Predict Genders for Vietnamese Names

Nhóm 8
Nguyễn Minh Quân - 19522082
Nguyễn Khắc Phương - 19522063
Võ Thành Phúc - 19522047

Trường Đại Học Công Nghệ Thông Tin
Đại Học Quốc Gia Thành Phố Hồ Chí Minh

Tóm tắt nội dung Vì giới tính sinh học là một trong những khía cạnh thể hiện cá nhân con người, nhiều báo cáo đã được thực hiện dựa trên phân loại giới trên tên người. Các đề xuất cho tiếng Anh và tiếng Trung ngôn ngữ là vô cùng lớn; có một vài báo cáo được thực hiện cho người Việt Nam từ trước đến nay. Trong bài báo cáo này chúng tôi sử dụng tệp dữ liệu UIT-ViNames với hơn 26000 cái tên đã được gán nhãn giới tính của các tác giả: Huy Quoc To và các tác giả. Trong bài báo cáo thực nghiệm năm mô hình học máy bao gồm Naive Bayes, Support Vector Machines kết hợp với GridSearchCV, Logistic Regression, Decision Tree,K-nearest Neighbors. Sau khi thực nghiệm với các mô hình thì việc sử dụng Logistic Regression với RandomizedSearchCV đạt được kết quả rất tốt lên tới 95.41%,sau đó là Support Vector Machine kết hợp với GridSearchCV nên tới 95.34% .

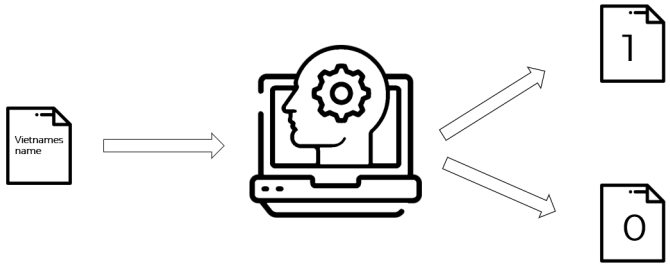
Keywords: Classifier · machine learning · GridSearchCV · RandomizedSearchCV.

1 MỞ ĐẦU

Bài toán phân lớp là bài toán phổ biến và tự nhiên nhất trong cuộc sống hằng ngày. Chúng ta thấy rằng, từ khi con người xuất hiện và sinh sống thì việc phân loại các sự vật, hiện tượng đã luôn là một trong những nhu cầu phổ biến và cấp bách. Nhưng không phải ai cũng có thể phân loại và lặp đi lặp lại một hành động, nên nhu cầu giải quyết bài toán phân loại ngày càng nhiều và bài toán ngày càng phức tạp hơn. Và việc tự động phân loại giới tính sinh học của con người qua tên cũng ngày càng được nhiều người quan tâm nghiên cứu và công bố báo cáo.
Đã có rất nhiều bài báo cáo nghiên cứu, ứng dụng các mô hình phân lớp vào việc xác định giới tính sinh học. Đằng sau những nghiên cứu này là lợi ích tiềm năng của việc dự đoán hai giới tính sinh học. Hiện tại có rất nhiều ứng dụng thực tế việc dự đoán giới tính sinh học của con người. Và trong bài báo cáo này sẽ là tập trung vào việc phân tích, đưa ra các mô hình phân loại phù hợp với bài toán phân loại giới tính sinh học của người Việt Nam.

2 GIỚI THIỆU BÀI TOÁN

Bài toán phân loại giới tính sinh học dựa vào tên người Việt Nam thuộc bài toán phân lớp nhị phân (dự đoán 2 lớp): Với Dữ liệu đầu vào là họ và tên người Việt Nam. Đầu ra là nhãn của họ và tên gồm 2 nhãn là 0 (nữ) và 1 (nam).



Hình 1. Mô tả bài toán

Kinh nghiệm của bài toán dựa trên tệp dữ liệu với thuộc tính Full_Names đã được gán nhãn Genders sẵn.
Với bài toán liên quan là Gender Prediction Based on Vietnamese Names with Machine Learning Techniques của tác giả Quoc Huy To và các đồng tác giả. Bài báo cáo của chúng tôi nhằm hướng đến việc phân tích lại một lần nữa về các lỗi mà các mô hình còn mắc phải, đồng thời tinh chỉnh siêu tham số các mô hình truyền thống và so sánh với kết quả của các mô hình truyền thống có trong bài báo cáo của tác giả Quoc Huy To và các đồng tác giả.

3 TỆP DỮ LIỆU

Tệp dữ liệu UIT-ViNames bao gồm 26.850 tên, trong đó tỷ lệ của hai nhãn là nam và nữ lần lượt là 57,71% và 42,29%. Tỉ lệ nam,nữ không cân bằng, ảnh hưởng đáng kể đến việc dự đoán. Tên người thường có ba phần là [Họ] - [Tên Đệm] - [Tên Chính]. Trong bối cảnh dùng họ và tên để dự đoán giới tính thì nhân tố [Họ] không ảnh hưởng quá lớn từ giới tính vì [Họ] được đặt theo họ của bố(cha), nên cả nam và nữ đều có thể mang cùng họ. Có những nghiên cứu chỉ ra rằng có tới 38.4% người mang họ Nguyễn, chiếm tới 4/10 trên tổng số gần 1000 Họ tại Việt Nam.
Tới hiện tại việc đặt tên người vẫn ảnh hưởng khá nhiều từ thời phong kiến "Văn" cho nam "Thị" cho nữ,việc này vẫn được áp dụng rộng rãi cho "tên đệm",

Bảng 1. Tên chính có 2 giới tính.

	Male	Female
Name	Quỳnh Văn Quỳnh	Thị Quỳnh
	Văn Văn Văn	Thị Văn

ngoài ra còn những tên đệm vừa dùng cho cả nam và nữ là "Ngọc Thủy","Ngọc Anh".Tên riêng ảnh hưởng rất nhiều từ giới tính,có rất nhiều ví dụ chỉ ra rằng một số tên riêng có thể sử dụng cho cả nam và nữ "Anh" - "Tuấn Anh" - "Lan Anh"; "Thủy" - "Ngọc Thủy" - "Kim Thủy", có thể thấy đối với những tên có thể dùng cho cả nam và nữ thì tên đệm đóng góp rất nhiều vào việc dự đoán giới tính.

4 MÔ HÌNH

4.1 Multinomial NB

Mô hình phân loại Naive Bayes phù hợp để phân loại với các tính năng rời rạc (ví dụ: số lượng từ để phân loại văn bản). Phân phối đa thức thường yêu cầu số tính năng số nguyên. Ví dụ thực tế về bài toán: “Nguyễn Thị Vân” và “Nguyễn Văn Vân” đều có họ và tên giống nhau, nhưng lại khác tên đệm là “Văn” thường được sử dụng cho nam giới và “Thị” lại dùng cho nữ giới.Vì vậy, mô hình NB là một lựa chọn phù hợp ứng dụng cho bài toán này.

4.2 Support Vector Machine

SVM là một trong những thuật toán máy học có Giám sát phổ biến nhất, được sử dụng cho các bài toán Phân loại. Mục tiêu của thuật toán SVM là tạo đường hoặc ranh giới quyết định tốt nhất có thể phân tách không gian n chiều thành các lớp để chúng ta có thể dễ dàng đặc điểm dữ liệu mới vào đúng danh mục trong tương lai. Ranh giới quyết định tốt nhất này được gọi là siêu phẳng.Vì vậy trong bài toán này SVM sẽ đưa ra kết quả dự đoán tốt hơn.

4.3 K-nearest Neighbors

KNN là một trong những thuật toán supervised-learning đơn giản nhất trong machine learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training, mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. Trong mô hình này, nhãn của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set và nhãn của tập test được quyết định bằng các h bầu chọn theo số phiếu giữa các điểm gần nhất.

4.4 Logistic Regression

Là một trong những thuật toán Học máy đơn giản và được sử dụng phổ biến nhất để phân loại hai lớp và để gán các đối tượng cho một tập hợp giá trị rời rạc. Có thể được sử dụng làm cơ sở cho bất kỳ vấn đề phân loại nhị phân nào. Ở đây áp dụng cho nhiệm vụ phân loại tên người Việt Nam vì cách sử dụng của nó là diễn đạt mối quan hệ giữa các biến nhị phân phụ thuộc và các biến độc lập.

4.5 Decision Tree

Cây quyết định là công cụ mạnh mẽ và phổ biến nhất để phân loại và dự đoán. Cây quyết định là một lưu đồ giống như cấu trúc cây, trong đó mỗi nút bên trong biểu thị một phép thử trên một thuộc tính, mỗi nhánh biểu thị một kết quả của phép thử và mỗi nút lá (nút đầu cuối) chứa một nhãn lớp.Trong bài toán, mỗi từ riêng lẻ trong mỗi cái tên sẽ ảnh hưởng quan trọng đến giới tính.

5 QUÁ TRÌNH THỰC HIỆN

5.1 CHUẨN BỊ DỮ LIỆU

Dữ liệu được xử lý chính tả và đưa về chữ thường. Trong bài báo cáo này chúng tôi sử dụng File_Full chia thành 7/1/2 lần lượt là 3 tệp File_Train dùng để huấn luyện mô hình, File_Dev sử dụng để phát triển mô hình và File_Test dùng để kiểm tra dự đoán.

Bảng 2. Ví dụ về dữ liệu

No.	Full_Names	Genders	Fa_Names	Mid_Names	Last_Names
1	Ngô Xuân Tùng	1	Ngô	Xuân	Tùng
2	Nguyễn Thị Hồng Diệp	0	Nguyễn	Thị Hồng	Diệp

Nhãn giới tính nam được quy định là 1 và nữ là 0 [Bảng 2].Dữ liệu các tệp sẽ được vector hóa và mã hóa bởi CountVector.
Để phục vụ việc ghép các trường hợp trong bộ 3 [Họ][Tên Đệm][Tên Chính] thuộc tính Full_Names sẽ được cắt thành 3 thuộc tính Fa_Names, Mid_Names,First_Names và cùng vector hóa và mã hóa bởi CountVector.

5.2 THỬ NGHIỆM MÔ HÌNH

Sử dụng 5 mô hình Multinomial NB, Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Decision Tree.
Các mô hình sẽ được cài đặt và được tinh chỉnh các siêu thông số bởi iGridSearchC và Randomized-SearchCV. Cấu hình đường cơ sở cho CountVector.Sau những lần thử nghiệm mô hình, kết quả F1 sẽ được tính và ghi lại, đồng thời cũng lưu lại các ma trận dự đoán của từng mô hình.

5.3 PHÂN TÍCH KẾT QUẢ

Sau khi hoàn thành huấn luyện và tinh chỉnh siêu tham số, chúng tôi thu thập kết quả của các mô hình. Nhận thấy kết quả của mô hình Logistic Regression đạt hiệu suất cao nhất[Bảng 3], không những thế Logistic Regression còn có điểm F1[Bảng 4] tốt nhất là 95.53%. Bên cạnh đó mô hình Support Vector Machine cũng có hiệu suất gần như tương đồng với mô hình Logistic Regression.

Bảng 3. Điểm F1 của các mô hình.

Model	Male	Female	Avg
Multinomial NB	95.59	94.06	94.82
Support Vector Machine	96.07	94.61	95.34
Logistic Regression	**96.15**	**94.68**	**95.53**
K-Nearest Neighbors	93.38	90.67	92.25
Decision Tree	94.62	92.47	93.72

Bảng 4. Hiệu suất của các mô hình trên tập dữ liệu UIT-ViName.

Model	Precision	Recall	F1-score
Multinomial NB	94.82	94.82	94.82
Support Vector Machine	95.49	95.21	95.34
Logistic Regression	**95.61**	**95.25**	**95.41**
K-Nearest Neighbors	92.40	91.74	92.02
Decision Tree	93.87	93.30	93.55

Bảng 5. So sánh với mô hình bài toán liên quan

Model	Trước	Mới
Logistic Regression	**95.14**	**95.41**
Support Vector Machine	95.28	95.34

Có thể thấy mô hình truyền thống Logistic Regression và Support Vector Machine của chúng tôi có kết quả cao hơn so với mô hình của bài toán liên quan[Hình 5], với siêu tham số của từng mô hình là Logistic Regression [C= 1.5103373646043785,penalty='l2'], Support Vector Machine [C=100.1, gamma=0.01, kernel= 'rbf'].

5.4 PHÂN TÍCH LỖI

Các mô hình đều có những dự đoán không chính xác, đa phần các dự đoán sai xuất phát từ nhãn của tệp train và test. .

Bảng 6. Lỗi do tệp kiểm tra dự đoán.

Full_Names	Predicted	True	Actual
H Joan Hwing	1	0	#NA
Đàng Thị Kim Oanh	0	1	0
Nguyễn Thị Tuyến	0	1	0
Võ Tiến Dũng	1	0	1

Trong đó có rất nhiều nhãn sai xuất phát từ train và test. Tiêu biểu là "H Joan Hwing" tên không thuần Việt xuất phát từ tệp test, "Đàng Thị Kim Oanh" tên này trên thực tế là nữ nhưng tệp test và train lại gán nhãn nam và mô hình dự đoán đúng với thực tế [Bảng 6]. Không chỉ những trường hợp trên mà còn rất nhiều trường hợp khác. Những lỗi xuất phát từ tệp test và train khiến cho việc dự đoán giảm hiệu quả.

Bảng 7. Dự đoán sai với tên không có tên đệm.

Full_Names	Predicted	True
Trần Hiếu	1	0
Lê Vy	1	0
Quát	1	0

Các mô hình còn dự đoán sai với các tên chỉ có [Họ] và [Tên chính] hoặc [tên] mà không có [tên đệm], hầu hết các dự đoán đều đưa về 1 (Nam) [Bảng 7]. Từ việc dự đoán sai này chúng ta có thêm bằng chứng để khẳng định [têm đệm] ảnh hưởng rất nhiều đến dự đoán giới tính tên Việt Nam.

Bảng 8. Dự đoán sai.

Full_Names	**Predicted**	**True**
Đoàn Hương Quân	0	1
Bạch Hồng Thái	0	1
Trần Nhã Hoài An	0	1
Trương Xuân Nguyên	1	0
Nguyễn Minh Châu	1	0
Nguyễn Thục Anh	1	0

Vẫn còn nhiều dự đoán sai khi xuất hiện những tên đệm thuộc về nữ "Hương", "Hồng", "Nhã", " Hoài" và ngược lại "Thục", "Minh"[Bảng 8] thường xuất hiện giới tính nam chứ không phải là nữ.
Ngoài những lỗi trên còn nhiều thứ ảnh hưởng đến kết quả và hiệu suất của các mô hình có thể là siêu tham số hay tệp dữ liệu chưa được xử lý và khai thác một cách hiệu quả.

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Dự đoán giới tính sinh học qua tên người Việt Nam là một trong những bài toán phổ biến và có tính ứng dụng cao, trong bài báo này, chúng tôi đã trình bày một số cách tiếp cận cho nhiệm vụ phân loại giới tính dựa trên tên tiếng Việt. Sau khi thử nghiệm trên 5 mô hình phân lớp học máy truyền thống. Mô hình học phân lớp Logistic Regression đạt kết quả khá cao trên tập dữ liệu UIT-ViNames, tỷ lệ này lên tới 95.41%. Qua những bảng và phân tích trên cũng khẳng định được phần nào đó là [Tên đệm] ảnh hưởng lớn đến dự đoán giới tính.
Trong tương lai chúng tôi sẽ tập chung chỉnh sửa một số lỗi sai đến từ tệp dữ liệu về đúng với thực tế. Đồng thời cố gắng tìm ra siêu tham số tối ưu hóa nhất cho bài toán, bằng cách mở rộng từ điển siêu tham số củaGridSearchC và RandomizedSearchCV.

Tài liệu

1. Pitsillides91.: Introduction to ML - Binary Logistic Regression Example for Beginners (2020)
2. Huynh Chi Trung.: Giới thiệu về Support Vector Machine (SVM) (2020)
3. Jason Brownlee.: Tune Hyperparameters for Classification Machine Learning Algorithms (2019) http://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
4. LNCS Homepage. http://www.springer.com/lncs. Last accessed 4 Oct 2017
5. hoavanhoc-ngonngu. http://www.khoavanhoc-ngonngu.edu.vn/nghien-cuu/ngon-ngu-hoc/5953-tai-sao-o-viet-nam,-so-nguoi-mang-ho-nguyen-nhieu-nhat.html
6. Huy Quoc To: Gender Prediction Based on Vietnamese Names with Machine Learning Techniques.https://sci-hub.se/10.1145/3443279.3443309
7. Data 360: Logistic Regression Loss Function – Hyper Parameter Tuning Evaluation Metrics – Part 3 (2020) https://www.youtube.com/watch?v=0HDy6n3UD5Mt=2303s