Phân Loại Ảnh Một Số Diễn Viên Nổi Tiếng Và So Sánh Các Phương Pháp Máy Học

Nguyễn Thanh Thiện Quá  
 *Đại Học Công Nghệ Thông Tin- ĐHQG-HCM-KHDL2020*   
20521783@gm.uit.edu.vn  
Nguyễn Hiếu Nghĩa  
 *Đại Học Công Nghệ Thông Tin- ĐHQG-HCM-KHDL2020*   
20521654@gm.uit.edu.vn

*Sơ lược*—Bài báo cáo này cung cấp thông tin về phân loại hình ảnh một số diễn viên nổi tiếng dùng các mô hình máy học khác nhau và sau đó so sánh giữa các mô hình máy học với nhau. Trong dự án này, chúng tôi dùng các mô hình SUPPORT VECTOR MACHINE (SVM), RANDOM FOREST (RF) và LOGISTIC REGRESSION (LR).

Các diễn viên nổi tiếng bao gồm 5 người: Angelina Jolie, Chris Evans, Dwayne Johnson, Emma Waston, Will Smith (ba nam và hai nữ).

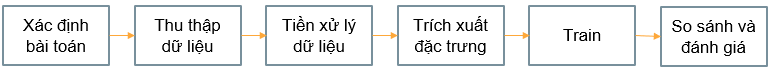
# GIỚI THIỆU

Trong dự án này, chúng tôi dùng các mô hình máy học đơn giản bao gồm SUPPORT VECTOR MACHINE (SVM)[1], RANDOM FOREST (RF)[2] và LOGISTIC REGRESSION (LR)[3], và so sánh hiệu quả giữa các mô hình để hiểu thêm về sự khác biệt giữa các mô hình trong bài toán phân loại hình ảnh nói riêng và bài toán phân loại nói chung. Chúng tôi cũng dùng GridSearchCV để tinh chỉnh từng mô hình để tìm ra kết quả tốt nhất.

Dự án được thực thi hoàn toàn bằng Python. Bài toán phân loại hình ảnh không còn mới ở thời điểm hiện tại, nhưng nó rất phù hợp cho những người mới tìm hiểu về machine learning và các thư viện của python. Bên cạnh đó, bài toán còn có tính ứng dụng cao trong thực tế, có thể kết hợp bài toán phân loại hình ảnh với các bài máy học khác để phục vụ bài toán cao hơn là bài toán phân loại vật thể. Hệ thống phát hiện tội phạm, camera được đặt ở một số địa điểm công cộng như: siêu thi, trạm xe, sân bay, … Khi phát hiện sự xuất hiện của các đối tượng được phân loại là tôi phạm, hệ thống sẽ gởi thông tin về trung tâm xử lý. Các hệ thống bảo mật khuôn mặt,…

A collage of people

Description automatically generated with medium confidenceĐầu vào của bài toán là một tấm ảnh chứa các khuôn mặt con người (có thể là 1 trong 5 diễn viên trên hoặc cũng có thể là bất kì một ai). Đầu ra của bài toán là tỷ lệ phần trăm tỷ lệ giống 1 trong các diễn viên trên (nếu không nhận dạng được thì kết quả sẽ là “không thể phân loại”)

Qui trình thực hiện dự án:

Sơ đồ 1: Qui trình thực hiện dự án

# DỮ LIỆU

## Thu Thập Dữ Liệu

Chúng tôi đã tìm hiểu một số cách thu thập như sau:

Đầu tiên: Thu thập thủ công, lần lượt tải từng hình ảnh từ Google Hình ảnh về. Cách này có ưu điểm khi chúng tôi có thể kiểm soát và chọn lọc từng điểm dữ liệu nhưng có nhiều bất lợi khi phải tốn nhiều thời gian và công sức để thu thập.

Thứ hai: Dùng python và web scrapping. Cách này có ưu điểm là nhanh chóng thu thập được dữ liệu nhưng không phải lúc nào đoạn script[4] cũng chạy được và cần am hiểu kiến thức về web.

Thứ ba: Dùng extension của Google. Cách này nhanh chóng và dễ dàng có được dữ liệu mà chúng tôi cần. Vì những thuận lợi và bất lợi của các cách thu thập trên nên chúng tôi quyết định dùng cách thứ 3 và extention mà chúng tôi sử dụng là Download All Images[5].

## Tiền Xử Lý Dữ Liệu

Do hình ảnh được thu thập hoàn toàn trên Google Hình ảnh nên không có độ tin cậy, cùng một từ khóa tìm kiếm nhưng có thể ra những hình ảnh khác nhau không liên quan. Để giải quyết vấn đề này, chúng tôi tiến hành lọc dữ liệu không liên quan bằng thủ công nhằm tăng độ tin cậy của bộ dữ liệu.

Khi chúng tôi download ảnh từ internet, những tấm ảnh đó có thể có rất nhiều vấn đề. Bây giờ, khi chúng tôi muốn phát hiện một người từ tấm ảnh đó, nếu phát hiện được tay và chân chúng ta có thể nhận diện được người đó nhưng quyết định cuối cùng để nhận biết được một người chủ yếu tập trung ở khuôn mặt. Vì thế, chúng tôi sẽ áp dụng việc cố gắng phát hiện khuôn mặt trên mỗi hình ảnh, nhưng đôi khi khuôn mặt có thể bị che khuất. Để giải quyết vấn đề này, khi nhìn vào 1 bức ảnh, chúng tôi tìm kiếm khuôn mặt với 2 mắt rõ ràng và sẽ giữ lại những hình ảnh đó, ngược lại sẽ xóa nó đi. Để phát hiện khuôn mặt với rõ ràng 2 mắt, chúng tôi dùng OpenCV, đây là thư viện xử lý ảnh nổi tiếng của python, và kĩ thuật được gọi là Haar Cascades[6].

Vì những điều kiện trên khi xử lý dữ liệu, bộ dữ liệu hiện tại của chúng tôi được chia thành 3 nhóm chính được thể hiện ở bảng bên dưới và các mô hình chỉ làm việc với những hình ảnh khuôn mặt đã được cắt ra từ ảnh gốc.

Bảng 1: Các nhóm dữ liệu

Nhóm 1-Những hình ảnh chỉ có 1 khuôn mặt với 2 mắt rõ ràng. Dữ liệu thuộc nhóm này rất dễ xử lý và phát hiện khuôn mặt.

A collage of a person

Description automatically generated with medium confidence

Hình 1: Quá trình xử lý hình ảnh thuộc nhóm 1

Nhóm 2-Là những hình ảnh có từ 2 khuôn mặt khác nhau trở lên. Cần phát hiện tất cả các khuôn mặt có trong hình, sau đó chỉ giữ lại những hình là khuôn mặt của các diễn viên đã nêu ở trên và xóa tất cả các ảnh còn lại để tránh ảnh hưởng đến kết quả mô hình sau khi xây dựng. Tất cả các quá trình xóa khuôn A picture containing clothing

Description automatically generatedmặt này đều được thực hiện thủ công.

Hình 2: Qui trình xử lý hình ảnh thuộc nhóm 2

A picture containing text

Description automatically generatedNhóm 3-Là nhóm chứa các hình ảnh không phát hiện được khuôn mặt hoặc khuôn mặt bị khuất hay chỉ nhìn thấy góc nghiên. Dữ liệu thuộc nhóm này không dùng để xây dựng mô hình máy học.

Hình 3: Qui trình xử lý hình ảnh thuộc nhóm 3

Graphical user interface, table

Description automatically generatedChart, bar chart

Description automatically generatedTổng dữ liệu thô ban đầu gồm 1436 hình ảnh cả 5 diễn viên, mỗi diễn viên có số lượng ảnh không đều nhau. Trãi qua các bước tiền xử lý, tổng dữ liệu sạch còn lại 668 hình ảnh, với lượng dữ liệu còn lại này không nhiều (chỉ gần bằng một nữa so với ban đầu) vẫn không thể chia đều cho mỗi diễn viên, sự chênh lệch này khá cao lên đến 172 ảnh giữa số lượng diễn viên cao nhất và thấp nhất. Và đã loại bỏ 768 ảnh không phù hợp với điều kiện đã nêu ở trên. Sự phân bố dữ liệu được thể hiện ở biểu đồ phía dưới.

Biểu đồ 1: Sự phân bố dữ liệu trước và sau khi tiền xử lý

## Trích Xuất Đặc Trưng

Trên khuôn mặt mỗi người có rất nhiều đặc trưng để có thể nhận diện người đó, những đặc trưng cụ thể như vầng trán, mắt, sống mũi, gò má, đôi môi, cằm và có thể là có thêm râu. Để trích xuất các đặc trưng quan trọng trên khuôn mặt mỗi người trên mỗi hình ảnh là một việc khá phức tạp đối với máy học. Có một số cách có thể trích xuất các đặc trưng đó như fourier transform[7], nhưng đối với một số tài liệu xử lý hình ảnh[8], cách hiệu quả nhất để trích xuất các đặc trưng quan trọng là Wavelet Transform[9], wavelet transform được xử dụng rộng rãi trong máy học như là một kĩ thuật xử lý hình ảnh có tác dụng phát hiện vật thể và phân loại, wavelet cũng được áp dụng một phần trong phân tích hình ảnh và phân tích kết cấu vật thể và phân loại hình ảnh[10], rất nhiều dự án đã áp dụng phương pháp này, cụ thể [10,11]. Để sử dụng wavelet transform, chúng tôi sử dựng thư viện pywt (python wavelet transform) của python, nó sẽ trả về 1 tấm ảnh đã được chuyển đổi dễ dàng trích xuất đặc trưng.

Trong dự án này, bằng cách dùng wavelet transform cho phép chúng tôi trích xuất các đặc trưng quan trọng từ hình ảnh, khi nhìn vào hình ảnh đã chuyển đổi (hình 5), ta có thể nhận thấy được có rất nhiều vùng khác biệt trên khuôn mặt, nó gần giống như một ảnh trắng-đen cung cấp các thông tin chi tiết quan trọng để nhận diện khuôn mặt người đó.

Thuật toán wavelet transform dùng để phân loại hình ảnh khuôn mặt được minh họa trong sơ đồ 2. Nó chia thành năm bước thực thi lần lượt. Khi có 1 ảnh góc -nó là ảnh với đầy đủ màu sắc RGB (Red Green Blue), và có thể có nhiều vật thể, nó trở thành một nhiệm vụ thật sự khó khăn để xác định khuôn mặt và phân loại. Vì thế bước đầu tiên là chuyển đổi màu sắc hình ảnh sang thang màu xám (greyscale)-hình 4. Bước thứ 2 là resize hình ảnh sang kích thức là 32x32 px nhằm đảm bảo rằng tất cả hình ảnh để xây dựng mô hình đều có kích thức bằng nhau. Trên cơ sở này để tiếp tục bước 3, tạo hệ số wavelet được thực hiện bằng cách dùng 2D CWT[12], ở level 5 trên tất cả các hình ảnh. Ở bước tiếp theo, chúng tôi tiến hành ghép ảnh góc ban đầu đã được resize có kích thước 32x32x3 (ảnh có kích 32x32 px và có độ màu RGB) và ảnh đã qua wavelet transform theo chiều dọc có kích thước là 4069 (32x32x3+32x32) -Hình 6. Chúng tối làm như vậy vì ảnh góc ban đầu có thể chứa một số thông tin nhất định về tấm ảnh cần phần loại, ảnh ghép vào theo chiều dọc dùng để trích xuất các thông tin quan trọng của khuôn mặt mỗi người. Và cuối cùng là mã hóa ảnh vừa ghép thành dãy số thực. Áp dụng thuật toán này trên tất cả các hình ảnh đã tiền xử lý trước đó tạo thành bộ dữ liệu sẵn sàng để đào tạo mô hình máy học, vì thế đầu vào của mô hình máy học là các hình ảnh như hình 6, sau khi hoàn thành tất cả các bước trên, chúng tôi có được một mảng 2 chiều có kích kích (668, 4096).

A picture containing text

Description automatically generatedA black and white photo of a person's face

Description automatically generated with medium confidenceA picture containing text, person, black, posing

Description automatically generatedSơ đồ 2: Các bước tiến hành trích xuất đặc trưng

Hình 4 Hình 5

Hình 6

Trong đó: Hình 4-hình ảnh đã được greyscale, hình 5-ảnh đã qua wavelet transform, hình 6-hình ghép từ anh góc với ảnh đã chuyển đổi theo chiều dọc

## Qui Trình Gắn Nhãn Và Chuẩn Hóa Dữ Liệu

Sau khi dữ liệu đã được làm sạch, chúng tôi tiến hành gắn nhãn cho bộ dữ liệu. Phương pháp gắn nhãn rất đơn giãn vì mỗi hình ảnh đã được phân loại theo từng diễn viên khi tải về, nên nhãn của mỗi tấm hình cũng chính là tên diễn viên đó đã được chuẩn hóa lại như bảng 2 bên dưới. Do đó, Tập X (tập dữ liệu) là một mảng 2 chiều chứa những đặc trưng đã xử lý ở trên, tập y (tập nhãn) chứa nhãn của những hình ảnh tương ứng tạo thành bộ dữ liệu sẵn sàng để đào tạo mô hình.

|  |  |
| --- | --- |
| Tên diễn viên | Nhãn |
| Angelina Jolie | 0 |
| Chris Evans | 1 |
| Dwayne Johnson | 2 |
| Emma Waston | 3 |
| Will Smith | 4 |

Bảng 2: Nhãn của những hình ảnh được chuẩn hóa

# CÁC PHƯƠNG PHÁP MÁY HỌC

Table

Description automatically generatedChúng tôi sẽ áp dụng một số mô hình cơ bản như SUPPORT VECTOR MACHINE (SVM), RANDOM FOREST (RF) và LOGISTIC REGRESSION (LR). Ở mỗi phương pháp, chúng tôi sẽ sử dụng các độ đo cụ thể (Bảng 3), và tìm ra các tham số và siêu tham số tốt nhất cho mô hình đó.

Chart, scatter chart

Description automatically generatedBảng 3: Các độ đo và phương pháp tính

Trong đó:

•True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)

•True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)

•False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error

•False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error

A picture containing text, clock, watch, gauge

Description automatically generatedMean squared error: cho biết khác biệt bình phương trung bình giữa các giá trị ước tính và giá trị thực tế. MSE là một hàm rủi ro , tương ứng với giá trị dự kiến ​​của tổn thất lỗi bình phương.

Trong đó: n là số điểm dữ liệu quan sát. Yi là điểm dữ liệu quan sát được dự đoán, là giá trị được dự đoán tại điểm dữ liệu i.

## Mô Hình Support Vector Machine (SVM)

1. Cơ sở lý thuyết

Support Vector Machine là một thuật toán học máy có giám sát (Supervised Learning), có khả năng thực hiện phân loại tuyến tính hoặc phi tuyến tính, hồi qui thậm chí là phát hiện ngoại vi. Đây là một trong những mô hình máy học phổ biến, đặc biệt rất thích hợp cho việc phân loại những tập dữ liệu phức tạp nhưng có kích thức vừa và nhỏ[13]. SVM có hai kiểu phân loại: phân loại nhị phân và phân loại đa lớp. Đối với phân loại nhị phân, thuật toán chỉ có thể phân loại đối tượng thuộc một trong hai lớp; yes/no, 1/0, true/false, … Phân loại đa lớp là đối tượng được phân loại một trong ba lớp hoặc nhiều hơn, một số ví dụ như phân loại văn bản tích cực, tiêu cực hay trung tính, xác định giống chó trong ảnh, và bài toán của chúng tôi thuộc phân loại đa lớp. Nói một cách đơn giản, SVM thực hiện các phép biến đổi dữ liệu phức tạp tùy thuộc vào hàm kernel mà bạn chọn, dựa trên các phép biến đổi đó, nó cố gắng tối đa hóa ranh giới phân tách giữa các điểm dữ liệu tùy thuộc vào các lớp đã xác định. Vậy SVM hoạt động như thế nào? Ở dạng cơ sở, phân lớp tuyến tính. SVM cố gắng tìm một đường tối đa hóa sự phân tách giữa tập dữ liệu 2 lớp của các điểm dữ liệu trong không gian 2 chiều. Nói một cách tổng quat, mục tiêu là tìm ra một siêu phẳng tối đa hóa việc phân tách các điểm dữ liệu thành các lớp tiềm năng của chúng trong không gian n chiều. Các điểm dữ liệu có khoảng cách nhỏ nhất đến siêu phảng (các điểm gần siêu phẳng nhất) được gọi là Support Vector. Trong hình 7 bên dưới, Support Vector là ba điểm (2 xanh dương và 1 xanh lá) nằm trên các đường phân tách và siêu phẳng phân tách là đường liền màu đỏ có phương trình w\*x-b=0[13].

Hình 7: Minh họa phân tách dữ liệu bằng đường thẳng tuyến tính

1. Kết quả thử nghiệm

Chart

Description automatically generated with medium confidenceSau quá trình đào tạo mô hình, chúng tôi thu được các quả như sau:

|  |  |
| --- | --- |
| Các độ đo và tham số | Giá trị |
| Siêu tham số C | 15 |
| Kernel function | Radial Basis Function |
| Accuracy | 0.8607 |
| Precision | 0.8844 |
| Recall | 0.8342 |
| MSE | 0.97 |

Bảng 4: Bảng kết quả thuật toán SVM Hình 8:Confusion matrix SVM

(average=macro)

Nhận xét: Điểm mạnh của thuật toán này là có khả năng xử lý trên không gian chiều cao, đặc biệt là phân loại hình ảnh hay văn bản. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn. Từ bảng 4, ta nhận thấy các độ đo đều khá cao (83% trở lên), bên cạnh đó thì giá trị MSE khá thấp, điều này chứng minh thuật toán có khả năng phân loại cao, hoạt động tốt. Khi xem xét ma trận nhầm lần từ hình 8, nhãn 0 bị phân loại sai tận 13 điểm dữ liệu là do lớp này có số lượng điểm dữ liệu trong tập kiểm thử khá cao so với các nhãn còn lại, đường chéo chính của ma trận có số lượng nhiều hơn vượt trội so với các ô còn lại với độ chính xác lên đến 86.07%, mô hình hoạt động hiệu quả trên bộ dữ liệu hiện tại.

## Mô Hình Random Forest (RF)

1. Cơ sở lý thuyết

Random Forest là thuật toán phân loại có giám sát, nó được xây dựng dựa trên khái niệm về Decision Tree (DT)[13], RF dựa trên nhiều cây quyết định tự học (tức là rừng). Ý tưởng đằng sau việc sử dụng nhiều cây quyết định (tập hợp) là nhiều Diagram

Description automatically generatedDT học cơ sở có thể dẫn đến một quyết định mạnh mẽ và chắc chắn hơn so với một DT đơn lẽ-Sơ đồ hình 9. Những cây này tự động xác định các qui tắc tại mỗi nút dựa trên tập dữ liệu huấn luyện. RF tìm cách giảm thiểu sự không đồng nhất của 2 tập con, kết quả của dữ liệu được tạo bởi qui tắc tương ứng. Hơn thê nữa, mỗi cây quyết định là ngẫu nhiên (Random). RF có 2 lớp ngẫu nhiên. Đầu tiên, nó sử dụng một mẫu ngẫu nhiên của tập dữ liệu đào tạo (Bootstrap sample) để phát triển từng cây quyết định riêng lẽ. Thành phần ngẫu nhiên thứ 2 là lựa chọn ngẫu nhiên các đặc trưng được xem xét tại mỗi nút để xác định qui tắc tốt nhất cho việc tách dữ liệu và cuối cùng là xác định nhãn. Nhiều cây có thể tạo ra các nhãn khác nhau cho cùng một điểm dữ liệu. Do đó, việc ấn định lớp cuối cùng của mỗi pixel hình ảnh dựa trên đa số phiếu bầu giữa tất cả các cây Diagram

Description automatically generatedtrong RF.

Hinh 9: Sơ đồ cấu trúc thuật toán RF

1. Kết quả thử nghiệm

|  |  |
| --- | --- |
| Các độ đo và tham số | Giá trị |
| N\_estimators | 5 |
| Accuracy | 0.5970 |
| Precision | 0.6004 |
| Recall | 0.5158 |
| MSE | 2.8 |

**Chart

Description automatically generated**Sau quá trình đào tạo, chúng tôi thu được kết quả như sau: Bảng 5: Bảng kết quả thuật toán RF (average=macro) Hình 10: Confusion matrix RF

Nhận xét: Các độ đo -bảng 5, của mô hình RF với n\_estimators = 5 và average macro cho ra kết quả khá thấp. Đường chéo chính của ma trận -hình 10, không thấy rõ ràng khi xét về phương diện màu sắc, phân loại sai khá nhiều vào các nhãn khác, số nhãn phân loại đúng không cao, độ chính xác (accuracy) không cao, MSE còn tương đối cao. Từ đó ta có thể kết luận rằng mô hình RF hoạt động hiệu quả nhưng phân loại khá yếu trên bộ dữ liệu hiện tại.

## Mô hình Logistic Regression (LR).

1. Cơ sở lý thuyết

LR [13] là mô hình phân loại có giám sát. Trong đó sử dụng các kỹ thuật của mô hình hồi quy tuyến tính trong các giai đoạn ban đầu để tính toán logits (Điểm). Vì vậy, về mặt kỹ thuật, chúng ta có thể gọi mô hình hồi quy logistic là mô hình tuyến tính. các mô hình hồi quy logistic được phân loại dựa trên số lượng các lớp mục tiêu và sử dụng các chức năng thích hợp như hàm sigmoid hoặc softmax[14] để dự đoán lớp mục tiêu. Nói ngắn gọn: Hàm Sigmoid: được sử dụng trong mô hình hồi quy logistic để phân loại nhị phân. Hàm Softmax: được sử dụng trong mô hình hồi quy logistic để đa phân loại. Nếu mô hình hồi quy logistic được sử dụng để giải quyết loại vấn đề phân loại nhị phân thì nó được gọi là bộ phân loại hồi quy logistic nhị phân (hình 11).

Hình 11: Ví dụ biểu đồ LR

Trong khi mô hình hồi quy logistic được sử dụng cho các loại vấn đề đa phân loại, nó được gọi là bộ phân loại hồi quy logistic đa thức. Hồi quy logistic đa thức cũng là một thuật toán phân loại giống như hồi quy logistic cho phân loại nhị phân. Trong khi trong hồi quy logistic cho phân loại nhị phân, nhiệm vụ phân loại là dự đoán lớp mục tiêu thuộc loại nhị phân. Như Có / KHÔNG, 0/1, Nam / Nữ. Khi nói đến hồi quy logistic đa thức. Ý tưởng là sử dụng kỹ thuật hồi quy logistic để dự đoán lớp mục tiêu (nhiều hơn 2 lớp mục tiêu) cho đến khi tính toán xác suất cho từng mục tiêu. Khi các xác suất đã được tính toán. Chúng tôi cần chuyển chúng thành one hot encoding và sử dụng phương pháp cross entropy[13] trong quá trình đào tạo để tính toán trọng lượng được tối ưu hóa thích hợp.

1. Kết quả thực nghiệm

|  |  |
| --- | --- |
| Các độ đo và tham số | Giá trị |
| Tham số C | 5 |
| Accuracy | 0.8557 |
| Precision | 0.8727 |
| Recall | 0.8419 |
| MSE | 1.0448 |

Chart

Description automatically generated Nhận xét: Dựa vào bảng 6, các độ đo đánh giá mô hình đều ở mức khá cao, thấp nhất là recall (84.19%), MSE chỉ bằng 1.0448 khá thấp, khi nhìn vào ma trận nhầm lẫn (hình 12) cho ta biết các điểm dữ liệu bị phân loại sai thấp, đường chéo của ma trận có màu sắc đậm hơn các ô còn lại, độ chính xác (accracy) là 85.57%. Từ đó ta có thể kết luận rằng mô hình phân loại mạnh, khá chính xác và hiệu quả trên bộ dữ liệu hiện tại.

Bảng 6: Bảng kết quả thuật toán LR Hình 10: Confusion matrix LR

(average=macro)

# So sánh Và Tinh Chỉnh Mô Hình

## So Sánh Các Mô Hình Máy Học

Sau quá trình đào tạo và thống kê các độ đo, cả ba mô hình đều phân loại hiệu quả, tuy nhiên khi nhìn trực quan độ lỗi (MSE) của mô hình Random Forest lớn hơn vượt trội so với hai mô hình còn lại-Hình 11b, và tương tự với các độ đo accuracy, recall, recision đều thấp hơn đáng kể-Hình 11a. Từ kết quả đó cho thấy mô hình Random Forest phân loại chưa mạnh trên bộ dữ liệu hiện tại. Hai mô hình Logistic Regression và SVM còn lại có độ lỗi (MSE) thấp và các độ đo khá cao, từ 83% trở lên. Hai mô hình LR và SVM có khả năng phân loại mạnh. Để kiểm tra sự khác biệt giữa ba mô hình này là thực sự hay do ngẫu nhiên, chúng tôi tiến hành dùng K-cross validation chia tập train thành 5 tập con bằng nhau và thu thập kết quả để xây dựng bảng ANOVA khoảng tin cậy 95%. Các giả thuyết kiểm tra được đặt ra như sau:

H0: “Không có sự khác biệt giữa ba mô hình”

H1: “Có sự khác biệt giauwx ba mô hình”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Groups | Diff | Lwr | Upr | P\_adj |
| RF-LR | -0.2677 | -0.2646 | -0.1708 | 0.00002 |
| SVM-LR | -0.0384 | -0.1353 | -0.0585 | 0.55669 |
| SVM-RF | 0.2292 | 0.1324 | 0.3261 | 0.00011 |

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedBảng 7: Kết quả phân tích ANOVA với k-cross validation

a b

Hình 11: a) Biểu đồ so sánh ba giá trị độ đo của ba mô hình. b) Biểu đồ so sánh MSE giữa ba mô hình

Kết quả-Bảng 7, cho chúng ta biết có sự khác biệt thực sự giữa mô hình RF-LR và giữa mô hình SVM-RF khi p\_value rất bé (p\_value < 0.05) vì mô hình RF hoạt động khá tệ so với hai mô hình còn lại. Không có sự khác biệt giữa hai mô hình SVM-LR (p\_value > 0.05) vì cả hai mô hình đều hoạt động hiệu quả. Tóm lại, Mô hình RF hoạt động tệ nhất, trong khi SVM và LR hoạt động rất tốt.

## Tinh Chỉnh Mô Hình

Độ chính xác của các mô hình máy học phân loại có thể được cải thiện bằng cách tỉnh chỉnh mô hình[15], quá trình này chúng tôi sử dụng Grid Sreach (GS) và K -Cross Validation (K-CV) để tìm ra các siêu tham số bị giới hạn trong phạm vi rộng thay vì một giá trị mặc định. Trong phạm vi siêu tham số được chỉ định, các tham số được điều chỉnh bằng cách lấy tổ hợp lần lượt các giá trị đưa vào. Sau đó mô hình được đào tạo với các siêu tham số đã điều chỉnh để tìm kiếm sự kết hợp tốt nhất với độ chính xác cao nhất trên tập đánh giá. GS thường được sử dụng với K-CV, có thể tạo ra chỉ số đánh giá mô hình. Khi mô hình được huấn luyện, tập dữ liệu được chia thành hai phần: Tập huấn luyện và tập kiểm tra. K-CV là phương pháp thường sử dụng cho các tập hợp huấn luyện-Hình 12. K-CV chia tập huấn luyện thành K phần có kích thước bằng nhau. Sau đó huấn luyện trên K-1 tập hợp huấn luyện và tập còn lại dùng để kiểm tra mô hình. Có thể thu được độ chính xác và hiệu suất của K mô hình sau đó đánh giá dựa trên độ chính xác trung bình của K mô hình. Tiếp theo các siêu tham số của bộ phân loại được thay đổi dựa trên GS và độ chính xác được tính toán lại. Quá trình tối ưu hóa cho GS và K-CV được trình bày trong hình 13. Kết quả cuối cùng, chúng tôi thu được các siêu tham số tốt nhất đối với từng mô hình thuật toán được trình bày trong bảng 8.

Graphical user interface, text, application

Description automatically generated

Bảng 8: Kết quả siêu tham số tìm được

Diagram

Description automatically generatedA picture containing diagram

Description automatically generated

Hình 12: Thủ tục K-CV

# PHÂN TÍCH LỖI VÀ HƯỚNG PHÁT TRIỂN

## Phân Tích Lỗi

Các mô hình hoạt động khá tốt, tuy nhiên hiệu suất của mô hình Random Forest vẫn chưa tốt như chúng tôi mong muốn, nguyên nhân có thể thuộc về bộ dữ liệu của chúng tôi. Ở bộ dữ liệu hiện tại, các điểm dữ liệu có nhãn là Angelina Jolie (nhãn 0) khá cao so với các nhãn còn lại -Biểu đồ 1, có nghĩa là bộ dữ liệu không cân bằng. Điều cần lưu ý là các thuật toán Classification thông thường thường hoạt động tốt nếu các class có lượng dữ liệu training tương đối như nhau. Nếu không, hiện tượng overfitting rất dễ xảy ra vì mô hình cố gắng ‘fit’ dữ liệu ở class trội hơn. Có một hướng tiếp cận được gọi là ‘Resampling’ để các classes có lượng dữ liệu tương đối như nhau. Cách thứ nhất là UNDER-sampling, tức chỉ chọn ra vài phần tử của class trội hơn và kết hợp với class còn lại để làm dữ liệu training. Cách thứ hai là OVER-sampling, tức có thể lặp lại dữ liệu, hoặc tìm cách kết hợp để tạo ra dữ liệu mới, của class ít hơn, và kết hợp với class còn lại để làm dữ liệu training. Chúng tôi đã thử nghiệm và nhận thấy UNDER-sampling khá hiệu quả.

## Hướng Phát Triển

Mô hình hiện tại chỉ phân loại được khuôn mặt rõ ràng, trên thực tế có rất nhiều trường hợp khuôn mặt được ghi nhận lại không rõ ràng, vì thế, đây là bước đầu tiên để tiến tới hệ thống phân loại các khuôn mặt hoàn chỉnh. Phân loại khuôn mặt chỉ là bước khởi đầu của dự án này, khuôn mặt người chỉ là một trong những đối tượng được phân loại, nhận diện, các đối tượng khác có thể được xác định, phân loại theo phương pháp tương tự. Ví dụ, hệ thống có thể phân loại các phương tiện đang di chuyển, vật dụng nội thất, động vật, … nếu mô hình được tạo ra và huấn luyện về dữ liệu với các đối tượng đó. Ngoài ra, có thể phát triển hệ thống sinh trắc học đa phương thức, kết hợp hai hoặc nhiều thông tin sinh trắc học (nhận diện khuôn mặt và giọng nói, …).

# Kết Luận

Trong báo cáo này, chúng tôi đã so sánh hiệu quả của một số phương pháp học máy với nhau trên cùng bài toán phân loại hình ảnh một số diễn viên nổi tiếng sử dụng bộ cơ sở dữ liệu hình ảnh tự thu thập từ Google của một số diễn viên Hollywood nổi tiếng. Dựa trên kết quả thực nghiệm trên tập dữ liệu này cho thấy mô hình Random Forest phân lớp với độ chính xác thấp hơn các mô hình học Support Vector Machine và Logistic Regression, vì có các điểm dữ liệu của nhãn nhiều hơn so với các nhãn còn lại. Sau quá trình so sánh, chúng tôi kết luận rằng mô hình máy học Random Forest cho kết quả phân lớp không tốt so với máy học Support Vector Machine và Logistic Regression. RF là một phương pháp kết hợp dựa trên mô hình Decision Tree, có thể nói mô hình FR phức tạp hơn LR và SVM, ta thấy được không phải mô hình phức tạp có thể tốt hơn mô hình đơn giản. Bên cạnh đó, khám phá cách phát triển hệ thống phân loại khuôn mặt từ vài mô hình cơ bản để xác định khuôn mặt từ hình ảnh. Ngoài ra, chúng tôi học được cách chuẩn bị tập dữ liệu nhận diện khuôn mặt bao gồm trích xuất khuôn mặt và trích xuất các đặc điểm khuôn mặt. Cách điều chỉnh, đánh giá một số mô hình máy học trên bộ dữ liệu sưu tầm được.

##### TÀI LIỆU THAM KHẢO

[1] F. Rustam *et al.*, “Vector mosquito image classification using novel RIFS feature selection and machine learning models for disease epidemiology,” *Saudi J. Biol. Sci.*, vol. 29, no. 1, pp. 583–594, Jan. 2022, doi: 10.1016/j.sjbs.2021.09.021.

[2] D. Liberda, K. Kosowska, P. Koziol, and T. P. Wrobel, “Spatial sampling effect on data structure and random forest classification of tissue types in High Definition and Standard Definition FT-IR imaging,” *Chemom. Intell. Lab. Syst.*, vol. 217, p. 104407, Oct. 2021, doi: 10.1016/j.chemolab.2021.104407.

[3] T. V. N. Prabhakar, G. Xavier, P. Geetha, and K. P. Soman, “Spatial Preprocessing Based Multinomial Logistic Regression for Hyperspectral Image Classification,” *Procedia Comput. Sci.*, vol. 46, pp. 1817–1826, Jan. 2015, doi: 10.1016/j.procs.2015.02.140.

[4] “Automate Downloading Web Images to any Folder,” *Computing Tidbits*. https://cects.com/automate-downloading-folder/

[5] “Download All Images - Nhận tiện ích mở rộng này cho Firefox

[6] “Face Detection using Haar Cascades — OpenCV-Python Tutorials beta documentation.” https://opencv24-python-tutorials.readthedocs.io/en/latest/py\_tutorials/py\_objdetect/py\_face\_detection/py\_face\_detection.html

[7] B. R. Jany, A. Janas, and F. Krok, “Automatic microscopic image analysis by moving window local Fourier Transform and Machine Learning,” *Micron*, vol. 130, p. 102800, Mar. 2020, doi: 10.1016/j.micron.2019.102800.

[8] H. Lu, L. Zhang, and S. Serikawa, “Maximum local energy: An effective approach for multisensor image fusion in beyond wavelet transform domain,” *Comput. Math. Appl.*, vol. 64, no. 5, pp. 996–1003, Sep. 2012, doi: 10.1016/j.camwa.2012.03.017.

[9] S. S. Tevaramani and R. J, “Image steganography performance analysis using discrete wavelet transform and alpha blending for secure communication,” *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 208–214, Jun. 2022, doi: 10.1016/j.gltp.2022.03.024.

[10] “(PDF) Feature Extraction Technique using Discrete Wavelet Transform for Image Classification.” https://www.researchgate.net/publication/4319558\_Feature\_Extraction\_Technique\_using\_Discrete\_Wavelet\_Transform\_for\_Image\_Classification

[11] “Jianying Hu and Mojsilovic - 2000 - Optimal color composition matching of images.pdf.” Available: http://www.engw.org/articles/2021/engw2021-005.pdf

[12] N. Kingsbury, “Complex Wavelets for Shift Invariant Analysis and Filtering of Signals,” *Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, May 2001, doi: 10.1006/acha.2000.0343.

[13] “2-Aurélien-Géron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow\_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O’Reilly-Media-2019.pdf.”

[14] “Difference Between Softmax Function and Sigmoid Function.” https://dataaspirant.com/difference-between-softmax-function-and-sigmoid-function/

[15] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, “Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm,” *J. Rock Mech. Geotech. Eng.*, Apr. 2022, doi: 10.1016/j.jrmge.2022.03.002.

Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| Công việc | 20521783 | 20521654 |
| Xác định yêu cầu | x | x |
| Thu thập dữ liệu | x | x |
| Chuẩn bị dữ liệu | x | x |
| Huấn luyện mô hình | x | x |
| Tinh chỉnh mô hình | x |  |
| Xây dựng demo | x |  |
| Viết báo cáo | x | x |
| Mức độ hoàn thành | 100% | 100% |

Link source code: <https://drive.google.com/file/d/1g9hYaf6Lrgx7mq6vvJgEFjU2qJPmMlbA/view?usp=sharing>