ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

VILAISAK SOPHABMIXAY

NGHÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN LỚP VÀ ỨNG DỤNG TRONG PHÂN LỚP DỮ LIỆU PROTEIN SUMO HÓA.

Chuyên ngành: **Khoa học máy tính**

Mã số chuyên ngành: **84 80 10 1**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. NGUYỄN VĂN NÚI

THÁI NGUYÊN - 2019

LỜI CAM ĐOAN

Luận văn này là công trình nghiên cứu của cá nhân tôi, được thực hiện dưới sự hướng dẫn khoa học của TS. Nguyễn Văn Núi. Các số liệu, những kết luận nghiên cứu được trình bày trong luận văn này hoàn toàn trung thực.

Học Viên

Vilaisak SOPHABMIXAY

LỜI CẢM ƠN

Để có thể hoàn thành đề tài luận văn thạc sĩ một cách hoàn chỉnh, bên cạnh sự nỗ lực cố gắng của bản thân còn có sự hướng dẫn nhiệt tình của quý Thầy Cô, cũng như sự động viên ủng hộ của gia đình và bạn bè trong suốt thời gian học tập nghiên cứu và thực hiện luận văn thạc sĩ.

Xin chân thành bày tỏ lòng biết ơn đến Thầy TS. Nguyễn Văn Núi người đã hết lòng giúp đỡ và tạo mọi điều kiện tốt nhất cho em hoàn thành luận văn này. Xin chân thành bày tỏ lòng biết ơn đến toàn thể quý thầy cô trong khoa học máy tính nói riêng và trường Đại học Công Nghệ Thông Tin và Truyền Thông Thái Nguyên nói chung đã dạy bảo, cung cấp những kiến thức quý báu cho em trong suốt quá trình học tập và nghiên cứu tại trường.

Cuối cùng, tôi xin chân thành cảm ơn đến gia đình, các anh chị và các bạn đồng nghiệp đã hỗ trợ cho tôi rất nhiều trong suốt quá trình học tập, nghiên cứu và thực hiện đề tài luận văn thạc sĩ một cách hoàn chỉnh.

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN iii
MỤC LỤC iv
DANH MỤC CÁC TỪ VIẾT TẮT vii
DANH MỤC CÁC BẢNG, BIỂU ix
DANH MỤC HÌNH VẼ x
MỞ ĐẦU 1
CHƯƠNG 1 TỔNG QUAN KHAI PHÁ DỮ LIỆU VÀ PHÁT HIỆN TRI THỨC 3
1.1 Giới thiệu chung 3
1.1.1 Khái niệm khai phá dữ liệu 3
1.1.2 Các bước của quá trình phát hiện tri thức 4
1.2 Tổng quan các kỹ thuật khai phá dữ liệu cơ bản 5
1.2.1 Khai phá dữ liệu dự đoán 6
1.2.1.1 Phân lớp 6
1.2.1.2 Hồi quy 7
1.2.2 Khai phá dữ liệu mô tả 7
1.2.2.1 Phân cụm 7
1.2.2.2 Luật kết hợp 8
1.3. Phân tích, so sánh với các phương pháp cơ bản khác 8
1.3.1 So sánh với phương pháp hệ chuyên gia (Expert Systems) 9
1.3.2 So sánh với phương pháp thống kê (Statistics) 9
1.3.3 So sánh với phương pháp học máy (Machine Learning) 10
1.3.4 So sánh với phương pháp học sâu (Deep Learning) 10
CHƯƠNG 2 CÁC PHƯƠNG PHÁP VÀ KỸ THUẬT PHÂN LỚP DỮ LIỆU 12
2.1 Tổng quan về phân lớp dữ liệu 13
2.2 Phân lớp dữ liệu bằng cây quyết định 15
2.2.1 Cây quyết định quy nạp 16

2.2.2 Cây cắt tỉa 20
2.2.3 Trích luật phân lớp từ các cây quyết định 20
2.2.4 Cải tiến cây quyết định quy nạp cơ bản 21
2.2.5 Khả năng mở rộng và cây quyết định quy nạp 22
2.3 Phân lớp dữ liệu Bayesian 23
2.3.1 Định lý Bayes 24
2.3.2 Phân lớp Bayesian ngây thơ 25
2.3.3 Các mạng belief Bayesian 27
2.3.4 Huấn luyện các mạng belief Bayesian 29
2.4 Phân lớp dữ liệu với Random Forest (rừng ngẫu nhiên) 30
2.5 Phân lớp dữ liệu sử dụng máy hỗ trợ vector 33
2.5.1 SVM cho bài toán phân lớp tuyến tính 33
2.5.2 SVM cho phân lớp phi tuyến 37
2.6 Một số phương pháp phân lớp dữ liệu khác 41
2.6.1 Các classifier k-láng giềng gần nhất 42
2.6.2 Lập luận dựa trên tình huống 42
2.7 Vấn đề đánh giá độ chính xác của phương pháp phân lớp dữ liệu 43
2.7.1 Đánh giá độ chính xác classifier 44
2.7.2 Gia tăng độ chính xác classifier 45
2.7.3 Độ chính xác có đủ để đánh giá một classifier hay không? 46
CHƯƠNG 3 KẾT QUẢ THỬ NGHIỆM 47
3.1 Giới thiệu bài toán phân lớp dữ liệu protein SUMO hóa (SUMOylation) 48
3.1.1 Giới thiệu về protein SUMO hóa (SUMOylation) 48
3.1.2 Thu thập và tiền xử lý dữ liệu 48
3.1.3 Trích chọn đặc trưng và mã hóa dữ liệu 53
3.2 Giới thiệu về phân lớp dữ liệu sử dụng công cụ Weka 55
3.2.1 Thuật toán Hồi quy logistic (Logistic Regression) 56
3.2.2 Thuật toán Naive Bayes 58
3.2.3 Thuật toán Cây quyết định (Decision Tree) 60

3.2.4 Thuật toán k-Nearest Neighbors 63
3.2.5 Thuật toán Máy hỗ trợ Vector (Support Vector Machines) 65
3.3 Kết quả phân lớp dữ liệu vị trí protein SUMOylation 68
KẾT LUẬN 70
TÀI LIỆU THAM KHẢO 71
Tiếng Việt: 71
Tiếng Anh: 71

DANH MỤC CÁC TỪ VIẾT TẮT

TT	Từ viết tắt	Tên đầy đủ	Chú thích
1.	SUMO	Small Ubiquitin-like MOdifier	Thành phần sửa đổi tương tựa như một Ubiquitin nhỏ
2.	KDD	Knowlegde Discovery in Databases	Phát hiện tri thức
3.	SVM	Support Vector Machine	Máy hỗ trợ vector
4.	AAC	Amino Axit Composition	Đặc trưng: AAC
5.	AAPC	Amino Axit Pairwise Composition	Đặc trưng: AAPC
6.	TP	True Positive	Đúng là dữ liệu Positive
7.	FP	False Positive	Không phải dữ liệu Positive
8.	TN	True Negative	Đúng là dữ liệu Negative
9.	FN	False Negative	Không phải dữ liệu Negative
10.	SEN	Sensitivity: SEN=TP/(TP+FN)	Tỷ lệ dự đoán đúng dữ liệu Positive
11.	SPE	Specificity: SPE=TN/(TN+FP)	Tỷ lệ dự đoán đúng dữ liệu Negative
12.	ACC	Accuracy	Độ chính xác
13.	MCC	Mathews Correlation Coefficient	Hệ số tương quan Mathews

14.	SUMOylated protein	Protein mà trong đó có ít nhất một vị trí đã SUMO hóa
15.	SUMO-sites Lysine	1 vị trí amino axit Lysine (K) đã được xác định thực nghiệm là SUMO hóa
16.	Non-SUMO-sites Lysine	KHÔNG PHẢI là SUMO hóa

DANH MỤC CÁC BẢNG, BIỂU

Bảng 2. 1 Các bộ dữ liệu huấn luyện từ cơ sở dữ liệu khách hang AllElectronics 18
Bảng 2. 2. Dữ liệu mẫu cho lớp mua máy tính 23
Bảng 3. 1 Bảng tổng hợp dữ liệu thu thập từ các nguồn khác nhau 48
Bảng 3. 2 Bảng tổng hợp dữ liệu thu được sau khi loại bỏ dữ liệu dư thừa bởi công cụ CD-HIT 52
Bảng 3. 3. Hiệu năng của mô hình dự đoán, đánh giá bởi kiểm tra chéo 5 mặt (5-fold cross-validation) 68
Bảng 3. 4 Hiệu năng của mô hình dự đoán, đánh giá bởi dữ liệu kiểm thử độc lập 69

DANH MỤC HÌNH VẼ

Hinh 1. 1. Quá trình phát hiện tri thức.....4
Hinh 1. 2. Tập dữ liệu với 2 lớp: có và không có khả năng trả nợ.....6
Hinh 1. 3. Phân lớp được học bằng mạng nơron cho tập dữ liệu cho vay 7
Hinh 1. 4. Phân cụm tập dữ liệu cho vay vào trong 3 cụm 8
Hinh 2. 1. Xử lý phân lớp dữ liệu.....14
Hinh 2. 2. Cây quyết định cho khái niệm mua máy tính.....15
Hinh 2. 3. Thuộc tính tuổi có thông tin thu được cao nhất.....19
Hinh 2. 4. Các cấu trúc dữ liệu danh sách thuộc tính và danh sách lớp được dung trong SLIO cho dữ liệu mẫu trong bảng 2.2 23
Hinh 2. 5. a) Mạng belief Bayesian đơn giản, b) Bảng xác suất có điều kiện cho.....28
Hinh 2. 6. Mô hình Rừng ngẫu nhiên.....31
Hinh 2. 7. Một đường thẳng tuyến tính phân chia 2 lớp điểm (hình vuông và hình tròn) trong không gian hai chiều. Ranh giới quyết định chia không gian thành hai tập tùy thuộc vào dấu của hàm f (x) = <w, x> + b. 34
Hinh 2. 8. Độ rộng biên lớn nhất được tính toán bởi một SVMs tuyến tính. Khu vực giữa hai đường mảnh xác định miền biên với -1 ≤ <w, x> + b ≤ 1. Những điểm sáng hơn với chấm đen ở giữa gọi là các điểm support vectors, đó là những điểm gần biên quyết định nhất. Ở đây, có ba support vectors trên các cạnh của vùng biên (f(x) = -1 hoặc f (x)=1). 34
Hinh 2. 9. Ảnh hưởng của hằng số biên mềm C trên ranh giới quyết định.....36
Hinh 2. 10. Mức độ tác động của kernel đa thức. Kernel đa thức dẫn đến một sự phân tách tuyến tính (A). Kernel đa thức cho phép một ranh giới quyết định linh hoạt hơn (B - C). 38
Hinh 2. 11. Ảnh hưởng của số chiều Gaussian kernel (σ) cho một giá trị cố định của các hằng số biên mềm. Đối với giá trị của σ (A) lớn quyết định ranh giới là gần như tuyến tính. Khi giảm σ tính linh hoạt của ranh giới quyết định tăng (B). Giá trị σ nhỏ dẫn đến học quá (overfitting) (C) 41