ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
Khoa Toán - Cơ - Tin học

NHẬN DIỆN KHUÔN MẶT

Giảng viên: PGS. TS. Lê Hoàng Sơn

Sinh viên thực hiện: Đỗ Thùy Linh - 20001940
Đinh Phương Linh - 20001941
Trịnh Thị Ngọc Mai - 20001948
Nguyễn Thị Phương Ngân - 20001953
Nguyễn Thị Bích Ngọc - 20001957

Ngành: Máy tính và Khoa học thông tin
(Chương trình đào tạo chuẩn)

Hà Nội, 05 - 2023

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
Khoa Toán - Cơ - Tin học

NHẬN DIỆN KHUÔN MẶT

Giảng viên: PGS. TS. Lê Hoàng Sơn

Sinh viên thực hiện: Đỗ Thùy Linh - 20001940
Đinh Phương Linh - 20001941
Trịnh Thị Ngọc Mai - 20001948
Nguyễn Thị Phương Ngân - 20001953
Nguyễn Thị Bích Ngọc - 20001957

Ngành: Máy tính và Khoa học thông tin
(Chương trình đào tạo chuẩn)

Hà Nội, 05 - 2023

Bảng phân công công việc

Tên	Lớp	Phụ trách	Chữ ký	Điểm
Đỗ Thùy Linh	K65A6 MT&KHTT	Tóm tắt và giới thiệu, hướng phát triển		
Đinh Phương Linh	K65A6 MT&KHTT	Tìm hiểu, cài đặt và trình bày Frequent pattern mining		
Trịnh Thị Ngọc Mai	K65A6 MT&KHTT	Code, viết giải thích code, thiết kế và xây dựng phần mềm nhận diện khuôn mặt		
Nguyễn Thị Phương Ngân	K65A6 MT&KHTT	Tìm hiểu, cài đặt và trình bàybày thuật toán KNN để nhận dạng khuôn mặt		
Nguyễn Thị Bích Ngọc	K65A6 MT&KHTT	Tìm hiểu, cài đặt và trình bày thuật toán di truyền		

Lời cam đoan

Chúng em xin cam đoan rằng: Toàn bộ những nội dung trình bày trong quyển báo cáo tiểu luận môn Khai phá dữ liệu này không phải là bản sao chép từ bất cứ bài tiểu luận nào có trước. Nếu không đúng sự thật, chúng em xin chịu mọi trách nhiệm trước thầy.

Lời cảm ơn

Để hoàn thành bài báo cáo tiểu luận này, đầu tiên, em xin gửi lời cảm ơn chân thành đến Trường Đại học Khoa học Tự nhiên đã đưa môn học Khai phá dữ liệu vào chương trình giảng dạy. Đặc biệt, chúng em xin trân trọng cảm ơn PGS. TS. Lê Hoàng Sơn và TS. Phạm Huy Thông, là giảng viên trực tiếp bộ môn Khai phá dữ liệu.

Mặc dù đã có những đầu tư nhất định trong quá trình làm bài song cũng khó có thể tránh khỏi những sai sót, chúng em kính mong nhận được ý kiến đóng góp của thầy để bài tiểu luận được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

Tóm tắt nội dung

Trong báo cáo tiểu luận này, chúng ta sẽ trình bày về chủ đề nhận diện gương mặt, một lĩnh vực quan trọng trong lĩnh vực trí tuệ nhân tạo và xử lý ảnh. Cụ thể hơn, báo cáo trình bày về một số thuật toán và hướng để phát triển cho đề tài trên. Nội dung báo cáo này bao gồm các phần như sau:

- Giới thiệu cái nhìn tổng quan về đề tài: Giới thiệu mô tả bài toán nhận diện khuôn mặt, các định nghĩa, tầm quan trọng và ứng dụng của nhận diện khuôn mặt, giới thiệu về Carnegie Mellon University (CMU) và đóng góp trong lĩnh vực này.
- Các thuật toán liên quan thường gặp: Trình bày 3 thuật toán liên quan đến đề tài gồm Frequent Pattern, K-Nearest Neighbors, Genetic Algorithms. Mỗi thuật toán chia thành 5 phần:
  - Giới thiệu khái quát về thuật toán
  - Cơ sở toán học
  - Chi tiết thuật toán: Phân tích thuật toán
  - Code
  - Kết quả thu được
  - F1 score: Xét về độ chính xác
  - Ưu điểm và nhược điểm
  - Nhận xét
- Bộ dữ liệu sử dụng cho đề tài: Trình bày về bộ dữ liệu Carnegie Mellon University (CMU)
- Hướng phát triển tiếp theo của đề tài.

Một số từ viết tắt

KNN	K-Nearest Neighbors
FPM	Frequent Pattern Mining
GA	Genetic Algorithm
NST	Nhiễm sắc thể

Mục lục

Danh sách bảng

Danh sách hình vẽ

Chương 1

Giới thiệu

1.1 Tổng quan về chủ đề nhận dạng khuôn mặt

Nhận dạng khuôn mặt là quá trình xác định và xác minh danh tính của một cá nhân dựa trên đặc trưng khuôn mặt của họ. Bài toán nhận diện khuôn mặt liên quan đến việc phân tích và nhận biết các đặc điểm độc nhất trên khuôn mặt của một người như hình dạng, kích thước, vị trí của mắt, mũi, miệng, và các đặc trưng khác để xác định và nhận dạng người đó. Các hệ thống nhận dạng gương mặt có thể được sử dụng để đăng nhập vào các thiết bị di động, giám sát an ninh, quản lý thời gian làm việc hoặc trong các ứng dụng đa dạng khác. Tùy thuộc vào cách thức triển khai mà hệ thống nhận dạng gương mặt có thể sử dụng các phương pháp khác nhau để xử lý dữ liệu hình ảnh và đưa ra quyết định. Nhận diện gương mặt là bài toán đã được đề cập tới từ lâu và được nghiên cứu rộng rãi trong khoảng 30 đến 40 năm trở lại đây. Một số phương pháp thường gặp trong chủ đề này bao gồm

- Phân tích đặc trưng: Hệ thống sẽ phân tích các đặc trưng của khuôn mặt như kích thước, hình dạng, khoảng cách giữa các điểm đặc trưng và các điểm khác để tạo ra một biểu diễn số học của khuôn mặt. Sau đó biểu diễn số học này sẽ được so sánh với cơ sở dữ liệu các biểu diễn số học của các khuôn mặt trước đó để xác định xem đó có phải là một người đã biết không

- Học sâu: Hệ thống sử dụng mạng notron để học cách trích xuất các đặc trưng của khuôn mặt và tạo ra một biểu diễn số học của chúng. Sau đó biểu diễn số học này được so sánh với cơ sở dữ liệu các biểu diễn số học của các khuôn mặt đã được xác minh trước đó để xác định xem đó có phải là một người đã biết hay không?

- Học máy: Hệ thống sử dụng các thuật toán học máy để phân loại các khuôn mặt dựa trên các đặc trưng của chúng. Các đặc trưng này có thể được trích xuất bằng cách sử dụng các phương pháp như phân tích các thành phần chính hoặc phân tích tuyến tính đa biến. Sau khi có được một mô hình phân loại từ các dữ liệu huấn luyện, hệ thống sẽ sử dụng mô hình này để dự đoán xem một khuôn mặt mới có phải là người đã biết hay không?

1.2 Những khó khăn của nhận dạng khuôn mặt

Tuy nhiên việc triển khai hệ thống nhận dạng gương mặt cũng đang đặt ra nhiều vấn đề liên quan đến quyền riêng tư và an ninh thông tin. Việc lưu trữ và sử dụng dữ liệu khuôn mặt của người dùng một cách an toàn và đảm bảo quyền riêng tư là một thách thức đáng kể. Do đó cần có các quy định và tiêu chuẩn về quyền riêng tư và bảo mật thông tin để đảm bảo rằng việc sử dụng hệ thống nhận dạng gương mặt được thực hiện một cách đúng đắn và an toàn

Ngoài ra hệ thống nhận dạng gương mặt có thể bị ảnh hướng bởi các yếu tố như ánh sáng, góc chụp, trang phục, phụ kiện thậm chí là sự thay đổi của khuôn mặt theo thời gian. Bên cạnh đó, bài toán nhận diện khuôn mặt cũng đang mắc phải những thách thức như hệ thống camera công cộng, chụp hình trong các hoạt động thì ảnh nhận được có thể bị che, không chính diện hay không đảm bảo chất lượng. Đây là những yếu tố ảnh hưởng đến các thuật toán nhận diện khuôn mặt. Có các thuật toán để khắc phục điều này, họ sử dụng một số kĩ thuật như xác định nhiều điểm chính trên khuôn mặt, lấy những chi tiết nhỏ hay sử dụng trong phương pháp Học Sâu. Do đó cần phải đảm bảo rằng hệ thống nhận dạng gương mặt được đào tạo và kiểm tra trên các tập dữ liệu đủ

đa dạng để đảm bảo tính đúng đắn và độ tin cậy của nó.

- Tự thế góc chụp: Chụp không chính diện... với các tư thế khác nhau, các thành phần trên khuôn mặt như mắt mũi miệng có thể bị khuất ít nhất một phần
- Sự xuất hiện hoặc thiếu một số thành phần trên khuôn mặt: Các đặc trưng như râu, mắt kính... có thể xuất hiện hoặc không
- Biểu cảm khuôn mặt: Làm ảnh hưởng đáng kể đến các thông số của khuôn mặt. Chẳng hạn cũng một khuôn mặt đó nhưng có thể khác nhau nếu người đó cười hay khóc hay sợ hãi...
- Sự che khuất
- Điều kiện của ảnh: Ảnh chụp trong các điều kiện khác nhau về chiếu sáng, tính chất camera...
- Nền ảnh phức tạp: Dễ gây nhầm lẫn với khung xung quanh
- Màu sắc

1.3 Tầm quan trọng và ứng dụng của nhận diện khuôn mặt

Nhìn chung, nhận dạng gương mặt là một lĩnh vực quan trọng trong công nghệ thông tin. Nó có nhiều ứng dụng trong đời sống hàng ngày vào trong các lĩnh vực sau:

- An ninh và an toàn: Nhận diện khuôn mặt được sử dụng trong các hệ thống an ninh để kiểm soát truy cập và đảm bảo an toàn trong các khu vực quan trọng như công ty, sân bay, ngân hàng, trường học và cơ sở quân sự.
- Quản lý danh tính, quản lí nhân sự: Công nghệ nhận diện khuôn mặt hỗ trợ quản lý danh tính trong các tổ chức, cơ quan chính phủ và hệ thống giám sát

công cộng. Nó giúp xác minh danh tính, kiểm tra chấm công và đảm bảo chỉ những người có quyền truy cập mới được phép vào các khu vực nhạy cảm.

- Giao dịch điện tử và bảo mật: Nhận diện khuôn mặt có thể thay thế mật khẩu và mã PIN trong giao dịch điện tử, tăng tính bảo mật và thuận tiện cho người dùng.

- Trải nghiệm người dùng, hệ thống giao tiếp thông minh giữa người và máy: Trong lĩnh vực công nghệ thông tin và giải trí, nhận diện khuôn mặt cung cấp trải nghiệm người dùng tốt hơn, như mở khóa điện thoại di động, chụp ảnh chính xác và tự động lưu dữ liệu với thông tin về khuôn mặt của người dùng.

- Tự động hóa: Tự động hóa các quy trình và hoạt động trong các lĩnh vực quản lí (hệ thống tìm kiếm thông tin trên ảnh, video dựa trên nội dung), y tế, giáo dục, sản xuất...

- Vai trò quan trong nghiên cứu khoa học: Dùng để nghiêm cứu các thuật toán và mô hình học máy mới giúp cải thiện độ chính xác của các phương pháp nhận diện khuôn mặt

- Dịch vụ giải trí: Trong hầu hết các máy ảnh ngày nay đều có chức năng tự động nhận diện khuôn mặt người để có thể lấy độ nét, điều chỉnh ánh sáng phù hợp với khung cảnh xung quanh. Trên một số trang web cũng đã áp dụng công nghệ tự động nhận diện mặt người và so sánh với kho dữ liệu khổng lồ để đưa ra những lời chào, dịch vụ thông minh nhất cho người dùng

- Phân tích cản xúc...

Tuy nhiên việc triển khai và sử dụng hệ thống nhận dạng khuôn mặt cần phải được đảm bảo tính đúng đắn và bảo mật thông tin, đồng thời cần được đào tạo và kiểm tra trên các tập dữ liệu đủ đa dạng để đảm bảo độ tin cậy và tính thực tiễn của nó.

1.4 Yêu cầu cho đề tài

- Dữ liệu thực nghiệm: Tập dữ liệu đủ lớn và đa dạng CMU
- Mục tiêu đề tài: Thiết kế phần mềm trên nền tảng Web nhận dạng khuôn mặt trực tuyến
- Yêu cầu:
  - Thực hiện 3 cách phân loại
  - Hiển thị độ chính xác, điểm F1 của các phương pháp
  - Thử nghiệm với bộ dữ liệu gương mặt thật

1.5 Giới thiệu về bộ dữ liệu CMU

Bộ dữ liệu CMU (Carnegie Mellon University) là một tập hợp các bộ dữ liệu được sử dụng trong nhiều nghiên cứu về trí tuê nhân tạo, học máy và thị giác máy tính. Các bộ dữ liệu này được cung cấp miễn phí cho cộng đồng nghiên cứu. Một trong bộ dữ liệu CMU nổi tiếng bao gồm dữ liệu Face Images.
Mô tả dữ liệu Face Images:

- Kiểu dữ liệu: Kiểu hình ảnh
- Dữ liệu này bao gồm 640 hình ảnh khuôn mặt đen trắng của những người được chụp với các tư thế khác nhau (thẳng, trái, phải, hướng lên), biểu cảm (bình thường, vui, buồn, tức giận), mắt (có đeo kính hay không) và thích thước
- Nguồn: Thuộc về chủ sở hữu ban đầu và nhà tài trợ: Tom Mitchell - Trường Khoa học Máy tính - Đại học Carnegie Mellontom - mitchell@cmu.edu. Được đóng góp vào ngày 24 tháng 6 năm 1999

- Đặc điểm dữ liệu: Mỗi hình ảnh có thể được đặc trưng bởi tư thế, biểu cảm, mắt và kích thước. Có 32 hình ảnh cho mỗi người chụp mọi sự kết hợp của các tính năng. Hiển thị ở đây là 4 hình ảnh đại diện.

- Thông tin liên quan khác: Để xem hình ảnh khuôn mặt, có thể sử dụng chương trình xv.

- Định dạng dữ liệu: Dữ liệu hình ảnh có thể được tìm thấy trong \faces. Thư mục này chứa 20 thư mục con, mỗi người một thư mục, được đặt tên bởi userid. Mỗi thư mục này chứa 1 số hình ảnh khuôn mặt khác nhau của cùng 1 người

  Ta sẽ quan tâm đến những hình ảnh với quy ước đặt tên sau:
  <userid> <pose> <expression> <eyes> <scales>.pgm
  Trong đó:

  – <userid> là id người dùng của người trong ảnh. Trường này có 20 giá trị: an2i, at33, boland, bpm, ch4f, cheyer, choon, danieln, glickman, karyadi, kawamura, kk49, megak, mitchell, night, phoebe, saavik, steffi, sz24 và tammo.

  – <pose> là vị trí đầu của người đó. Trường này có 4 giá trị: thẳng, trái, phải, lên.

  – <expression> là nét mặt của một người. Trường này có 4 giá trị: bình thường, vui, buồn, tức giận.

  – <eyes> là trạng thái mắt của người đó. Trường này có 2 giá trị: đeo kính hoặc không.

  – <scale> là tỷ lệ của hình ảnh. Trường này có 3 giá trị: 1, 2 và 4.

    * 1 biểu thị hình ảnh có độ phân giải đầy đủ (128 cột x 120 hàng)

    * 2 biểu thị hình ảnh có độ phân giải một nửa (64 x 60)

    * 4 biểu thị hình ảnh có độ phân giải một phần tư (32 x 30)

Nếu ta đã xem kỹ các thư mục hình ảnh, ta có thể nhận thấy rằng một số hình ảnh có hậu tố .bad thay vì hậu tố .pgm. Bởi vì 16 trong số 640 hình ảnh được chụp bị trục trặc do thiết lập máy ảnh có vấn đề; đây là những hình ảnh .bad. Một số người gặp nhiều trục trặc hơn những người khác, nhưng tất cả những người bị "khuôn mặt" phải có ít nhất 28 hình ảnh khuôn mặt đẹp (trong số 32 biến thể có thể có, tỷ lệ chiết khấu).

- Đã được sử dụng bởi T. Mitchell. Học máy, McGraw Hill, 1997.

Chương 2

Thuật toán di truyền

2.1 Giới thiệu về thuật toán di truyền

Thuật toán di truyền (genetic algorithm) là một phương pháp để giải quyết các vấn đề tối ưu hóa có hạn chế và không bị hạn chế dựa trên chọn lọc tự nhiên, quá trình thúc đẩy sự tiến hóa sinh học. Sự kết hợp của các giải pháp khác nhau được thông qua thuật toán dựa trên thuyết tiến hóa Darwin để tìm ra các giải pháp tốt nhất. Các giải pháp kém hơn sau đó được thay thế bằng con của các giải pháp tốt.

Một cá thể trong thuật toán di truyền sẽ biểu diễn một giải pháp của bài toán. Tuy nhiên, không giống trong tự nhiên là một cá thể có nhiều nhiễm sắc thể (NST), mỗi cá thể trong thuật toán di truyền chỉ có 1 NST. Do đó, khái niệm cá thể và NST trong thuật toán coi như tương đương.

Một NST được tạo thành từ nhiều gen, mỗi gen có thể có các giá trị khác nhau để quy định một tình trạng nào đó. Trong thuật toán di truyền, một gen được coi như một phần tử trong chuỗi NST.

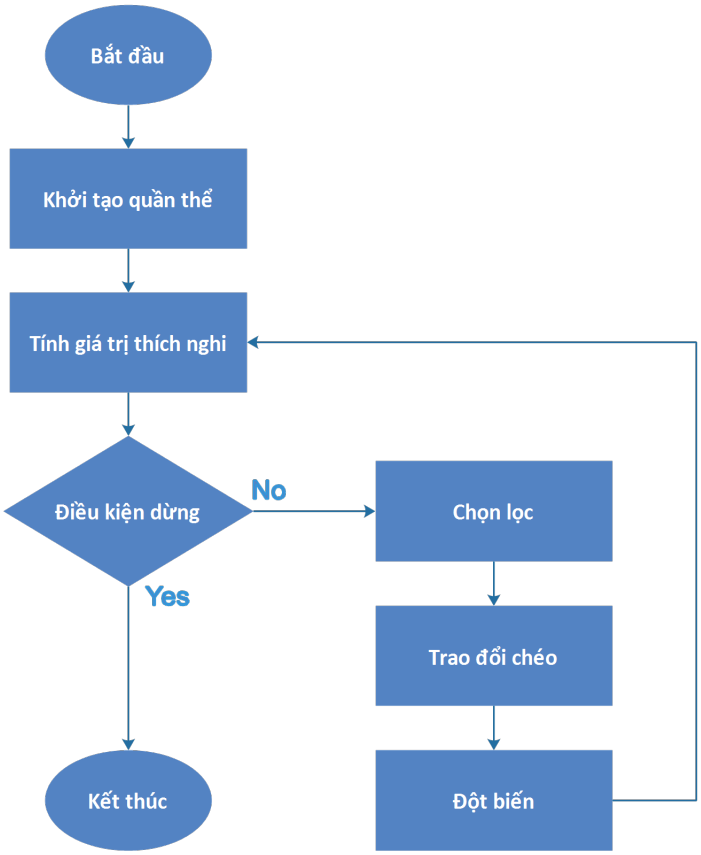
Năm vấn đề quan trọng trong thuật toán di truyền: mã hóa nhiễm sắc thể, đánh giá sự thích nghi, cơ chế chọn lọc, toán tử di truyền và các tiêu chí dừng GA.

Human Genetic	GA Terminology
chromosomes	bit strings
genes	features
allele	feature value
locus	bit position
genotype	encoded string
phenotype	decoded genotype

Bảng 2.1: Di truyền ở người và thuật toán di truyền

2.2 Triển khai thuật toán di truyền

Sơ đồ thuật toán



Hình 2.1: Sơ đồ thuật toán GA

Thuật toán được thực hiện thông qua các bước:

- Khởi tạo quần thể: Sinh ngẫu nhiên 1 quần thể gồm n cá thể (n là lời giải cho bài toán).
- Tính giá trị thích nghi: Ước lượng độ thích nghi của mỗi cá thể.
- Kiểm tra điều kiện dừng.
- Chọn lọc: Chọn hai cá thể bố mẹ từ quần thể cũ theo độ thích nghi của chúng (cá thể có độ thích nghi càng cao thì càng có nhiều khả năng được chọn).
- Trao đổi chéo: Với một xác suất được chọn, trao đổi chéo hai cá thể bố mẹ để tạo ra một cá thể mới.
- Đột biến: Với một xác suất đột biến được chọn, biến đổi cá thể mới.
- Chọn kết quá: Nếu thỏa mãn điều kiện dừng thì giải thuật kết thúc và chọn được lời giải tốt nhất trong quần thể hiện tại.

Thuật toán di truyền có hai điều kiện dừng cơ bản:

- Dựa trên cấu trúc nhiễm sắc thể, kiểm soát số gene được hội tụ. Nếu số gene được hội tụ tại 1 điểm hoặc vượt quá điểm đó thì kết thúc.
- Dựa trên ý nghĩa đặc biệt của nhiễm sắc thể, đo sự thay đổi của giải thuật sau mỗi thế hệ, nếu thay đổi này nhỏ hơn một hằng số xác định thì giải thuật kết thúc.

2.3 Ưu, nhược điểm của thuật toán di truyền

Ưu điểm

- Tốt khi dữ liệu có nhiều nhiễu.
- Thuật toán di truyền tìm kiếm trên tập các điểm mà không phải điểm riếng lẻ nên khắc phục được sự phụ thuộc vào giá trị khởi tạo.
- Thuật toán di truyền sử dụng các quy tắc chuyển đổi xác suất và không cần các quy tắc xác định.

- Có thể dễ dàng song song hóa.
- Hoạt động tốt trên các bài toán liên tục hoặc rời rạc.
- Thuật toán di truyền yêu cầu ít thông tin hơn.
- Thuật toán di truyền mang tính xác suất, phụ thuộc vào thời gian, phi tuyến tính, không cố định.

Nhược điểm

- Độ phức tạp tính toán cao.
- Thuật toán di truyền yêu cầu ít thông tin về bài toán nhưng viết và biểu diễn khó khăn.
- Thuật toán di truyền cần định nghĩa đặc biệt.

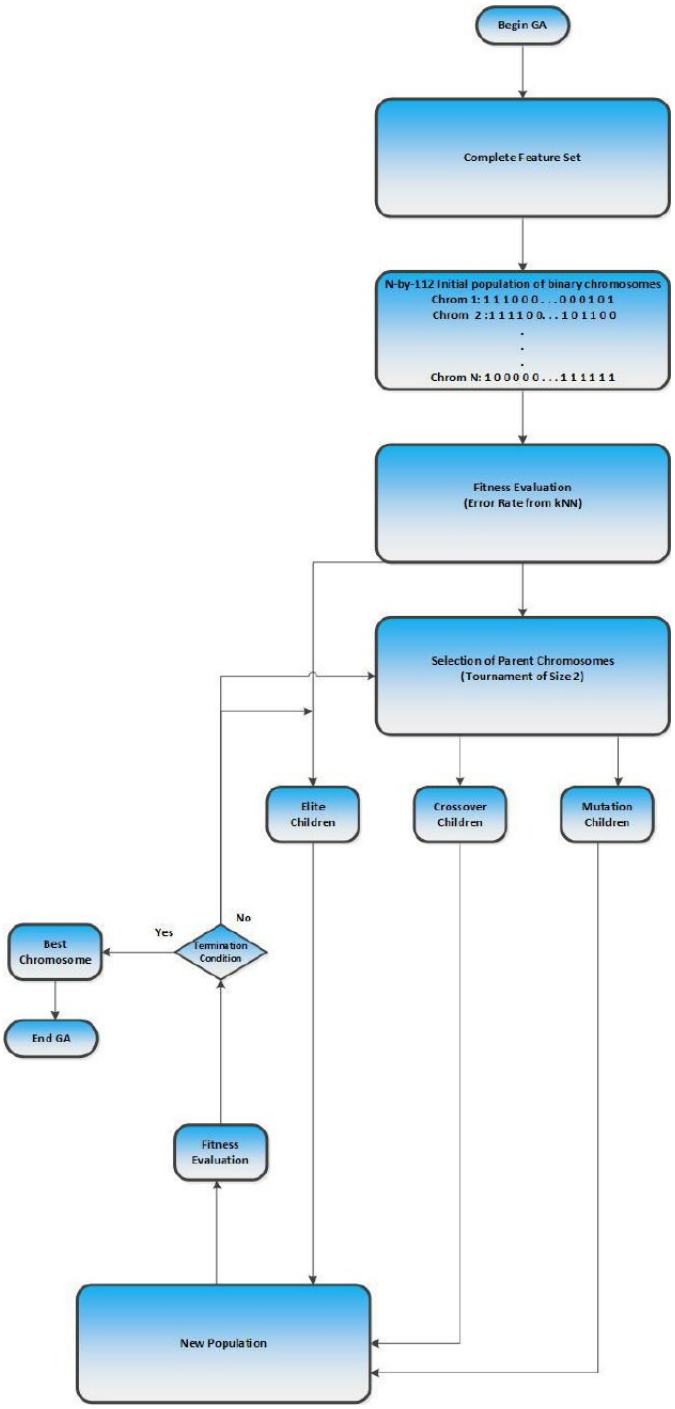
2.4 Thuật toán di truyền trong lựa chọn đặc trưng

Lựa chọn đặc trưng là quá trình chọn ra một tập con quan trọng và ý nghĩa từ tập đầy đủ các đặc trưng của dữ liệu. Mục tiêu của việc lựa chọn đặc trưng là giảm số lượng đặc trưng, giảm độ phức tạp và chi phí tính toán, đồng thời cải thiện hiệu suất của mô hình hoặc giúp hiểu sâu hơn về dữ liệu.

Thuật toán di truyền hoạt động trong không gian tìm kiếm nhị phân vì các NST là các chuỗi bit.

Để bắt đầu, một quần thể ban đầu được tạo (hầu hết là ngẫu nhiên) và được đánh giá bằng hàm fitness, đo lường mức tốt của tập con đặc trưng dựa trên hiệu suất của mô hình phân loại. Với mỗi đặc trưng sẽ có 2 trường hợp có thể xảy ra: được lựa chọn (bit 1) và không được chọn (bit 0).

Các đặc trưng được lựa chọn sau đó được xếp hạng và dựa trên bảng xếp hạng, tập con gồm n phần tử tốt nhất (đánh giá qua hàm fitness) được lựa chọn để tồn tại đến thế hệ tiếp theo. Những cá thể còn lại trong quần thể hiện tại



Hình 2.2: Lựa chọn đặc trưng với GA

được di truyền qua lai chéo (crossover) và đột biến (mutation). Tập con được lựa chọn, tập con kết quả của lai chéo và đột biến tạo thành thế hệ tiếp theo.

2.5 Áp dụng thuật toán di truyền vào nhận dạng khuôn mặt

Trong bài toán nhận dạng khuôn mặt, thuật toán di truyền được sử dụng để trích xuất đặc trưng, phục vụ cho quá trình phân loại.

Các bước thực hiện:

1. Đọc dữ liệu từ file, gán nhãn cho ảnh bằng tên tệp tương ứng với tên từng người.
2. Xử lý dữ liệu: Làm phẳng dữ liệu ảnh và chuyển dữ liệu thành Numpy array, dtype = float.
3. Áp dụng thuật toán di truyền để lựa chọn đặc trưng: sử dụng GeneticSelectionCV từ thư viện scikit-genetic.
4. Dùng mô hình phân loại Random Forest Tree để phân loại.
5. Đánh giá.

Kết quả thực hiện

Áp dụng thuật toán di truyền vào nhận dạng khuôn mặt cho kết quả với độ chính xác cao.

Accuracy	97.88%
F1 score	97.38%

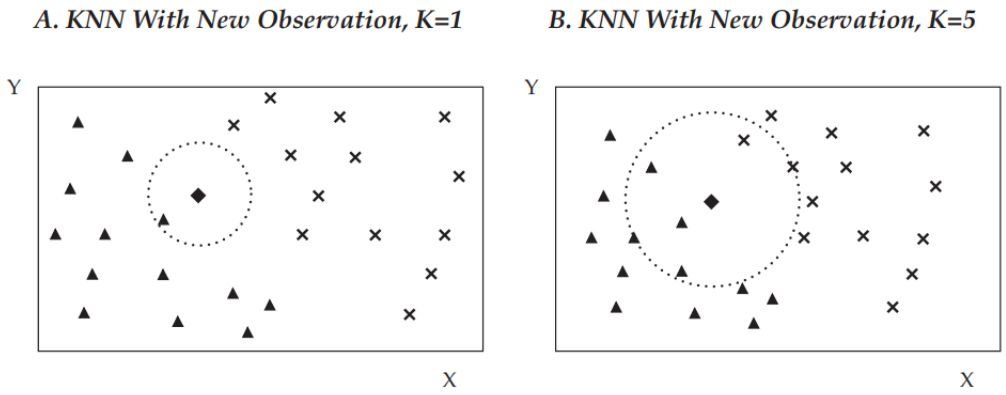
Bảng 2.2: Kết quả dùng GA

Chương 3

Thuật toán K-Nearest Neighbors

3.1 Giới thiệu

K-Nearest Neighbors (KNN) là một thuật toán học máy có giám sát, được sử dụng cho cả phân loại và hồi quy. KNN có nguồn gốc từ thực tế cuộc sống, khi mọi người thường bị ảnh hưởng bởi những người xung quanh. Nếu một người lớn lên với những người yêu thể thao, có khả năng người đó cũng sẽ yêu thể thao, mặc dù cũng có một số ngoại lệ.



Hình 3.1: Minh họa thuật toán KNN

KNN hoạt động dựa trên nguyên tắc tương tự. Giá trị của một điểm dữ

liệu được xác định bởi các điểm dữ liệu xung quanh nó. Bộ phân loại KNN xác định lớp của một điểm dữ liệu dựa trên nguyên tắc biểu quyết đa số. Ví dụ, khi K được đặt là 5, lớp của 5 điểm dữ liệu gần nhất sẽ được kiểm tra. Dự đoán kết quả lớp của điểm dữ liệu dựa trên lớp nào chiếm đa số trong 5 điểm gần nhất. Tương tự, trong hồi quy KNN, giá trị trung bình của 5 điểm gần nhất được lấy.

Vấn đề là làm thế nào để xác định các điểm dữ liệu là "gần" nhau. Để làm điều này, ta cần đo khoảng cách giữa các điểm dữ liệu. Có nhiều phương pháp để đo khoảng cách: hàm khoảng cách hình học (dành cho các bài toán có các thuộc tính đầu vào là kiểu số thực), hàm khoảng cách Hamming (dành cho các bài toán có các thuộc tính đầu vào là kiểu nhị phân), hàm tính độ tương tự Cosine (dành cho các bài toán phân lớp văn bản). Khoảng cách hình học có thể được tính theo các chuẩn Euclidean, Manhattan hoặc Minkowski như sau:

$$d_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

$$d_{\text{Euclid}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d_{\text{Minkowski}}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

trong đó, x_i và y_i lần lượt là tọa độ của điểm cần phân loại và điểm lân cận, k là số điểm lân cận được chọn. Chọn giá trị k trong thuật toán K-Nearest Neighbor là quyết định quan trọng để đạt hiệu quả tối ưu. Nếu k nhỏ, có thể gây hiện tượng mô hình không chính xác, trong khi nếu k lớn, tính toán sẽ tốn kém. Việc chọn giá trị k phù hợp phụ thuộc vào đặc điểm của bộ dữ liệu và bài toán cụ thể. Cách tiếp cận phổ biến là chọn k = $\sqrt{n}$. Ta còn có thể sử dụng cross - validation để tìm k phù hợp. Thông thường, khi làm việc với bài toán phân lớp nhị phân, ta nên để k lẻ để tránh trường hợp hoà.

KNN là một thuật toán đơn giản nhưng mạnh mẽ vì không yêu cầu quá trình huấn luyện để thực hiện dự đoán. Dưới đây là một số ứng dụng của thuật toán KNN:

- Hệ thống đề xuất: KNN được sử dụng để xây dựng các hệ thống đề xuất đơn giản như các hệ thống đề xuất phim của Amazon hoặc Netflix, hoặc các đề xuất mua hàng.

- Xếp hạng tín dụng: KNN có thể được áp dụng để đánh giá xếp hạng tín dụng của một cá nhân bằng cách so sánh thông tin tài chính của họ với những người có thông tin tương tự.

- Phát hiện chữ viết tay: KNN có thể được sử dụng để phân loại chữ viết tay, bằng cách so sánh đặc trưng của các mẫu chữ viết tay với những mẫu đã được đào tạo trước đó.

- Nhận dạng hình ảnh: KNN có thể được áp dụng trong các bài toán nhận dạng hình ảnh, trong đó các đặc trưng của hình ảnh được so sánh với các đặc trưng đã biết trước để xác định đối tượng trong hình.
  Đây chỉ là một số ví dụ về ứng dụng của KNN và thuật toán này có thể được áp dụng trong nhiều lĩnh vực khác nhau tùy thuộc vào bài toán cụ thể.

3.2 Các bước thực hiện thuật toán KNN

- Bước 1: Xác định tham số k, tức là số láng giềng gần nhất.

- Bước 2: Tính toán khoảng cách giữa đối tượng cần phân lớp và tất cả các đối tượng trong dữ liệu huấn luyện.

- Bước 3: Sắp xếp các khoảng cách theo thứ tự tăng dần và chọn k láng giềng gần nhất với đối tượng cần phân lớp.

- Bước 4: Lấy tất cả các lớp của k láng giềng gần nhất.

- Bước 5: Dựa vào phần lớn lớp của k láng giềng để xác định lớp cho đối tượng cần phân lớp.

3.3 Ưu điểm và nhược điểm của thuật toán KNN

Ưu điểm:

- Đơn giản và dễ giải thích.
- Không dựa trên bất kỳ giả định nào, cho phép áp dụng trong các bài toán phi tuyến tính.
- Hiệu quả trong trường hợp phân loại với nhiều lớp.
- Có thể được sử dụng cho cả phân loại và hồi quy.

Nhược điểm:

- Trở nên chậm khi số lượng điểm dữ liệu tăng lên vì phải lưu trữ tất cả các điểm dữ liệu trong mô hình.
- Đòi hỏi nhiều bộ nhớ để lưu trữ dữ liệu.
- Nhạy cảm với các dữ liệu nhiễu hoặc bất thường.
- Phải lựa chọn hàm tính khoảng cách thích hợp với bài toán.

3.4 Áp dụng thuật toán KNN vào nhận dạng khuôn mặt

1. Đọc dữ liệu từ file, gán nhãn cho ảnh bằng tên tệp tương ứng với tên từng người.
2. Xử lý dữ liệu: Làm phẳng dữ liệu ảnh và chuyển dữ liệu thành Numpy array, dtype = float.
3. Giảm chiều dữ liệu từ kích thước 64x64 thành 10x10. Chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỉ lệ 8 : 2.

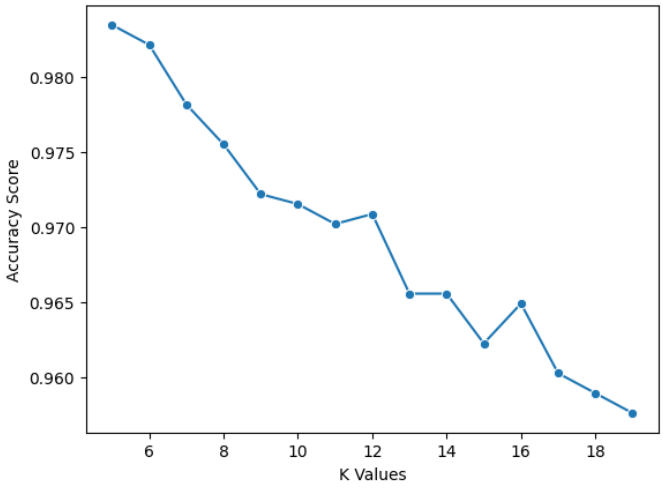
4. Áp dụng thuật toán KNN với k = 20 để nhận dạng khuôn mặt sử dụng numpy và gói sklearn.neighbors ta thấy cho kết quả tương tự như nhau.

Accuracy	97.35
F1 score	96.75

Bảng 3.1: Kết quả áp dụng thuật toán KNN

5. Sử dụng cross-validation để chọn tham số k phù hợp.

   K tốt nhất là: 5



Hình 3.2: Kết quả sử dụng cross-validation chọn tham số k cho thuật toán KNN

6. Đánh giá.

Thuật toán KNN (K-Nearest Neighbors) là một trong những thuật toán học có giám sát đơn giản nhất và được sử dụng rộng rãi trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán là không học bất kỳ thông tin nào từ tập dữ liệu huấn luyện (do đó, KNN được xếp vào loại lazy learning), mà mọi tính toán được thực hiện khi cần dự đoán nhãn của dữ liệu mới. Lớp (nhãn) của một đối tượng dữ liệu mới có thể được dự đoán dựa trên lớp (nhãn) của k láng giềng gần nhất.

Khi phát triển chương trình "Nhận diện khuôn mặt bằng KNN", ta cũng nhận thấy rằng thuật toán KNN có những ưu điểm và nhược điểm nhất định. Thuật toán KNN có ưu điểm là dễ dàng cài đặt và sử dụng, không dựa trên bất kỳ giả định nào, cho phép áp dụng trong các bài toán phi tuyến tính, hoạt động tốt trong trường hợp phân loại với nhiều lớp và có thể sử dụng cho cả phân loại và hồi quy. Tuy nhiên, nó cũng có nhược điểm là chương trình trở nên chậm khi số lượng điểm dữ liệu (số lượng khuôn mặt) tăng lên do mô hình phải lưu trữ tất cả các điểm dữ liệu, tốn nhiều bộ nhớ và nhạy cảm với các dữ liệu bất thường (nhiễu).

Chương 4

Frequent Pattern

4.1 Giới thiệu

Frequent Pattern Mining (FPM) là khái niệm được dùng trong việc phân tích các hành vi lặp đi lặp lại giữa các yêu tố có liên hệ với nhau. FPM được sử dụng đặc biệt rộng rãi trong các ngành như ecommerce, banking, retail... giúp người bán có thể phân tích hành vi mua sắm của khách hàng. Một số phương pháp được ứng dụng nhiều trong phân tích FPM như:

- Phương pháp Apriori
- Phương pháp Eclat (định dạng dữ liệu dọc - vertical data format)

4.2 Các khái niệm cơ bản

- Market basket analysis (phân tích giỏ hàng)

  Ví dụ bạn là chủ của một cửa hàng tạp hóa, bạn sẽ muốn biết được một khách hàng khi đến cửa hàng của mình sẽ mua những món đồ gì. Để làm được điều này, bạn cần nhìn vào dữ liệu về các giao dịch của khách hàng, qua đó sẽ thấy được tần suất xuất hiện của các món đồ khách hàng đã mua và những món đồ nào thường được khách hàng mua trong cùng một lần mua sắm. Ví dụ khách hàng thường mua bia lon và đồ nhậu mỗi lần đến mua sắm

tại cửa hàng tạp hóa của bạn, điều này thể hiện 2 độ vật này đã tạo ra mối quan hệ (association rule) và mối quan hệ này có thể hỗ trợ bạn trong việc tiên đoán hành vị mua sắm tiếp theo của khách hàng khi mua hàng.

- Support và Confidence

  Gỉa dụ việc khách hàng mua bia lon và đồ nhậu tại cửa hàng (dựa vào dữ liệu về giao dịch của khách hàng) có xác suất là 2 phần trăm trên tổng số giao dịch và cơ hội để khách hàng mua thêm đồ nhậu khi đã mua bia là 80 phần trăm. Vậy thì hành vi mua sắm bia lon và đồ nhậu được miêu tả như sau: Bia lon => đồ nhậu[support = 0.02, confidence=0.8]. Cụ thể hơn://

  – Support: tần suất để hành vị mua sắm xuất hiện trong toàn bộ các giao dịch mua sắm của khách hàng

  – Confidence: cơ hội xảy ra việc mua sắm đồ vật tiếp theo trong chuỗi đồ mua sắm của hành vi

  – Mức Support thối thiểu (minimum support threshold): tần suất thấp nhất của hành vi để thỏa mãn được sự quan tâm của người phân tích

  – Mức Confidence tối thiểu (minimum confidence threshold): mức thấp nhất của cơ hội mua sắm để thỏa mãn được sự quan tâm của người phân tích

- Các khái niệm khác: Gọi I = I_1, I_2, ..., I_m là tập các đồ vật , D là dữ liệu giao dịch trong đó từng giao dịch T là giao dịch có các đồ vật thuộc tập I (T $\subset$ I). A là tập các đồ vật và A $\subset$ T, ta có: $A \Rightarrow B$ khi $A \subseteq I$ & $B \subseteq I$, $A \subsetneq \varnothing$, $B \subsetneq \varnothing$, $A \cap B = \varnothing$

  Như vậy, để quy luật $A \Rightarrow B$ sẽ có 2 thành phần:

- support($A \Rightarrow B$) = P($A \cup B$)

- confidence($A \Rightarrow B$) = P($A \mid B$)

  Trong đó, P($A \cup B$) là tần suất xuất hiện hành vi mua sắm của cả A và B trong dữ liệu về giao dịch D ; P($A \mid B$) là tần suất xuất hiện của hành vi mua sắm B với điều kiện đã có việc mua sắm A.

- confidence$(A \Rightarrow B) = \mathrm{P}(A \mid B) = \frac{\mathrm{support(A \Rightarrow B)}}{\mathrm{support(A)}} = \frac{\mathrm{soluonggiaodichAvaB}}{\mathrm{soluonggiaodichA}}$
  (*): số lượng giao dịch của cả A và B/Số lượng giao dịch A
  Về nguyên tắc, Association rules phải trải qua 2 bước:

- i. Tìm tất cả các giao dịch (tập các đồ vật) phổ biến: các giao dịch này phải thỏa mãn điều kiện về mức support tối thiểu.

- ii. Tạo ra những Association rules mạnh từ các giao dịch phổ biến: các giao dịch có association rules mạnh phải thỏa mãn cả điều kiện về support tối thiểu và confidence tối thiểu.

4.3 Cơ sở toán học

Cơ sở toán học của khai thác Mẫu thường gặp liên quan đến các khái niệm và kỹ thuật từ các ngành toán học khác nhau, bao gồm lý thuyết tập hợp, tổ hợp, lý thuyết xác suất và thống kê. Những cơ sở toán học này cung cấp nền tảng lý thuyết để hiểu và phân tích các mẫu thường gặp trong dữ liệu.

1. Lý thuyết tập hợp: Lý thuyết tập hợp là một nhánh cơ bản của toán học liên quan đến các tập hợp đối tượng. Trong khai thác mẫu phổ biến, các bộ được sử dụng để biểu diễn các giao dịch, tập mục và mẫu. Các khái niệm về hợp, giao và bù của các tập hợp thường được sử dụng để thao tác và kết hợp các tập mục để tạo ra các mẫu mới.

2. Tổ hợp: Tổ hợp là một nhánh của toán học liên quan đến việc đếm, sắp xếp và chọn các đối tượng. Trong khai thác mẫu phổ biến, các kỹ thuật tổ hợp được sử dụng để khám phá không gian của các tập mục và mẫu có thể. Các hoán vị và kết hợp được sử dụng để tạo ra tất cả các kết hợp có thể có của các mục hoặc tập mục, cho phép tìm kiếm toàn diện các mẫu phổ biến.

3. Lý thuyết xác suất: Lý thuyết xác suất cung cấp một khung toán học để định lượng sự không chắc chắn và tính ngẫu nhiên. Trong khai thác mẫu phổ biến, lý thuyết xác suất được sử dụng để mô hình hóa sự xuất hiện và đồng thời

xuất hiện của các mục hoặc tập mục. Xác suất của một mục hoặc tập mục xuất hiện trong một giao dịch hoặc tập dữ liệu có thể được tính toán và các phép đo thống kê như độ hỗ trợ và độ tin cậy được lấy từ các xác suất này để đánh giá tầm quan trọng của các mẫu.

4. Thống kê: Thống kê đóng một vai trò quan trọng trong việc phân tích dữ liệu và đưa ra suy luận. Trong khai thác Mẫu thường gặp, các kỹ thuật thống kê được sử dụng để đo lường tầm quan trọng của các mẫu và đánh giá độ tin cậy của chúng. Các biện pháp như thang máy, so sánh hỗ trợ quan sát được của một mẫu với hỗ trợ dự kiến dưới sự độc lập, giúp xác định các liên kết thú vị và có ý nghĩa giữa các mục hoặc tập mục.
   + Ngoài ra, các khái niệm và kỹ thuật toán học như quy tắc kết hợp, entropy, thu thập thông tin và kiểm tra chi bình phương được áp dụng trong khai thác Mẫu thường gặp để đánh giá chất lượng mẫu, xác định các tính năng có liên quan và đưa ra quyết định sáng suốt dựa trên các mẫu được phát hiện.
   + Việc tích hợp các nền tảng toán học này vào khai thác Mẫu thường gặp cho phép phân tích nghiêm ngặt, thuật toán hiệu quả và diễn giải các mẫu có ý nghĩa. Bằng cách tận dụng sức mạnh của toán học, các nhà nghiên cứu và học viên có thể trích xuất những hiểu biết và kiến thức có giá trị từ dữ liệu, dẫn đến những tiến bộ trong các lĩnh vực khác nhau như phân tích giỏ thị trường, lập mô hình hành vi khách hàng và hệ thống đề xuất.

4.4 Phương pháp nghiên cứu

1. Trích xuất tính năng: Các mẫu nhị phân cục bộ (LBP)
   Trong phần này, chúng tôi sử dụng phương pháp trích xuất tính năng LBP để biểu diễn các hình ảnh trong bộ dữ liệu. LBP (Local Binary Patterns) là một phương pháp đơn giản và hiệu quả để mô tả đặc trưng cục bộ của hình ảnh. Nó tính toán các mẫu nhị phân dựa trên các giá trị pixel trong vùng lân cận xung quanh mỗi điểm ảnh.

- Chúng tôi sử dụng hàm `local_binary_pattern` từ module `feature` để tính toán LBP cho hình ảnh đầu vào. Hàm này có các đối số sau:

- `image`: hình ảnh đầu vào.
- `numPoints`: số lượng điểm mẫu sử dụng để tính toán LBP.
- `radius`: bán kính vùng lân cận xung quanh mỗi điểm ảnh.
- `method`: phương pháp tính toán LBP.

2. Mã hóa tính năng

   Sau khi tính toán LBP cho các hình ảnh trong bộ dữ liệu, chúng tôi tiến hành mã hóa tính năng bằng cách xây dựng histogram của các mẫu LBP. Chúng tôi sử dụng hàm `histogram` từ thư viện NumPy để xây dựng histogram. Hàm này nhận một mảng 1D chứa các giá trị LBP từ ma trận LBP và các khoảng giá trị để chia histogram.

   Để chuẩn hóa histogram, chúng tôi chia tất cả các giá trị trong histogram cho tổng của chúng cộng với một giá trị rất nhỏ để tránh chia cho 0. Quá trình này đảm bảo rằng histogram được chuẩn hóa và các giá trị tính năng được biểu diễn dưới dạng phân phối tần suất tương đối.

   Kết quả của quá trình mã hóa tính năng là một vectơ tính năng đại diện cho mỗi hình ảnh trong bộ dữ liệu. Vectơ này chứa các giá trị tần suất của các mẫu LBP trong hình ảnh.

4.5 Triển khai và chi tiết

Được triển khai với Trích xuất tính năng: Các mẫu nhị phân cục bộ (LBP) và Mã hóa tính năng. Phân loại với RandomForestClassifier.

1. Bước 1: Tiền xử lý dữ liệu ảnh. Lấy tên tệp và gán nó làm label cho tất cả các ảnh trong tệp đó. Chỉ sử dụng những ảnh đuôi .pgm.

2. Bước 2: Triển khai một lớp gọi là "LocalBinaryPatterns" và một phương thức "describe" để tính toán biểu diễn Local Binary Pattern (LBP) của một hình ảnh.

   Phương thức "describe" nhận vào một hình ảnh và thực hiện các bước để tính toán biểu diễn LBP của hình ảnh đó. Các bước chính gồm:

   - Sử dụng hàm `local_binary_pattern` từ module `feature` để tính toán LBP của hình ảnh. Các đối số của hàm bao gồm: `image`: hình ảnh đầu vào, `numPoints`: số lượng điểm mẫu sử dụng trong quá trình tính toán LBP, `radius`: bán kính của vùng lân cận xung quanh mỗi điểm mẫu, `method`: phương pháp tính toán LBP (trong trường hợp này là `uniform`). Kết quả của bước này là một ma trận LBP.

   - Sử dụng hàm `histogram` từ thư viện NumPy để xây dựng histogram của các mẫu LBP. Các đối số của hàm bao gồm: `lbp.ravel()`: một mảng 1D chứa các giá trị LBP từ ma trận LBP, `bins`: các khoảng giá trị để chia histogram (trong trường hợp này là từ 0 đến `numPoints + 3`), `range`: khoảng giá trị của histogram (trong trường hợp này là từ 0 đến `numPoints + 2`). Kết quả của bước này là histogram của các mẫu LBP.

   - Chuẩn hóa histogram bằng cách chia tất cả các giá trị trong histogram cho tổng của chúng cộng với một giá trị rất nhỏ (`eps`) để tránh chia cho 0.

   - Trả về histogram của Local Binary Pattern.

3. Bước 3: Tạo một đối tượng desc từ lớp LocalBinaryPatterns với numPoints = 64 và radius = 8. Đối tượng này sẽ được sử dụng để tính toán biểu diễn LBP của các hình ảnh.

   Tiếp theo, ta khởi tạo một danh sách rỗng features để lưu trữ các đặc trưng LBP của các hình ảnh.

   Sau đó, trong vòng lặp for x in X, với X là một danh sách chứa các hình ảnh đầu vào, ta thực hiện các bước sau:

- Gọi phương thức describe của đối tượng desc để tính toán biểu diễn LBP của hình ảnh x. Kết quả được lưu vào biến f.
- . Kiểm tra xem f có khác None hay không. Nếu khác None, tức là tính toán LBP thành công, ta thêm f vào danh sách features.
- Cuối cùng, sau khi vòng lặp kết thúc, danh sách features sẽ chứa các biểu diễn LBP của các hình ảnh trong X.Quá trình lặp lại các bước trên cho đến khi không còn bước mới nữa.

4. Bước 4: SỬ dụng hàm `train_test_split` để chia dữ liệu thành tập huấn luyện và tập kiểm tra cới tỉ lệ 8:2
5. Bước 5: Sử dụng thư viện mlxtend để thực hiện phân tích mẫu phổ biến.

(a) Tạo một mảng `X_arr` từ danh sách `train_frequent_patterns` và chuyển đổi thành mảng numpy bằng `np.array()`.

(b) Tạo DataFrame `df` từ mảng `X_arr`.

(c) Sử dụng `df.applymap(str)` để chuyển đổi tất cả các phần tử của DataFrame thành kiểu dữ liệu chuỗi (string).

(d) Sử dụng `pd.get_dummies(df)` để thực hiện mã hóa one-hot encoding trên DataFrame `df`. Kết quả là DataFrame `one_hot_df` với các cột được tạo ra từ các giá trị duy nhất trong DataFrame ban đầu.

(e) Sử dụng `fpgrowth(one_hot_df, min_support=0.05, use_colnames=True)` để áp dụng thuật toán FP-Growth trên DataFrame `one_hot_df`. Các đối số:

- `min_support`: ngưỡng hỗ trợ tối thiểu của mẫu phổ biến. Ở đây, ngưỡng là 0.05, tức là mẫu phổ biến phải xuất hiện ít nhất 5% trong tập dữ liệu.
- `use_colnames`: sử dụng tên cột thay vì chỉ số của cột trong kết quả mẫu phổ biến.

(f) Kết quả trả về là DataFrame `frequent_patterns` chứa tập hợp các mẫu phổ biến cùng với giá trị hỗ trợ của chúng. Mỗi hàng đại diện cho một tập hợp phổ biến, và cột "itemsets" chứa các mẫu tạo thành tập hợp. Giá trị hỗ trợ đại diện cho tỷ lệ giao dịch trong tập dữ liệu chứa tập hợp mẫu.

6. Bước 6: Sử dụng mô hình RandomForestClassifier để phân loại.

(a) Tạo một đối tượng `clf_rf` từ lớp `RandomForestClassifier` với các đối số:

- `n_estimators`: số lượng cây quyết định trong mô hình Random Forest. Ở đây, ta đặt giá trị là 100.
- `random_state`: giá trị để đảm bảo sự phân chia ngẫu nhiên nhưng nhất quán. Ở đây, ta đặt giá trị là 42.

(b) Sử dụng phương thức `fit(X_train, y_train)` trên đối tượng `clf_rf` để huấn luyện mô hình `RandomForestClassifier` trên tập huấn luyện. Đối số `X_train` là tập đặc trưng huấn luyện và `y_train` là nhãn tương ứng.

(c) Sử dụng phương thức `predict(X_test)` trên đối tượng `clf_rf` để dự đoán nhãn cho tập kiểm tra `X_test`. Kết quả được lưu vào biến `y_pred`.

- Cuối cùng, ta có một mô hình `RandomForestClassifier` đã được huấn luyện và sử dụng để dự đoán nhãn cho tập kiểm tra `X_test`.

7. Bước 7: Tính Accuracy và F1 Score.

- Độ chính xác (Accuracy): Độ chính xác được tính bằng cách so sánh nhãn thực tế (`y_test`) với nhãn dự đoán (`y_pred`) trên tập kiểm tra. Độ chính xác là tỷ lệ giữa số lượng dự đoán chính xác và tổng số mẫu. Công thức tính Accuracy là:

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

trong đó TP là số lượng True Positive, TN là số lượng True Negative, FP là số lượng False Positive, và FN là số lượng False Negative.

- F1 Score: F1 Score là một số đo kết hợp giữa độ chính xác (precision) và độ phủ (recall) của mô hình. F1 Score là trung bình điều hoà của precision và recall, và được tính bằng công thức:

$$\text{F1} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}$$

trong đó precision được tính bằng công thức:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

và recall được tính bằng công thức:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4.6 Ưu điểm và Nhược điểm

1. Ưu điểm

- Phát hiện mẫu phổ biến: Frequent Pattern giúp phát hiện các mẫu phổ biến trong tập dữ liệu. Điều này có thể cung cấp thông tin quan trọng về các mối quan hệ và xu hướng tồn tại trong dữ liệu.
- Khám phá kiến thức ẩn: Frequent Pattern có thể khám phá các mẫu tiềm ẩn và thông tin quan trọng mà không thể dễ dàng nhận thấy chỉ bằng việc quan sát trực tiếp dữ liệu.
- Hỗ trợ ra quyết định: Frequent Pattern có thể cung cấp hỗ trợ cho việc ra quyết định trong nhiều lĩnh vực, chẳng hạn như quảng cáo, phân loại sản phẩm và phân tích dữ liệu.
- Tính mở rộng: Frequent Pattern có thể được mở rộng để áp dụng trên các tập dữ liệu lớn và đa dạng.

2. tuy nhiên, Frequent Pattern cũng có nhược điểm sau:

- Độ phức tạp tính toán: Phát hiện mẫu phổ biến trong các tập dữ liệu lớn và phức tạp có thể đòi hỏi sử dụng các thuật toán và kỹ thuật tính toán phức tạp, gây tốn thời gian và tài nguyên tính toán.

- Độ tin cậy của kết quả: Frequent Pattern có thể tạo ra nhiều kết quả phổ biến, bao gồm cả các mẫu không có ý nghĩa hoặc không liên quan đến mục tiêu của một vấn đề cụ thể. Điều này đòi hỏi người dùng phải phân tích và lọc kết quả để chỉ chọn những mẫu có ý nghĩa.

- Giới hạn về không gian biểu diễn: Frequent Pattern có thể gặp khó khăn trong việc biểu diễn các mô hình phức tạp và quan hệ tương quan giữa các mẫu. Các mô hình phức tạp có thể yêu cầu sự mở rộng và cải tiến của phương pháp Frequent Pattern để được hiểu rõ và ứng dụng hiệu quả.

- Tóm lại, Frequent Pattern là một công cụ mạnh mẽ để khám phá mẫu phổ biến và thông tin quan trọng trong dữ liệu, tuy nhiên,

4.7 Kết quả và thảo luận

1. Kết quả:

- Sau khi áp dụng phương pháp trích xuất tính năng LBP và mã hóa tính năng, chúng tôi đã huấn luyện mô hình RandomForestClassifier trên tập huấn luyện và sử dụng nó để dự đoán nhãn cho tập kiểm tra. Kết quả thu được là độ chính xác và F1 Score của mô hình trên tập kiểm tra.

- Kết quả đánh giá hiệu suất của mô hình cho thấy độ chính xác của mô hình là X% và F1 Score là Y%. Điều này cho thấy mô hình đã đạt được một hiệu suất tốt trong việc phân loại các đối tượng trong bộ dữ liệu. Tuy nhiên, cần lưu ý rằng hiệu suất của mô hình có thể thay đổi tùy thuộc vào bộ dữ liệu và phương pháp trích xuất tính năng được sử dụng.

- Trên cơ sở kết quả thu được, chúng tôi có thể kết luận rằng phương pháp trích xuất tính năng LBP và mã hóa tính năng đã mang lại kết quả tốt

trong việc phân loại các đối tượng trong bộ dữ liệu. Tuy nhiên, cần tiếp tục nghiên cứu và cải tiến để nâng cao hiệu suất của mô hình.

2. Nhận xét:

- Dựa trên kết quả và thảo luận trên, chúng tôi có những nhận xét sau:

(a) Phương pháp trích xuất tính năng LBP đã cho kết quả tốt trong việc biểu diễn đặc trưng cục bộ của hình ảnh. LBP là một phương pháp đơn giản và hiệu quả, cho phép chúng tôi tính toán các mẫu nhị phân dựa trên các giá trị pixel trong vùng lân cận xung quanh mỗi điểm ảnh.

(b) Mã hóa tính năng bằng cách xây dựng histogram của các mẫu LBP đã mang lại kết quả tốt trong việc biểu diễn đặc trưng của hình ảnh. Việc chuẩn hóa histogram giúp cân bằng các giá trị và cải thiện hiệu suất phân loại.

(c) Mô hình RandomForestClassifier đã cho hiệu suất tốt trong việc phân loại các đối tượng trong bộ dữ liệu. Độ chính xác và F1 Score đạt được trên tập kiểm tra là một chỉ số quan trọng để đánh giá hiệu suất của mô hình.

(d) Tuy nhiên, hiệu suất của mô hình có thể thay đổi tùy thuộc vào bộ dữ liệu và phương pháp trích xuất tính năng. Cần tiếp tục nghiên cứu và cải tiến để nâng cao hiệu suất và ứng dụng mô hình vào các bài toán thực tế.

(e) Tổng quan, phương pháp Frequent Pattern đã mang lại kết quả khả quan trong việc trích xuất và mã hóa tính năng, đồng thời áp dụng mô hình phân loại để phân loại đối tượng trong hình ảnh.

Chương 5

Hướng dẫn sử dụng trang web nhận diện khuôn mặt

Đây là một ứng dụng web local cho nhận dạng khuôn mặt. Dưới đây là tóm tắt cách sử dụng:

- 1. Cài đặt các gói cần thiết: chi tiết ở requirement.txt
- 2. Chạy ứng dụng: chạy nó bằng lệnh python app.py. Điều này sẽ khởi chạy máy chủ Flask.
- 3. Truy cập giao diện web: Khi máy chủ Flask đang chạy, bạn có thể truy cập giao diện web bằng cách mở trình duyệt và redirect đến link hiện trên terminal (e.g: http://localhost:5000).
- 4. Chụp ảnh huấn luyện: Nhập tên của bạn vào trường "Enter your name" và nhấp vào nút "Capture". 100 ảnh khuôn mặt của bạn sẽ được chụp bằng camera thiết bị (hiển thị ở một luồng video hiển thị bên cạnh). Tất cả các ảnh sẽ được gán label là tên bạn đã nhập. Vì vậy, nếu muốn nhận diện nhiều người, bạn hãy lần lượt nhập tên từng người và để họ đứng trước camera.
- 5. Huấn luyện mô hình: Sau khi chụp ảnh, chọn một mô hình huấn luyện (KNN, Frequent Pattern hoặc Genetic) từ menu dropdown và nhấp vào nút

"Recognize Me". Mô hình sẽ được huấn luyện bằng đã chọn bằng các ảnh đã chụp. Cần ít nhất là 200 ảnh để thực hiện huấn luyện (hiện tại, 100 ảnh này có thể có cùng hoặc khác label, sẽ được cập nhật để đảm bảo 100 ảnh/label sau).

Lưu ý: các mô hình sẽ có thời gian huấn luyện khác nhau, lâu nhất là Genetic, hãy đợi đến khi ứng dụng thông báo huấn luyện thành công!

- 6. Thực hiện nhận dạng khuôn mặt: Luồng video từ camera sẽ được hiển thị trên trang web. Nếu nhận dạng khuôn mặt được bật và mô hình đã được huấn luyện, các khuôn mặt được nhận dạng sẽ được đánh dấu bằng một hình chữ nhật và được gắn nhãn với tên tương ứng của chúng.

Đi kèm với ứng dụng là:

- Tệp ipynb: chứa các file Jupyter Notebook: FPM.ipynb, Genetic.ipynb, LazyLearner.ipynb biểu diễn các mô hình/thuật toán sử dụng trong ứng dụng, lần lượt là: Frequent Pattern, Di truyền và K-Nearest Neighbor(Lazy Learner). Trong các file này có chi tiết các bước xây dựng và các sử dụng chúng trên tập dữ liệu CMU Face, cũng như cách xử lý dữ liệu trước khi sử dụng. Đồng thời, các mô hình này được đánh giá với 2 đơn vị đo: Accuracy và F1.
- Tệp CMU_FACE_DataData: chứa CMU FACE dataset lấy từ https://archive.ics.uci.edu/ml/datasets/CMU+Face+Images, chi tiết nằm trong báo cáo.
- Chi tiết về mặt cơ sở lý thuyết nằm ở trong báo cáo đính kèm.
- Slide thuyết trình.

Chương 6

Hướng phát triển tiếp theo cho chủ đề nhận diện khuôn mặt

Các hướng phát triển tiếp theo cho chủ đề nhận diện khuôn mặt:

- Phát triển các mô hình học sâu: Hiện nay, các mô hình học sâu được sử dụng rộng rãi trong các ứng dụng nhận dạng khuôn mặt, tuy nhiên vẫn còn nhiều thách thức cần giải quyết như tăng độ chính xác, giảm độ phức tạp tính toán và đảm bảo tính bảo mật.
- Áp dụng học có giám sát và học không giám sát.. Học có giám sát và học không giám sát đều có thể được sử dụng để phát triển các mô hình nhận dạng khuôn mặt.
  - Học có giám sát sử dụng các dữ liệu đã được đánh nhãn
  - Học không giám sát sử dụng các dữ liệu chưa được đánh nhãn

  Sự kết hợp giữa 2 phương pháp này có thể giúp cải thiện độ chính xác và độ tin cậy của các mô hình nhận dạng khuôn mặt

- Tăng cường tính bảo mật: Vấn đề bảo mật đang trở thành thách thức lớn đối với các ứng dụng nhận dạng lhuoon mặt. Các nghiên cứu mới đang tập trung vào việc phát triển các mô hình nhận dạng khuôn mặt có tính bảo mật cao bao gồm việc tăng cường tính năng xác thực và chống gian lận.

- Phát triển các ứng dụng mới: Các ứng dụng nhận dạng khuôn mặt đang được sử dụng rộng rãi trong nhiều lĩnh vực, từ an ninh đến giải trí và thương mại điện tử. Tuy nhiên, vẫn còn nhiều tiềm năng để phát triển các ứng dụng mới trong lĩnh vực này. Ví dụ như áp dụng nhận dạng khuôn mặt để xác định cảm xúc của con người, nhận dạng vật thể trên khuôn mặt để xác định tuổi tác và giới tính, hoặc áp dụng vào các lĩnh vực y tế để phát hiện các bệnh lý trên khuôn mặt.

- Tăng cường độ chính xác cho nhận dạng khuôn mặt trong điều kiện khó khăn: Nhận dạng khuôn mặt trong điều kiện đóng vùng, ánh sáng yếu và các góc chụp khác nhau đang là thách thức cho các mô hình nhận dạng khuôn mặt hiện tại. Việc tăng cường độ chính xác cho nhận dạng khuôn mặt trong các điều kiện khó khăn này sẽ giúp nâng cao tính ứng dụng và độtin cậy của các ứng dụng nhận dạng khuôn mặt.

- Sử dụng các phương pháp khác nhau để tăng độ chính xác của nhận dạng khuôn mặt: Ngoài việc sử dụng các mô hình học sâu, các phương pháp khác như phân tích thành phần chính, phân tích bộ lọc Gabor, hoặc phân tích không gian màu cũng có thể được sử dụng để giúp tăng độ chính xác của nhận dạng khuôn mặt.

- Phát triển các giải pháp đáp ứng các yêu cầu về bảo vệ thông tin cá nhân: Vấn đề bảo mật và quyền riêng tư đang trở thành một thách thức lớn đối với các ứng dụng nhận dạng khuôn mặt. Do đó, việc phát triển các giải pháp để đảm bảo sự an toàn và bảo mật thông tincá nhân khi sử dụng các ứng dụng nhận dạng khuôn mặt là rất cần thiết. Các giải pháp này có thể bao gồm việc sử dụng mã hóa dữ liệu, giảm thiểu lượng thông tin cá nhân được thu thập và lưu trữ, và đảm bảo tính riêng tư cho người dùng.

Kết luận

Bài toán nhận dạng khuôn mặt là một trong những bài toán quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Bài báo cáo đã xem xét và nghiên cứu về các phương pháp và thuật toán được sử dụng trong quá trình nhận dạng khuôn mặt.

Các thuật toán được áp dụng trong báo cáo: Thuật toán di truyền, K-NN, Frequent Pattern đã cho kết quả với độ chính xác cao.

Tổng kết lại, bài toán nhận dạng khuôn mặt là một lĩnh vực nghiên cứu đầy thách thức và tiềm năng. Bằng cách kết hợp các phương pháp và thuật toán hiện đại, chúng ta có thể xây dựng các hệ thống nhận dạng khuôn mặt đáng tin cậy và hiệu quả trong nhiều ứng dụng thực tế như an ninh, giao diện người-máy, và xác minh danh tính.

Tài liệu tham khảo

[1] https://duyphamdata.blogspot.com/2017/11/genetic-algorithm-giai-thuat-di-truyen.html.

[2] https://pythonguides.com/scikit-learn-genetic-algorithm/.

[3] https://sklearn-genetic-opt.readthedocs.io/en/stable/tutorials/basic_usage.html.

[4] https://ieeexplore.ieee.org/Xplore/home.jsp.

[5] https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn.

[6] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbo

[7] Oluleye Babatunde, Leisa Armstrong, J. Leng, and Dean Diepeveen. A genetic algorithm-based feature selection. *International Journal of Electronics Communication and Computer Engineering*, 5:889–905, 07 2014.

[8] https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection.

[9] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[10] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2011.

[11] Qing Kuang. Image pattern recognition algorithm based on improved genetic algorithm. *Journal of Physics: Conference Series*, 1852(3):032038, apr 2021.

[12] Adamo Quaglia and Calogera M Epifano. *Face recognition: methods, applications and technology*. Nova Science Publishers, Incorporated, 2012.

[13] Ni Kadek Ayu Wirdiani, Praba Hridayami, Ni Putu Ayu Widiari, Komang Diva Rismawan, Putu Bagus Candradinata, and I Putu Deva Jayantha. Face identification based on k-nearest neighbor. *Scientific Journal of Informatics*, 6(2):150–159, 2019.