân, Quản Thành Thơ	
75.	XẤP XỈ ĐA THỨC TRONG TÁI TẠO ĐƯỜNG TẦN SỐ CƠ BẢN F0 CHO CÁC TỪ GHÉP TIẾNG VIỆT DỰA TRÊN MÔ HÌNH QTA VÀ THUẬT TOÁN TỐI ƯU HÀM MỤC TIÊU	602
	Tạ Yên Thái, Vũ Thị Hải Hà, Đào Văn Tuyết, Ngô Hoàng Huy, Nguyễn Văn Hùng, Đoàn Văn Hòa, Trần Công Hùng	
76.	XÂY DỰNG KHO NGỮ LIỆU ĐA NGỮ NHỜ VÀO UNL	608
	Võ Trung Hùng, Phan Thị Lệ Thuyền, Ninh Khánh Chi	
77.	AN AUTOMATED COMPARISON OF THE STRUCTURE BETWEEN VIETNAMESE WORDNET (VIETNET) AND WORDNET FOR ABSTRACT NOUNS	614
	Phan Thi My Trang, Duong Thi An, Tran Thi Minh Phuong	
78.	VICALLIGRAPHY: TẬP DỮ LIỆU CHO BÀI TOÁN NHẬN DIỆN CHỮ THƯ PHÁP TIẾNG VIỆT VÀ MỘT SỐ ĐÁNH GIÁ	620
	Lê Xuân Tùng, Nguyễn Xuân Trường, Trần Thanh Tùng, Trần Trọng Khiêm, Đỗ Văn Tiến, Ngô Đức Thành	
79.	QUERY2TREE: MÔ HÌNH TRẢ LỜI CÂU HỎI LẬP LUẬN THEO TIÉP CẬN NHÚNG ĐỒ THỊ TRI THỨC LỚN	628
	Phan Hồ Viết Trường, Đỗ Phúc	
80.	THIẾT KỂ VÀ XÂY DỰNG HỆ THỐNG CHỮA CHÁY TỰ ĐỘNG HAI TRỤC – NHẬN DẠNG LỬA BẰNG CAMERA	636
	Võ Duy, Liêu Xuân Hiền, Trần Anh Khoa, Lê Nguyễn Tín Huy, Mai Xuân Phú, Võ Đông Hưng, Lê Thế Thông, Võ Đình Ngọc Uyển, Bùi Sỹ Quân, Nguyễn Phúc Song Toàn, Trần Quang Nguyên, Nguyễn Mạnh Bảo	
81.	GIẢI PHÁP NÂNG CAO KHẢ NĂNG GIẤU TIN TRONG ẢNH NHỊ PHÂN DỰA VÀO KỸ THUẬT DỊCH CHUYỂN LƯU ĐÔ	643

NGHIÊN CỨU MỘT SỐ MÔ HÌNH HỎI ĐÁP TỰ ĐỘNG TRÊN MIỀN DỮ LIỆU DU LỊCH

Trần Thanh Phước, Nguyễn Duy Khanh, Trần Thanh Trâm

Huỳnh Văn Thanh, Võ Phước Hưng, Nguyễn Thái Sơn

558

69.

82.	GIẦU TIN THUẬN NGHỊCH CHO ẢNH NỘI SUY DỰA TRÊN MODULO VÀ DỊCH CHUYỀN LƯU ĐỔ	650
	Trâm Hoàng Nam, Nguyễn Thái Sơn, Hồ Ngọc Huyền	
83.	ỨNG DỤNG LƯỢC ĐỔ ZKP FIAT-SHAMIR VÀO GIAO THỨC TRAO ĐỔI KHÓA DIFFIE-HELLMAN TRÊN TRƯỜNG HỮU HẠN	657
	Nguyễn Văn Nghị, Lê Minh Hiếu, Nguyễn Văn Duẩn, Lê Thị Bích Hằng, Đinh Văn Hùng	
84.	ỨNG DỤNG MẠNG NƠRON HỒI QUY TRONG PHÂN LOẠI TIN GIẢ	665
	Khánh Chi Ninh, Khắc Nghĩa Từ, Trung Hùng Võ, Khánh Duy Ninh	
85.	GIẦU TIN THUẬN NGHỊCH TRONG ẢNH NHỊ PHÂN DỰA TRÊN DỊCH CHUYỂN BIỂU ĐỒ ĐỘ LẬT ĐỐI XỬNG	674
	Dương Ngọc Vân Khanh, Huỳnh Văn Thanh, Nguyễn Thái Sơn	
86.	BUILDING KEY-DEPENDENT XOR TABLES FOR AES BASED ON HADAMARD MATRICES	684
	Truong Minh Phuong, Tran Thi Luong, Hoang Dinh Linh, Nguyen Van Long	
87.	CẢI THIỆN HIỆU NĂNG FANET ỨNG DỤNG TRONG TÌM KIẾM CỨU NẠN: THAM SỐ MÔ HÌNH DI ĐỘNG VÀ CHIẾN LƯỢC TRIỂN KHAI	691
	Mai Cường Thọ, Nguyễn Thị Hương Lý, Nguyễn Quốc Cường, Lê Hữu Bình, Võ Thanh Tú	
88.	ĐÁNH GIÁ HIỆU NĂNG HỆ THỐNG CÂN BẰNG TẢI CHO MẠNG TRUYỀN THỐNG VÀ MẠNG ĐỊNH NGHĨA BẰNG PHẦN MỀM	700
	Vũ Đức Như, Dương Văn Dần, Trần Nam Khánh, Tạ Minh Thanh	
89.	ĐÁNH GIÁ HIỆU QUẢ MỘT SỐ PHƯƠNG PHÁP TỐI ƯU TRONG QUY HOẠCH MẠNG RFID Văn Hòa Lê, Trung Đức Phạm, Văn Tùng Nguyễn	707
90.	KIỂM CHỨNG THÔNG TIN DỰA TRÊN DỮ LIỆU THU THẬP CÓ TRÍCH DẪN TỪ INTERNET Dương Tổ Hương, Hồ Hải Văn, Đỗ Phúc	716
91.	NÂNG CAO HIỆU QUẢ ĐỊNH VỊ TRONG NHÀ SỬ DUNG MẠNG HỌC SÂU KẾT HỢP Vũ Văn Hiệu, Nguyễn Thị Tính, Ngô Văn Bình	725
92.	LẬP QUỸ ĐẠO CHUYỂN ĐỘNG CHO CÁNH TAY ROBOT DẠNG KHUNG XƯƠNG SỬ DỤNG MẠNG NƠRON NHÂN TẠO	733
	Vũ Duy Giang, Phạm Văn Hà	
93.	RLMR: MỘT PHƯƠNG PHÁP ÁP DỤNG Q-LEARNING CHO ĐỊNH TUYẾN TRONG MẠNG TÙY BIẾN DI ĐỘNG	741
	Nguyễn Quốc Cường, Mai Cường Thọ, Lê Hữu Bình, Võ Thanh Tú	
94.	TĂNG CƯỜNG BẢO MẬT CHO CƠ SỞ DỮ LIỆU QUAN HỆ ĐIỆN TOÁN ĐÁM MÂY BẰNG CÔNG NGHỆ BLOCKCHAIN	749
	Từ Quốc Huy, Nguyễn Đình Thuân	
95.	MÔ HÌNH MỚI SỬ DỤNG KỸ THUẬT SO KHỚP ĐỂ PHÁT HIỆN HÀNH VI BẤT THƯỜNG CỦA CON NGƯỜI	757
	Lê Hồng Lam, Lê Tiến Hiếu, Hà Huy Công, Bùi Xuân Vinh, Phạm Thành Công,	
	Nguyễn Đức Nhân, Đinh Văn Châu, Nguyễn Hà Nam	

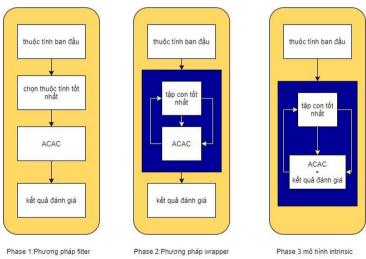
HƯỚNG TIẾP CẬN HIỆU QUẢ CHO VIỆC PHÂN LỚP DỰA TRÊN LUÂT KẾT HƠP

TÓM TẮT: Phân lớp dựa trên luật kết hợp là một cách tiếp cận thú vị trong khai thác dữ liệu để tạo ra các hệ thống dự đoán chính xác và dễ dàng, dễ giải thích hơn. Cách tiếp cận này thường được xây dựng dựa trên cả kỹ thuật khai thác luật kết hợp và phân lớp, để tìm ra tập luật gọi là luật kết hợp dùng cho phân lớp (CAR) thuộc tính nhãn. Thuật toán ACAC thuộc họ bài toán CPAR nhưng độ chính xác phân loại còn thấp trên dữ liệu lớn. Bài báo này đề xuất ba bước tìm tập thuộc tính tối ưu để nâng cao đáng kể hiệu suất của thuật toán ACAC về mặt thời gian cũng như độ chính xác trên dữ liệu lớn. Kết quả thực nghiệm cho thấy việc chọn tập dữ liệu tối ưu giúp cho thuật toán ACAC trở nên hữu ích hơn trên dữ liệu lớn, đặc biệt chi phí tối ưu tập dữ liệu theo ba bước này hoàn toàn hợp lý trong thực tế.

Từ khóa: CAR, CPAR, Lựa chọn đặc trưng, Entropy, Leo đồi, Dữ liệu lớn.

I. GIỚI THIỆU

Bộ phân loại kết hợp là một mô hình học có giám sát dùng luật kết hợp để gán nhãn. Mô hình có được sau khi huấn luyện bao gồm các luật kết hợp được tạo ra dùng để gán nhãn dữ liệu mới. Vì vậy mô hình cũng có thể được xem là một danh sách các mệnh đề "if-then": nếu dòng dữ liệu mới đáp ứng các tiêu chí nhất định (chứa đủ các thuộc tính phân lớp bên trái), thì dòng dữ liệu này sẽ được phân lớp theo đúng giá trị bên phải của luật [7, 8]. Bộ phân lớp kết hợp kế thừa độ Hỗ trợ và độ Tin cậy để lọc bớt các luật không cần thiết từ dữ liệu huấn luyện. CPAR là một loại phân lớp kết hợp dựa trên các luật kết hợp dựa đoán, phương pháp này kết hợp các ưu điểm của phân lớp kết hợp và phân lớp dựa trên quy tắc truyền thống khi vừa sinh luật trực tiếp trong giai đoạn huấn luyện vừa kiểm tra để tránh bỏ sót các luật quan trọng. Vào năm 2003, Omiecinski đưa ra cách tính All-confidence rất hiệu quả trong khai thác các mẫu kết hợp. Dựa vào độ đo này, nhóm nghiên cứu của Zaixiang Huang đề xuất thuật toán ACAC [1] sử dụng cả độ đo All-confidence và độ hỗ trợ để khai thác những tập không những phổ biến mà các tập mục con còn có mối quan hệ hai chiều, tính chất này rất cần thiết cho việc phân lớp dựa trên luật kết hợp, và đây cũng là một thuật toán tiêu biểu theo phương pháp CPAR [4, 5, 6, 9]. Tuy nhiên thuật toán này vẫn gặp nhiều hạn chế về mặt thời gian và độ chính xác phân lớp khi số thuộc tính dữ liệu lớn. Bài báo này đề xuất phương pháp loại bỏ những thuộc tính không quan trọng giúp cải tiến phương pháp ACAC theo ba giai đoạn như trong Hình 1. Việc cải tiến này giúp tăng tốc độ xử lý đồng thời tăng độ chính xác lên đáng kể.



Hình 1. Ba giai đoạn chọn tập thuộc tính tối ưu

Giai đoạn 1 chọn tập thuộc tính ban đầu theo độ đo Entropy. Ở giai đoạn này, các thuộc tính bị loại bỏ dựa trên mối quan hệ Entropy của thuộc tính đó với thuộc tính nhãn. Độ đo Entropy thể hiện mối tương quan để kiểm tra xem các thuộc tính có đóng vai trò phân lớp cao hay thấp với thuộc tính nhãn và quyết định loại bỏ hay không các thuộc tính tương ứng.

Giai đoạn 2 tìm tập thuộc tính tối ưu xuất phát từ tập thuộc tính trong giai đoạn 1. Tập thuộc tính được chia thành các tập hợp con và huấn luyện mô hình dựa trên những tập con dữ liệu này. Dựa trên độ chính xác của mô hình, ta có thể thêm và bốt các thuộc tính trên tập thuộc tính có được trong giai đoạn 1 và huấn luyện lại mô hình. Phương pháp này hình thành các tập hợp

con bằng cách sử dụng cách tiếp cận tham lam và đánh giá tính chính xác của tất cả các tổ hợp xuất phát từ tập thuộc tính trong giai đoạn 1 thay vì thử hết trên tất cả các tổ hợp thuộc tính có thể có.

Giai đoạn 3 mở rộng tập thuộc tính tối ưu để tăng độ chính xác phân lớp. Mở rộng bằng tập thuộc tính tối ưu trong giai đoan 2 kết hợp với từng thuộc tính còn lai để tìm tập tối ưu mới có độ chính xác cao hơn.

Đóng góp chính của bài báo tập trung vào Mục II.B và Mục 3. Nội dung bài báo được sắp xếp như sau: Mục II.A sẽ đề cập lại việc phân lớp dựa trên luật kết hợp và tính chất độ đo Entropy, Mục II.B trình bày phương pháp ba giai đoạn, kết quả việc thực nghiệm và minh chứng xác định tính hiệu quả được mô tả trong Mục III, và kết luận việc nghiên cứu được trình bày trong Mục IV.

II. PHÂN LỚP THEO CÁC THUỘC TÍNH CÓ LIÊN QUAN

Theo kiểu phân lớp này, dữ liệu huấn luyện T có m thuộc tính A1, A2, ..., Am, và thuộc tính phân lớp. Các giá trị thuộc tính có thể liên tục hay rời rạc. Đối với thuộc tính có giá trị rời rạc, ta có thể dùng các số nguyên dương liên tiếp để biểu diễn. Đối với thuộc tính có giá trị liên tục, ta rời rạc hóa các giá trị này trước và dùng các số nguyên dương liên tiếp để biểu diễn các giá trị rời rạc này. Một tập mục $X=a_{i_1},...,a_{i_j},...,a_{i_k}$ là tập các giá trị của các thuộc tính khác nhau. Như vậy tập mục 1-item được xác định bằng một giá trị của một thuộc tính, và chính là a_{i_j} . Số dòng dự liệu trong T khóp với tập mục X được gọi là độ hỗ trợ của X, biểu diễn là sup(X). All-confidence [10] của tập mục $X=a_{i_1},...,a_{i_j},...,a_{i_k}$ được định nghĩa như sau:

$$allconf(X) = \frac{\sup(X)}{\max(\sup(a_{i_i}),...,\sup(a_{i_i}))}$$

Công thức này xác định độ tin cậy tối thiểu của tất cả những luật được sinh ra từ một tập mục X.

Cho tập dữ liệu huấn luyện T, gọi c là nhãn lớp. Một ruleitem có dạng $\{\text{condset}, c\}$, với condset là một tập mục. Mỗi ruleitem biểu diễn cơ bản một luật: $condset \rightarrow c$. Ruleitem mà condset có k item thì gọi là k-ruleitem. Độ hỗ trợ của condset (gọi là consup) là số lần xuất hiện của condset trong T. Độ hỗ trợ của một luật (gọi là rulesup) là số lần xuất hiện của cả condset và c trong T.

A. Phương pháp truyền thống

Hầu hết các thuật toán phân lớp dựa trên luật kết hợp có ba giai đoạn:

- Bỏ đi những luật có thể gây ra quá khớp hoặc dư thừa.
- Phân lớp dữ liêu mới.

Để cho dễ hình dung, ta xem ví dụ trên tập T cho trước như Bảng 1. Trong ví dụ này, ngưỡng của độ hỗ trợ là 2, ngưỡng all-confidence là 50%, và ngưỡng độ tin cậy là 100%.

A	В	C	D	class
32	55	80	83	90
32 33	52	80	85	89
33	52	80	85	89
33	55	79	82	90
34 34 32	55	79	82	89
34	55	77	82	89
32	55	80	88	90
33	55	79	82	90

Bảng 1. Dữ liệu huấn luyện T

Trong bảng này, ta chọn các ruleitem thỏa ngưỡng condsup và all-confidence và đưa vào trong tập ứng viên 1-ruleitem F1. Từ tập F1, ta chọn những ruleitem nào thỏa ngưỡng confidence và đưa vào tập R1, rồi xóa tập F1 đi.

Bảng 2. Tập ứng viên 1-ruleitem F1

itemset	class	allconf	conf	condsup	rulesup
33	89	100	50	4	2
33	90	100	50	4	2
55	89	100	67	6	2
55	90	100	33	6	4
79	89	100	33	3	1

79	90	100	67	3	2
80	89	100	50	4	2
80	90	100	50	4	2
82	89	100	50	4	2
82	90	100	50	4	2

Bảng 3. Tập luật R1

itemset	class	allconf	conf	condsup	rulesup
32	90	100	100	2	2
34	89	100	100	2	2
52	89	100	100	2	2
85	89	100	100	2	2

Các ứng viên 2-ruleitem được kết hợp từ tập F1, giả sử trong F 1 có 2 luật:

Ruleitem{(55), 90} có rulesup là 4

Ruleitem{(79), 90} có rulesup là 2

Thì luật sinh ra sẽ là $\{(55, 79), 90\}$ có rulesup là 2, all-conf = $2/\max(4,2) = 50\%$. Kết quả các 2-ruleitem được đưa vào bảng F2, và bảng R2 là bảng chứa những 2-ruleitem nào thỏa ngưỡng confidence.

Bảng 4. Tập ứng viên 2-ruleitem F2

itemset	class	allconf	conf	condsup	rulesup
55 79	90	50	67	3	2
55 82	89	50	50	4	2
55 82	90	100	50	4	2
79 82	90	100	67	3	2

Bảng 5. Tập luật R2

itemset	class	allconf	conf	condsup	rulesup
33 55	90	100	100	2	2
33 79	90	100	100	2	2
33 80	89	100	100	2	2
33 82	90	100	100	2	2
55 80	90	100	100	2	2

Quá trình này được lặp đi lặp lại cho đến khi tập ứng viên k-ruleitem rỗng. Thuật toán 1 có tên là ACAC-RG, trong đó tập ứng viên k-itemset Ck, tập k-itemset phổ biến Fk. R là tập các luật. Dòng 1-2 tính condsup và rulesup của từng item đồng thời, rồi hàm ruleSelection được gọi tại mỗi lần lặp (dòng 6). Trong mỗi lần lặp, thuật toán thực hiện 3 thao tác chính. Đầu tiên, các itemset phổ biến Fk-1 được tìm thấy trong lần lặp (k-1) dùng để sinh các itemset ứng viên Ck theo như hàm candidateGen (dòng 4), hàm này tương tự với hàm Apriori-Gen. Tiếp theo, hàm supportCount quét tập dữ liệu và tính condsup và rulesup của các ứng viên trong Ck (dòng 5). Cuối cùng thì thực hiện hàm ruleSelection.

Hàm ruleSelection tính all-confidence của mỗi itemset ứng viên trong mỗi lớp và độ tin cậy của mỗi luật ứng viên (dòng 1 đến 3). Nếu các luật ứng viên thỏa ngưỡng độ hỗ trợ, ngưỡng all-confidence, và ngưỡng độ tin cậy, thì luật này sẽ đưa vào tập R. Nếu luật ứng viên chỉ thỏa ngưỡng độ hỗ trợ, ngưỡng all-confidence, thì đưa nó vào tập Fk (dòng 4 đến 10). Khi một luật ứng viên được chon, condset của nó sẽ không được mở rộng trong vòng lặp con tiếp theo.

Algorithm ACAC-RG(T)

```
Input: tập dữ liệu huấn luyện, minSup, minConf,
minAllConf
Output: R: tập luật cho mô hình dự đoán nhãn
1: C_1 \leftarrow init - pass(T);
2: ruleSelection(C_1, F_1, R);
3: for(k=2; F_{k-1} \neq \emptyset; k++)
          Ck\leftarrow candidateGen(F_{k-1});
4:
5:
          supportCount(C_k);
6:
          ruleSelection(C_k, F_k, R);
7: end for
8: return R:
```

Algorithm ACPredict(R,D)

```
Input:
- R là tập luật được khai thác,
- Dnew (dữ liệu mới chưa có nhãn),
Output: D' (dữ liệu được gán nhãn)
1:D' = empty
2:foreach t<sub>i</sub> in D:
3:
           v_1 = S(R,t_i,Yes),
4:
           v_2 = S(R,t_i,N_0),
5:
           if(v1>v2) then:
6:
                         t_i [Label] \leftarrow Yes;//gán nhãn là
Yes
7:
           else.
8:
                      t<sub>i</sub> [Label] ← No;//gán nhãn là No
           D' = D' + t_i;
9.
10:return D';
```

Sau khi thực hiện thuật toán ACAC-RG, ta thu được tập luật R. Dựa vào tập luật R, các bước phân lớp như sau:

Tính giá tri Info(entropy) cho từng thuộc tính theo công thức như sau:

Info(X)=
$$\frac{1}{\log 2k} \sum_{i=k}^{n} P(C_i|X) \log_2 P(C_i|X)$$

trong đó: k là số lượng lớp, $P(C_i|X)$ là khả năng phù hợp giữa C_i và X.

Đại lương entropy ở trên sẽ được sử dụng để tính S(r), 1 đại lượng đo lường đô phù hợp của các luật theo công thức sau:

$$S(r) = 0.9*(1 - \frac{\sum \sup(Xi) * Info(Xi)}{\sum \sup(Xi)}) + 0.1* \frac{n}{ntot}$$

 $S(r) = 0.9*(1 - \frac{\Sigma sup(Xi)*Info(Xi)}{\Sigma sup(Xi)}) + 0.1*\frac{n}{ntot}$ trong đó: Xi là tập luật con của r_i ($r_i \in R$), n_{tot} là tổng số luật phù hợp với đối tượng mới, n là số lượng luật trong tập R. Sau khi tính toán được giá tri S(r), ACAC dùng nhóm có S(r) cao nhất thể chon nhãn cho dữ liêu mới. Quá trình phân lớp theo công thức S(r) được mô tả trong thuật toán ACPredict.

B. Phương pháp cải tiến

Với tập dữ liệu lớn và nhiều thuộc tính, chi phí để chon ra một tập thuộc tính tối ưu là rất lớn. Phương pháp bài báo đề xuất kết hợp tính chất của độ đo Entropy và phương pháp Leo đồi để lựa chọn đặc trưng có ba giai đoạn. Giai đoạn 1 chủ yếu là lọc thuộc tính không quan trọng theo độ đo Entropy (giá trị Entropy giữa một thuộc tính và thuộc tính phân lớp) và phương pháp lựa chọn đặc trung theo hình thức Leo đổi. Độ chính xác của mô hình sau khi loại bỏ thuộc tính được kiểm nghiệm lại theo phương pháp kiểm tra chéo. Tập thuộc tính được chọn sẽ là tập thuộc tính tối ưu cho sự phân lớp. Giá trị Entropy nằm trong khoảng [0,1], những thuộc tính có giá tri Entropy lớn thì sẽ không có đóng góp nhiều thông tin cho việc phân lớp thì nên được loại bỏ. Trong những tập dữ liệu có ít thông tin, ta có thể tính tất cả các entropy của các thuộc tính và sắp xếp thứ tư từ cao đến thấp. Sau đó, tuần tư kiểm tra việc loại bỏ từng thuộc tính theo đô ưu tiên từ cao đến thấp và đánh giá đô chính xác của tập thuộc tính còn lại. Nếu độ chính xác của tập thuộc tính còn lại bằng hoặc lớn hơn giá trị cũ (khi chưa bỏ thuộc tính) thì thuộc tính được bỏ thật sự là nên bỏ. Vì đây là hướng tiếp cận từ dưới lên (bottom up) nên chi phí lại bỏ trong tập dữ liệu lớn ban đầu là không khả thi. Bài báo đề xuất ngưỡng Entropy ban đầu, thay vì phải thử Entropy theo thứ tự từ cao xuống thấp như trong thuật toán Phase 1.

Trong thuật toán Phase 1, dòng 13 chính là thuật toán ACAC chưa cải tiến, dòng 14 tính đô chính xác của mô hình có được từ dòng 13. Đoạn mã từ dòng 8 đến dòng 18 lặp đi lặp lại công việc cải tiến cho đến khi đô chính xác của mô hình đạt đỉnh. Trong mỗi lần lặp, dựa trên độ entropy lân cận entropy hiện hành (dòng 11, 12). Dựa vào độ entropy e', thuật toán xác định tập thuộc tính De bao gồm những thuộc tính có độ entropy so với lớp nhãn nhỏ hơn e'. Nói theo cách khác, trong lần lặp này thuật toán đã thử loại bỏ thuộc tính được cho là thừa ra khỏi tập huấn luyện. Dòng 15 kiểm tra xem việc loại bỏ liệu có đúng đắn hay không, nếu đúng thì cập nhật lại giá tri tốt hơn. Sau khi thoát khỏi vòng lặp, thuật toán tìm được tập thuộc tính tương đối tốt như dòng số 18. Trong dòng 2, D là tập dữ liệu được chia ra thành 2 nhóm Dtrain và Dtest theo tỷ lệ 80/20. Dựa vào Dtrain để tìm tập luật phân lớp, và Dtest dùng để kiểm thử độ chính xác của tập luật phân lớp.

Algorithm Phase1(FSAC)

Algorithm Phase 2 (Hill Climbing)

```
1:
      minSup, minConf, minAllConf,
                                                                             Input:
      D: dataset (D<sub>train</sub>, D<sub>test</sub>)
                                                                             - T là tập dataset
                                                                             - best_Attributes (tập thuộc tính ban đầu)
3:
     e := entropy threshold
     D_e := \{ \text{các thuộc tính } D_{\text{train }} a_i \} \text{ với entropy}(a_i, \text{label}) \leq e
                                                                             - sub Attributes (tập thuộc tính dùng để thay vào
4:
                                                                             best Attributes)
     neighbor = [e, e \pm \Delta]
5:
                                                                             - all Attributes (tất cả thuộc tính của dataset)
6:
      e' := random(neighbor)
                                                                             - initial Validation (chỉ số đánh giá ban đầu)
7:
      D_{e'} := \{ \text{các thuộc tính } D_{\text{train }} a_i \} \text{ với entropy}(a_i, \text{ label}) \leq
                                                                             Output: tập thuộc tính có chỉ số đánh giá tốt nhất
e'
                                                                              1: delta := \Delta;
8: do
                                                                             2: for(i=0; i \le n; i++)
9:
       model_e := ACAC-RG(D_e);
                                                                                     new Attributes:=GenerateAttributes(best Att
10:
       accuracy_e := model_e.predict(D_{test});
                                                                             ributes, sub Attributes, delta);
       neighbor = [e, e \pm \Delta]
11:
                                                                                      validation:=ACAC(new_Attributes);
                                                                             4.
12.
       e' := random(neighbor);
                                                                             5:
                                                                                      if (validation > initial_Validation) then:
13:
       model_{e'} := ACAC-RG(D_e);
                                                                             6:
                                                                                            initial_Validation := validation;
14:
       accuracy_e' := modele'.predict(D_{test});
                                                                             7:
                                                                                            i:=0;
15:
       if (accuracy<sub>e</sub> \leq accuracy<sub>e</sub>) then:
                                                                             8:
                                                                                            best Attributes := new Attributes
16:
            e := e';
                                                                             9:
                                                                                            sub Attributes:= all Attributes- best At
17:
        end if
                                                                             tributes
18: while (accuracy<sub>e</sub>) \leq accuracy<sub>e</sub>);
                                                                              10:
                                                                                       end if
19: return model<sub>e</sub>;
                                                                              11: end for
                                                                              12: return best_Attributes;
```

Kết quả của việc thực hiện Phase 1, ta thu được tập thuộc tính ban đầu tương đối tốt nhưng chưa thực sự tối ưu vì hạn chế của phương pháp chọn dữ liệu theo hướng lọc (Filter method) từng thuộc tính. Giai đoạn 2 mô phỏng phương pháp Leo đồi với nghiệm đầu tiên là tập thuộc tính ban đầu thu được từ giai đoạn 1. Giai đoạn này, thuộc tính đang được chia thành 2 nhóm thuộc tính (best_Attributes và Sub_attributes). Tập thuộc tính best_Attributes chứa các thuộc tính thu được từ giai đoạn 1, tạm cho là tập thuộc tính tốt. Tập thuộc tính sub_Attributes chứa các thuộc tính còn lại. Thuật toán Phase 2 thực hiện công việc tìm xem có tập nào tốt hơn tập best_Attributes hiện hành hay không, nếu có thì cập nhật lại tập best_Attributes, đây là công việc từ dòng 3 đến dòng 10 trong thuật toán Phase 2. Công việc này lặp đi lặp lại cho đến khi tìm được tập tốt nhất với số lượng thuộc tính vẫn không thay đổi so với tập thuộc tính thu được từ giai đoạn 1. Trong mỗi lần lặp như vậy là thử trên tập thuộc tính mới được sinh ra từ hàm *GenerateAttributes* trong dòng số 3, dựa vào giá trị *delta* để đưa ra số lượng *delta* thuộc tính ngẫu nhiên nào đó trong tập best_Attribute, và bổ sung số lượng *delta* thuộc tính ngẫu nhiên từ tập sub_Attributes vào tập best_Attributes. Nếu tập thuộc tính mới có kết quả tốt hơn tập thuộc tính best_Attributes hiện hành thì cập nhật lại (dòng 5 đến dòng 8 trong thuật toán Phase 2).

Agorithm Phase 3

```
1: best Attributes :={};
                                                          2: for(i=0;i< sub_Attributes.length;i++) do:
- initial Attributes (tâp thuộc tính ban đầu của
                                                                         new_Attributes := initial _Attributes +
                                                          3:
- sub Attributes (tập thuộc tính dùng để thay vào
                                                          sub_Attributes[i];
                                                          4:
                                                                    validation := ACAC(new Attributes);
best Attributes)
                                                          5:
                                                                    if (validation > initial Validation) then:
- T là tập dataset
                                                          6:
                                                                            initial Validation := validation
- initial Validation (chỉ số đánh giá ban đầu)
                                                                            best Attributes := new_Attributes;
                                                          7:
Output: tập thuộc tính có chỉ số đánh giá tốt nhất
                                                          8: return best Attributes
```

Sau khi xong giai đoạn 2 (Phase 2), ta được tập thuộc tính tối ưu với số lượng thuộc tính bằng với tấp thuộc tính thu được trong giai đoạn 1. Giai đoạn 3 là giai đoạn mở rộng số lượng tập thuộc tính nhằm tìm giá trị tối ưu hơn xuất phát từ tập thuộc tính thu được trong giai đoạn 2. Dòng 2 đến dòng 7 trong thuật toán Phase 3 tiến hành việc thử tìm tập mở rộng tối ưu hơn bằng việc ghép tập thu được trong giai đoạn 3 ghép với từng thuộc tính còn lại trong tập sub_Attributes.

III. THỰC NGHIỆM

Môi trường thực nghiệm được xử lý tập trung trên máy tính có cấu hình Intel(R), Core(TM) i7-6820HQ CPU @ 2.70 GHz (8 CPUs), ~2.7 GHz và hệ điều hành Windows 10. Trong bài báo này thực nghiệm được thực hiện trên tập dữ liệu Spam Emails Dataset [3] một tập dự liệu khá phức tạp, dữ liệu có 4602 dòng và 58 cột. Trong đó, thuộc tính cuối là thuộc tính nhãn có tên là 'spam' để phân loại email có là spam hay ham (spam = 1, ham = 0). Dữ liệu này cũng được dùng trong thuật toán ACAC. Trong 57 thuộc tính còn lại có nhiều thuộc tính không có giá trị trong việc phân lớp cần phải loại bỏ. Để loại bỏ, bài báo dùng độ entropy để xác định khả năng đóng góp trong việc phân lớp của từng thuộc tính. Thuộc tính nào có giá trị entropy cao sẽ được loại bỏ khỏi tập dữ liệu, tập dữ liệu mới được thử lại theo phương pháp ACAC để tìm tập luật kết hợp 5 dùng cho phân lớp. Tập luật này sẽ được kiểm thử theo phương pháp kiểm tra chéo (80% - 5000 dòng dữ liệu dùng cho việc huấn luyện, 20% - 1001 dòng dữ liệu dùng cho việc kiểm thử). Nếu độ chính xác tăng lên thì việc loại bỏ thuộc tính là đúng, và việc này được thực hiện cho đến khi độ chính xác đạt giá trị cao nhất. Nói một cách khác, công việc sẽ dừng khi độ chính xác bị suy giảm.

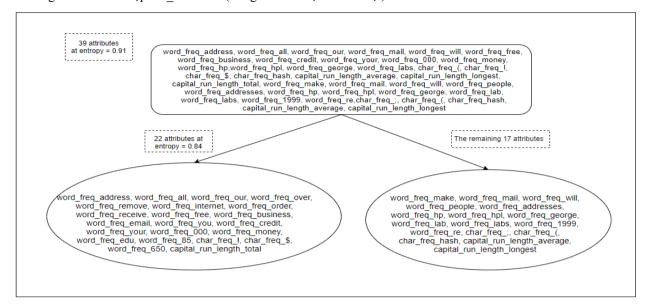
Với 57 thuộc tính phân lớp, việc phân lớp dựa trên thuật toán ACAC [1] là một công việc có chi phí thực hiện (thời gian và không gian lưu trữ) rất cao và các chỉ số đánh giá chưa chắc đã ở một mức cao. Bởi điểm hạn chế đó, ta có thể sử dụng độ đo entropy để loại bỏ bớt số luật rời rạc nhằm giảm thời gian thực hiện việc phân lớp như thuật toán Phase 1 thực hiện.

entropy	Attributes	Rules	Runtimes(ms)	Accuracy(%)
0,9	39	?	?	?
0,885	33	3108	7.78375E+13	94,2
0,88	31	2239	3.4375E+11	94,4
0,875	29	1612	11375000000	94,9
0,87	28	1254	261000000	94,06
0,865	27	989	51000000	93,04
0,86	26	692	10000000	92,4
0,85	24	338	2200000	90,99
0,84	22	109	35170	88,93
0,83	17	24	343	87,04
0,82	15	19	289	81,11
0,81	14	17	151	81,11
0,79	13	13	145	80,61

Bảng 6. Giá trị thực nghiệm của FSAC trên từng ngưỡng entropy

Để loại bỏ những thuộc tính không có vai trò phân lớp, thực nghiệm loại bỏ các thuộc tính có Entropy > 0,9. Qua đó số thuộc tính giảm từ 57 xuống còn 39 thuộc tính như trong Bảng 6, thời gian xử lý và cải thiện các chỉ số đánh giá. Mặc dù đã loại bỏ 18 thuộc tính nhưng số thuộc tính vẫn là quá lớn cho việc tìm ra tập luật kết hợp phân lớp. Trong giai đoạn 1, thực nghiệm

tìm ra tập thuộc tính ban đầu sao cho thời gian xử lý vẫn chấp nhận được và độ chính xác hợp lý. Và tập thuộc tính ban đầu tìm được từ thuật toán Phase 1 trong thực nghiệm này là 22 thuộc tính có độ Entropy \leq 0,84, tập 17 thuộc tính còn lại được dùng trong phép thử tối ưu trong giai đoạn 2. Phase 1 dùng tại giá trị Entropy = 0,84 và đạt 22 thuộc tính vì cận bằng giữa hiệu suất và thời gian xử lý, khi giá trị Entropy = 0,85 lúc này số thuộc tính là 24 (xem Bảng 6, cột 2) thì chi phí về thời gian thực hiện tăng lên đáng kể mặt dù độ chính xác vẫn còn tăng. Hình 2 mô tả kết quả sau khi thực hiện xong thuật toán Phase 1 là chia tập dữ liệu ban đầu thành 2 tập con. Tập best_Attributes (22 thuộc tính thu được từ thuật toán Phase 1) được cho là tập thuộc tính tương đối tốt nhưng chưa tối ưu và tập sub Attributes (bao gồm các thuộc tính còn lại).



Hình 2. Tập thuộc tính dữ liệu sau khi thực hiện Phase 1



Hình 3. Kết quả Phase 2 theo từng delta

Sau khi chọn được tập thuộc tính ban đầu (best_Attributes) là tập 22 thuộc tính tại ngưỡng entropy = 0,84. bước tiếp theo đó chính là tìm ra tập 22 thuộc tính tối ưu hơn 22 thuộc tính ban đầu. Để thực hiện được điều đó, phần thực nghiệm sẽ được tiến hành chạy thuật toán Phase 2. Phần thực nghiệm này chọn giá trị delta lần lượt là 1, 2 và 3 với mỗi giá trị delta khác nhau sẽ cho những kết quả khác nhau như trong Hình 3. Theo phương pháp Leo đồi vẫn còn nhiều hạn chế, chưa thể kết luận giá trị delta ảnh hưởng như thế nào đến việc tối ưu, nhưng trong thực nghiệm này giá trị delta = 1 mang lại kết quả tốt nhất với tập thuộc tính tối ưu có độ chính xác là 0,9327 so với các giá trị còn lại. Giá trị delta = 2 trong thực nghiệm này có kết quả kém nhất (tốn 88 lần lặp so với 31 lần lặp khi delta = 3) về mặt thời gian cũng như độ chính xác.

Qua giai đoạn 2, Ta đã chọn ra được tập luật gồm có 22 thuộc tính có độ chính xác = 0,9305 (tương đối tốt vì thời gian thực hiện của delta = 3 chỉ là 31 lần lặp so với thời gian delta = 1 mất 139 lần lặp) và 17 thuộc tính còn lại. Ở giai đoạn 3, thuật toán Phase 3 tìm ra tập 23 thuộc tính có độ chính xác cao hơn 0,9305. Để cụ thể hóa mục tiêu trên, phần thực nghiệm này sẽ kết hợp tập 22 thuộc tính và từng thuộc tính trong 17 thuộc tính còn lại. mỗi tập 23 thuộc tính sẽ cho những kết quả có thể khác nhau như trong Bảng 7. Khi thực nghiệm Phase 3, ta có được 2 thuộc tính "word_freq_mail" và thuộc tính "char_freq_;" giúp cho tập thuộc tính tối ưu có kết quả tốt hơn.

Index	entropy	Tên thuộc tính	Time(ms)	rules	Accuracy(%)
1	0,30840	capital_run_length_average	3195068	640	0,9305
2	0,65802	capital_run_length_longest	2991905	640	0,9305
3	0,78659	word_freq_hp	26753570	604	0,9305
4	0,78937	char_freq_(3680278	605	0,9305
5	0,83027	word_freq_george	31908135	604	0,9305
6	0,83569	word_freq_mail	54212858	957	0,9349
7	0,83673	word_freq_hpl	59496226	604	0,9305
8	0,83897	word_freq_will	23804363	917	0,9305
9	0,83950	char_freq_hash	14989621	861	0,9305
10	0,87196	word_freq_re	17217569	703	0,9305
11	0,87707	word_freq_people	25790388	700	0,9305
12	0,87805	word_freq_1999	20146323	604	0,9305
13	0,88005	word_freq_make	71690336	645	0,9305
14	0,88494	char_freq_;	19102413	1039	0,9392
15	0,88756	word_freq_addresses	22861710	725	0,9305
16	0,89806	word_freq_labs	32724505	604	0,9305
17	0,90796	word_freq_lab	28166730	604	0,9305

Bảng 7. Kết quả sự kết hợp từng thuộc tính trong tập sub_Attributes với tập best_Attributes

Có một điều thú vị là sau khi kết thúc giai đoạn 1, thuật toán Phase 1 dùng tại giá trị Entropy = 0,84. Trong Bảng 7, có 9 thuộc tính đầu tiên có giá trị Entropy < 0,84 và 8 thuộc tính cuối cùng > 0,84. Trong 9 thuộc tính đầu tiên có thuộc tính "word_freq_mail" có giá trị Entropy = 0,83569 giúp cho tập thuộc tính tối ưu từ Phase 2 tăng độ chính xác phân lớp đến 0,9349. Trong 8 thuộc tính cuối cùng, có thuộc tính "char_freq_;" có giá trị Entropy = 0,88494 giúp cho tập thuộc tính tối ưu từ Phase 2 tăng độ chính xác phân lớp đến 0,9392. Đây cũng là điều hạn chế của phương pháp lọc đặc trưng mà bài báo đã trình bày ở trên.

IV. KÉT LUẬN

Thuật toán ACAC chỉ hiệu quả trên tập dữ liệu nhỏ như Mushroom [2], đối với tập dữ liệu lớn và có nhiều thuộc tính như SpamMail [3] thuật toán bộc lộ hạn chế. Điều này làm cho thuật toán ACAC nói riêng và các thuật toán CAR nói chung không có tính thực tế khi phân lớp dữ liệu, trong khi các thuật toán CAR là loại thuộc toán phân lớp dễ giải thích (có tính explainable) lý do phân lớp cho người dùng cuối thấu hiểu nguyên nhân phân lớp.

Phương pháp ba giai đoạn trong bài báo đề cập là phương pháp lựa chọn đặc trưng theo hướng học có giám sát. Giai đoạn một thực hiệu việc lọc đặc trưng theo độ đo Entropy, giai đoạn hai thực hiện việc lựa chọn đặc trưng tối ưu theo nhóm nhỏ mô phỏng hình thức Leo đồi, giai đoạn ba mở rộng tập tối ưu nhằm tìm bộ phân lớp có kết quả chính xác hơn. Cả ba giai đoạn đều có điểm mạnh và điểm hạn chế riêng. Tuy nhiên, theo phương pháp thực hiện linh hoạt như mô tả rõ trong thực nghiệm. Kết quả cuối cùng vẫn tìm ra tập thuộc tính tối ưu về mặt phân lớp cũng như thời gian xử lý chấp nhân được.

Phương pháp này có thể mở rộng để đạt kết quả cao hơn khi vượt qua hạn chế của Leo đồi bằng các phương pháp khác như Luyện thép, Di truyền, Bầy đàn. Bên cạnh đó ta cũng có thể tiến hành trên các tập dữ liệu phức tạp hơn nhằm đánh giá độ ổn định cũng như tính chắc chắn của phương pháp.

TÀI LIỆU THAM KHẢO

- [1] Z. Huang, Z. Zhou, T. He, and X. Wang, "ACAC: Associative Classification Based on All-Confidence," *IEEE International Conference on Granular Computing*, pp. 289–293, 2011
- [2] https://www.kaggle.com/datasets/uciml/mushroom-classification
- [3] https://www.kaggle.com/datasets/yasserh/spamemailsdataset
- [4] H. F. Ong, C. Y. M. Neoh, V. K. Vijayaraj, Y. X. Low, "Information-Based Rule Ranking for Associative Classification," ISPACS, 2022.
- [5] M. Abrar, A. Tze and S. Abbas, "Associative Classification using Automata with Structure based Merging," *International Journal of Advanced Computer Science and Applications*, vol. 10, 2019
- [6] D. L. Olson and G. Lauhoff, "Market Basket Analysis," in *Descriptive Data Mining, Singapore*, Springer Singapore, pp. 31-44, 2019.

- [7] K. D. Rajab, "New Associative Classification Method Based on Rule Pruning for Classification of Datasets", *IEEE Access*, vol. 7, pp. 157783-157795, 2019.
- [8] H. F. Ong, N. Mustapha, H. Hamdan, R. Rosli and A. Mustapha, "Informative top-k class associative rule for cancer biomarker discovery on microarray data," *Expert Systems with Applications*, vol. 146, 2020.
- [9] Majid Seyf, Yue Xu, Richi Nayak, "DAC: Discriminative Associative Classification," SN Computer Science (2023) 4:401
- [10] E.R.Omiecinski, "Alternative interest measures for mining associations in databases," *IEEE Transactions on Knowledge and Data Engineering*, vol 15, pp. 57-69, 2003.

AN EFFICIENT APPROACH FOR ASSOCIATIVE CLASSIFICATION

ABSTRACT: Associative Classification is an interesting approach in data mining to create more accurate and easily interpretable predictive systems. This approach is often built on both association rule mining and classification techniques, to find a set of rules called association rules for classification (CAR) of label attributes. ACAC algorithm belongs to the family of CPAR problems but the classification accuracy is still low on big data. This paper proposes three phases to find the optimal attribute set to significantly improve the performance of ACAC algorithm in terms of time as well as accuracy on large data. Experimental results show that choosing the optimal data set makes the ACAC algorithm more useful on big data, especially the cost of optimizing the data set in these three phases is completely reasonable in practice.

NHÀ XUẤT BẢN KHOA HỌC TỰ NHIÊN VÀ CÔNG NGHỆ

Nhà A16 - Số 18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội Điên thoại: Phòng Phát hành: **024.22149040**;

Diện thoại: Phong Phát hành: 024.22149040; Phòng Biên tập: 024.37917148;

Phòng Quản lý Tổng hợp: **024.22149041**;

Fax: 024.37910147; Email: nxb@vap.ac.vn; Website: www.vap.ac.vn

FAIR

KỶ YẾU HỘI NGHỊ KHOA HỌC CÔNG NGHỆ QUỐC GIA LẦN THỨ XVI Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin

Proceedings of the 16th National Conference on Fundamental and Applied Information Technology Research (FAIR'2023)

Trường Đại học Sư phạm Kỹ thuật - Đại học Đà Nẵng, ngày 28-29/9/2023

LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT VIỆT NAM

Chịu trách nhiệm xuất bản Giám đốc, Tổng biên tập PHẠM THỊ HIẾU

Biên tập: Nguyễn Thị Chiên, Hà Thị Thu Trang

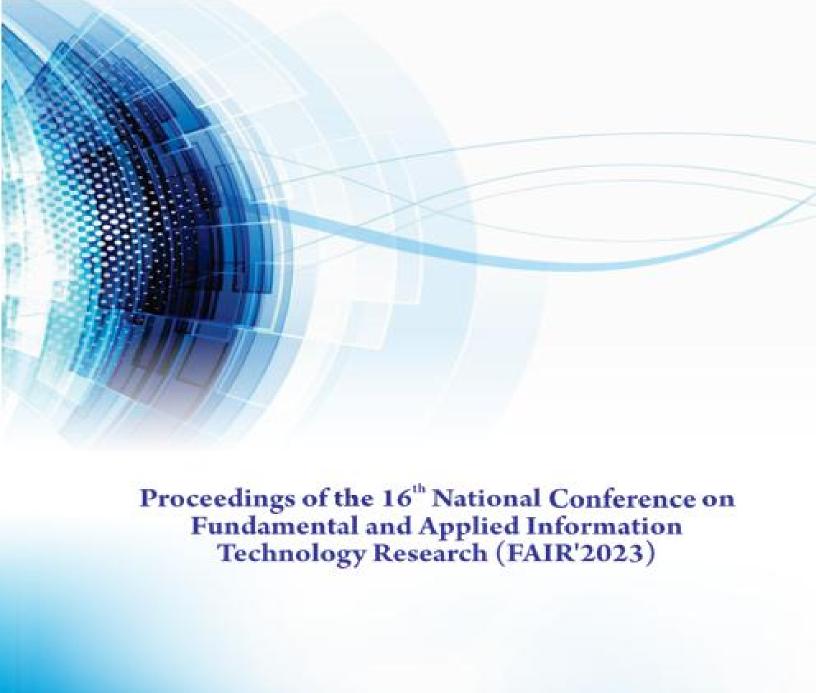
Trình bày kỹ thuật: Đỗ Hồng Ngân Trình bày bìa: Đỗ Hồng Ngân

Liên kết xuất bản:

Liên hiệp các Hội Khoa học và Kỹ thuật Việt Nam Địa chỉ: Lô D20, Khu Đô thị mới Cầu Giấy, ngõ 19 Phố Duy Tân, phường Dịch Vọng Hậu, quận Cầu Giấy, Hà Nội

ISBN: 978-604-357-201-8

In 100 cuốn, khổ 20x29,5 cm, tại Công ty Cổ phần Khoa học và Công nghệ Hoàng Quốc Việt. Địa chỉ: Số 11 ngách 1, ngõ 1 Võ Chí Công, phường Nghĩa Đô, quận Cầu Giấy, Hà Nội. Số xác nhận đăng ký xuất bản: 3686-2023/CXBIPH/04-40/KHTNVCN. Số quyết định xuất bản: 91/QĐ-KHTNCN, cấp ngày 21 tháng 12 năm 2023. In xong và nộp lưu chiều quý IV năm 2023.



9 78 60 4 3 5 7 2 0 1 8 SÁCH KHỐNG BÁN