

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

INTHAVONG SOUKSAKHONE

**NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN LỚP
DỮ LIỆU VÀ ỨNG DỤNG TRONG PHÂN LỚP NẤM
(MUSHROOM) VỚI CÔNG CỤ WEKA**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên – 2020

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

INTHAVONG SOUKSAKHONE

**NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN LỚP DỮ
LIỆU VÀ ỨNG DỤNG TRONG PHÂN LỚP NẤM
(MUSHROOM) VỚI CÔNG CỤ WEKA**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Chuyên ngành: **KHOA HỌC MÁY TÍNH**

Mã số: **84 8 01 01**

Người hướng dẫn khoa học: **TS. Nguyễn Văn Núi**

Thái Nguyên – 2020

LỜI CẢM ƠN

Trước tiên, tôi xin được gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Thầy giáo, **TS. Nguyễn Văn Núi** đã tận tình chỉ bảo, hướng dẫn, động viên và giúp đỡ tôi trong suốt quá trình tôi thực hiện luận văn tốt nghiệp.

Tôi xin gửi lời cảm ơn tới các thầy cô Trường Đại Học Công nghệ Thông Tin và Truyền Thông – Đại học Thái Nguyên, những người đã tận tình giúp đỡ, hướng dẫn trong quá trình tôi học tập tại trường.

Cuối cùng, tôi muốn gửi lời cảm ơn tới gia đình và bạn bè, những người thân yêu luôn bên cạnh, quan tâm, động viên tôi trong suốt quá trình học tập và thực hiện luận văn tốt nghiệp này.

Tôi xin chân thành cảm ơn!

Thái Nguyên, tháng 11 năm 2020

Học viên

Inthavong Souksakhone

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong Luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Những điều được trình bày trong nội dung Luận văn, hoặc là của cá nhân hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn đúng quy cách. Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Thái Nguyên, tháng 11 năm 2020

Tác giả luận văn

Inthavong Souksakhone

MỤC LỤC

LỜI CẢM ƠN	I
LỜI CAM ĐOAN	II
MỤC LỤC.....	III
DANH SÁNH BẢNG	VI
DANH SÁNH HÌNH VẼ.....	VII
DANH SÁCH TỪ VIẾT TẮT	IX
CHƯƠNG 1 TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ PHÁT HIỆN TRI THỨC.....	3
1.1 Giới thiệu tổng quan	3
1.1.1 Khái niệm khai phá dữ liệu.....	3
1.1.2 Nhiệm vụ của khai phá dữ liệu	4
1.1.3 Một số ứng dụng khai phá dữ liệu	4
1.1.4 Bước phát triển của việc tổ chức và khai thác các CSDL	5
1.1.5 Quá trình phát hiện tri thức.....	6
1.1.6 Các bước của quá trình KPDL.....	8
1.2. Một số kỹ thuật khai phá dữ liệu cơ bản.....	10
1.2.1 Khai phá dữ liệu dự đoán.....	10
1.2.1.1 Phân lớp (Classification)	10
1.2.1.2 Hồi quy (Regression).....	11
1.2.2 Khai phá dữ liệu mô tả.....	11
1.2.2.1 Phân cụm	11
1.2.2.2 Khai phá luật kết hợp	12
1.3 Một số so sánh giữa khai phá dữ liệu và các phương pháp cơ bản khác	12
1.3.1 So sánh với phương pháp hệ chuyên gia (Expert Systems)	13
1.3.2 So sánh với phương pháp thống kê (Statistics)	14
1.3.3 So sánh với phương pháp học máy (Machine Learning).....	14
1.3.4 So sánh với phương pháp học sâu (Deep Learning).....	15

1.4 Tổng kết chương	18
CHƯƠNG 2 MỘT SỐ PHƯƠNG PHÁP VÀ KỸ THUẬT PHÂN LỚP DỮ	
LIỆU	19
2.1 Tổng quan về phân lớp dữ liệu.....	19
2.2 Phân lớp dữ liệu bằng cây quyết định	22
2.2.1 Độ lợi thông tin.....	26
2.2.2 Tỉ số độ lợi.....	29
2.2.3 Chỉ số Gini.....	30
2.2.4 Tia cây quyết định	32
2.3 Phân lớp dữ liệu Bayesian	33
2.3.1 Định lý Bayes	33
2.3.2 Phân lớp Naïve Bayes.....	34
2.4. Phân lớp dữ liệu sử dụng máy hỗ trợ vector (SVM)	36
2.4.1 Phân lớp đa lớp với SVM	40
2.5. Phân lớp dữ liệu với Random Forest (rừng ngẫu nhiên)	40
2.6 Một số phương pháp phân lớp dữ liệu khác	44
2.6.1 Thuật toán phân lớp k-NN.....	44
2.7 Đánh giá mô hình phân lớp dữ liệu	44
2.8 Tổng kết chương	46
CHƯƠNG 3 ỨNG DỤNG PHÂN LỚP DỮ LIỆU MUSHROOM VỚI CÔNG	
CỤ WEKA VÀ MỘT SỐ THUẬT TOÁN CƠ BẢN.....	47
3.1 Giới thiệu bài toán phân lớp dữ liệu Mushroom.....	47
3.1.1 Giới thiệu về bài toán phân lớp dữ liệu Mushroom.....	47
3.1.2. Thu thập, tiền xử lý và mã hóa dữ liệu.....	47
3.1.3. Mô tả sơ lược về dữ liệu.....	51
3.2 Giới thiệu về công cụ Weka, cấu hình và ứng dụng phân lớp Mushroom	52
3.2.1 Môi trường Explorer.....	53

3.2.2 Khuôn dạng của tập dữ liệu	54
3.2.3 Tiền xử lý dữ liệu.....	54
3.2.4 Phân tích chức năng phân lớp (Classify)	54
3.2.5 Mô tả chức năng phân lớp (Classify).....	58
3.3 Áp dụng các phương pháp phân lớp trên tập dữ liệu Mushroom	60
3.3.1 Thực hiện phân lớp bằng thuật toán Naive Bayes	61
3.3.2 Thực hiện phân lớp bằng thuật toán k-Nearest neighbor	63
3.3.3 Thực hiện phân lớp bằng thuật toán Support Vector Machines	66
3.4 Đánh giá mô hình phân lớp dữ liệu Mushroom	70
3.4.1 Đánh giá mô hình bằng phương pháp Hold-out	70
3.4.2 Đánh giá mô hình bằng phương pháp k-fold Cross validation.....	71
3.5 Kết luận thực nghiệm phân lớp dữ liệu Mushroom.....	71
3.6 Tổng kết chương	72
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	73
TÀI LIỆU THAM KHẢO	74

DANH SÁNH BẢNG

Bảng 2.1: Bảng dữ liệu khách hàng	25
Bảng 2.3: Bảng biểu diễn ma trận nhầm lẫn	45
Bảng 3.1: Bảng tổng hợp dữ liệu thu thập	47
Bảng 3.2: Các tính năng dành cho các dữ liệu nầm	48
Bảng 3.3: Mô tả ý nghĩa các giá trị dữ liệu nầm	50
Bảng 3.4: Hiệu năng của mô hình dự đoán, đánh giá bởi kiểm tra 70%	70
Bảng 3.5: Hiệu năng của mô hình dự đoán, đánh giá bởi kiểm tra chéo mặt (fold=10 cross-validation)	71

DANH SÁNH HÌNH VẼ

Hình 1.1: Quá trình phát hiện tri thức	6
Hình 1.2: Quá trình khai phá dữ liệu (KPDL)	9
Hình 1.3: Phân cụm tập dữ liệu cho vay thành 3 cụm	12
Hình 1.4: Một số lĩnh vực ứng dụng của trí tuệ nhân tạo	13
Hình 1.5: Học sau nhận dạng khuôn mặt hoặc biểu hiện cảm xúc trên khuôn mặt	16
Hình 2.1: Quá trình phân lớp dữ liệu - (a) Bước xây dựng mô hình phân lớp	21
Hình 2.2 : Quá trình phân lớp dữ liệu - (b1) Ước lượng độ chính xác của mô hình	22
Hình 2.3: Quá trình phân lớp dữ liệu - (b2) Phân lớp dữ liệu mới	22
Hình 2.4: Phân lớp cho bài toán cho vay vốn của ngân hàng	23
Hình 2.5: Thuật toán xây dựng cây quyết định	24
Hình 2.6: Minh họa cây quyết định	26
Hình 2.7: Thuộc tính tuổi có thông tin thu được cao nhất	29
Hình 2.8 : Các điểm trong không gian D chiều	36
Hình 2.9: Siêu phẳng phân lớp các điểm trong không gian	37
Hình 2.10: Đồ thị biểu diễn các điểm trong mặt phẳng R^+	37
Hình 2.11: Các điểm lựa chọn cho siêu phẳng	38
Hình 2.12: Kiến trúc mô hình SVM	38
Hình 2.13: Đồ thị biểu diễn siêu phẳng tìm được	39
Hình 2.14: Mô hình rừng ngẫu nhiên	42
Hình 2.15: Mô hình chia tập dữ liệu Hold-out	45
Hình 2.16: Mô hình chia tập dữ liệu Cross validation	46
Hình 3.1: Sơ đồ Phương pháp phân lớp nấm (Mushroom)	49
Hình 3.2 : Load Mushroom data	51
Hình 3.3: Giao diện ban đầu Phần mềm WEKA	52
Hình 3.4: Giao diện của WEKA Explorer	53
Hình 3.5: Biểu diễn tập dữ liệu weather trong tập tin văn bản(text)	54
Hình 3.6: Biểu diễn đọc dữ liệu vào chương trình Weka	55

Hình 3.7: Biểu diễn chọn tab Classify để phân lớp.....	55
Hình 3.8: Biểu diễn chọn thuật toán phân lớp và xác định tham số	56
Hình 3.9: Biểu diễn chọn kiểu test	56
Hình 3.10: Chạy thuật toán phân lớp	57
Hình 3.11: Bảng lưu thông tin.....	57
Hình 3.12: Bảng kết quả sau chạy thuật toán phân lớp.....	58
Hình 3.13: Giải thích Running Information	58
Hình 3.14: Giải thích Classifier model (full training set)	59
Hình 3.15: Giải thích xem xét tổng kết số liệu thống kê tập dữ liệu	59
Hình 3.16: Xem độ chính xác chi tiết cho từng phân lớp	59
Hình 3.17: Confusion matrix của bộ phân lớp dữ liệu Mushroom	60
Hình 3.18: Sơ đồ tổng thể Mô hình phân lớp dự đoán nấm (mushroom).....	60
Hình 3.19: Cấu hình Weka cho thuật toán Naive Bayes.....	61
Hình 3.20: Kết quả phân lớp Weka cho thuật toán Naive Bayes với số 70% Split.....	62
Hình 3.21: Kết quả phân lớp Weka cho thuật toán Naive Bayes kiểm tra chéo 10 mặt.....	63
Hình 3.22: Cấu hình Weka cho thuật toán k-NN	64
Hình 3.23: Cấu hình Weka cho thuật toán tìm kiếm trong thuật toán k-NN	64
Hình 3.24: Kết quả phân lớp Weka cho thuật toán k-NN với số 70% Split	65
Hình 3.25: Kết quả phân lớp Weka cho thuật toán k-NN kiểm tra chéo 10 mặt.....	65
Hình 3.26: Cấu hình Weka cho thuật toán SVM	66
Hình 3.27: Kết quả phân lớp Weka cho thuật toán SVM với số 70% Split.....	67
Hình 3.28: Kết quả phân lớp Weka cho thuật toán SVM kiểm tra chéo 10 mặt	67
Hình 3.29: Cấu hình Weka cho thuật toán J48	68
Hình 3.30: Kết quả phân lớp Weka cho thuật toán J48 decision với số 70% Split	68
Hình 3.31: Kết quả phân lớp Weka cho thuật toán J48 kiểm tra chéo 10 mặt.....	69
Hình 3.32: Mô hình cây quyết định hiển thị bởi Hold-out J48	69
Hình 3.33: cây quyết định Visualization.....	70