**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

****

**LUẬN VĂN THẠC SĨ**

**Tối ưu hóa dữ liệu và mô hình phân lớp nhằm nâng cao chất lượng nhận dạng cảm xúc sử dụng bộ dữ liệu sinh học của MIT**

**NGUYỄN TIẾN NAM**

nguyennam291096@gmail.com

**Điều khiển và tự động hóa**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | TS. Cung Thành Long  Chữ ký của GVHD |
| **Viện:** | Điện |
|  |  |
| **HÀ NỘI, 3/2021** | |

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

**Độc lập – Tự do – Hạnh phúc**

**BẢN XÁC NHẬN CHỈNH SỬA LUẬN VĂN THẠC SĨ**

**Họ và tên tác giả luận văn :** …………………………………........……………..

**Đề tài luận văn:** ………………………………………….....……………...............….

**Chuyên ngành:**……………………………...…………………........................…..........

**Mã số SV**:………………………………….. …………………....................................…...

Tác giả, Người hướng dẫn khoa học và Hội đồng chấm luận văn xác nhận tác giả đã sửa chữa, bổ sung luận văn theo biên bản họp Hội đồng ngày….........................………… với các nội dung sau:

……………………………………………………………………………………………………..……………………………………………………………………………………………………..……………………………………………………………………………………………………..……………………………………………………………………………………………………..……………………………………………………………………………………………………..……………………………………………………………………………………………………..……………………………………………………………………………………..

Ngày tháng năm

**Giáo viên hướng dẫn Tác giả luận văn**

**CHỦ TỊCH HỘI ĐỒNG**

**ĐỀ TÀI LUẬN VĂN**

Biểu mẫu của Đề tài/Luận văn tốt nghiệp theo qui định của Viện, tuy nhiên cần đảm bảo giáo viên giao đề tài ký và ghi rõ họ và tên.

Giáo viên hướng dẫn

Ký và ghi rõ họ tên

**Lời cảm ơn**

Đây là mục tùy chọn, nên viết phần cảm ơn ngắn gọn, tránh dùng các từ sáo rỗng, giới hạn trong khoảng 100-150 từ.

**Tóm tắt nội dung luận văn**

Tóm tắt nội dung của luận văn thạc sĩ trong khoảng tối đa 500 chữ. Phần tóm tắt cần nêu được các ý: vấn đề cần thực hiện; phương pháp thực hiện; công cụ sử dụng (phần mềm, phần cứng…); kết quả của luận văn có phù hợp với các vấn đề đã đặt ra hay không; tính khoa học thực tiễn của luận văn, định hướng phát triển mở rộng của luận văn (nếu có);

HỌC VIÊN

Ký và ghi rõ họ tên

**MỤC LỤC**

[CHƯƠNG 1. TỔNG QUAN VỀ NHẬN DẠNG CẢM XÚC 2](#_Toc67472281)

[1.1 Bài toán nhận dạng cảm xúc 2](#_Toc67472282)

[1.1.1 Nhận dạng cảm xúc thông qua tín hiệu giọng nói 2](#_Toc67472283)

[1.1.2 Nhận dạng cảm xúc thông qua hình ảnh khuôn mặt 2](#_Toc67472284)

[1.1.3 Nhận dạng cảm xúc thông qua tín hiệu sinh học 2](#_Toc67472285)

[1.2 Bộ dữ liệu nhận dạng cảm xúc MIT 2](#_Toc67472286)

[CHƯƠNG 2. HỌC MÁY VÀ ỨNG DỤNG TRONG NHẬN DẠNG CẢM XÚC 2](#_Toc67472287)

[2.1 Học máy và các kỹ thuật sử dụng trong bài toán phân loại (classification) 2](#_Toc67472288)

[2.1.1 Lựa chọn đặc trưng 2](#_Toc67472289)

[2.1.2 Phân lớp và ra quyết định nhận dạng 2](#_Toc67472290)

[2.2 Áp dụng với bộ dữ liệu MIT 2](#_Toc67472291)

[CHƯƠNG 3. TỔ CHỨC DỮ LIỆU ĐẦU VÀO NÂNG CAO CHẤT LƯỢNG NHẬN DẠNG CẢM XÚC 2](#_Toc67472292)

[3.1 Thử nghiệm một số kỹ thuật nhận dạng lên bộ dữ liệu MIT 2](#_Toc67472293)

[3.2 Thay đổi tỉ lệ chồng chập 2](#_Toc67472294)

[3.3 Phân tích kết quả đạt được 2](#_Toc67472295)

**DANH MỤC HÌNH VẼ**

**No table of figures entries found.**

**DANH MỤC BẢNG**

**No table of figures entries found.**

**DANH MỤC TỪ VIẾT TẮT**

|  |  |
| --- | --- |
|  |  |

# TỔNG QUAN VỀ NHẬN DẠNG CẢM XÚC

## Bài toán nhận dạng cảm xúc

Cảm xúc, ảnh hưởng đến tâm sinh lý của mỗi người, đóng một vai trò rất quan trọng trong cuộc sống của con người. Cảm xúc tích cực giúp cải thiện sức khỏe con người và hiệu quả công việc, trong khi cảm xúc tiêu cực có thể gây ra các vấn đề về sức khỏe, quá nhiều cảm xúc sẽ gây những ảnh hưởng không tốt cho việc đưa ra quyết định. Cảm xúc tương tác với suy nghĩ theo những cách không rõ ràng nhưng lại đóng vai trò quan trọng đối với hoạt động đòi hỏi trí thông minh. Ngược lại, sự tích tụ lâu dài của những cảm xúc tiêu cực là yếu tố dẫn đến các bệnh lý nghiêm trọng. Cảm xúc phản ánh trạng thái tinh thần một cách tự phát chứ không thông qua nỗ lực có ý thức và đi kèm với những thay đổi về thể chất và sinh lý liên quan đến các cơ quan của con người như não, tim, da, lưu lượng máu, cơ, nét mặt, giọng nói,... Do mức độ phức tạp của sự tương tác lẫn nhau giữa sinh lý và tâm lý trong cảm xúc, việc nhận biết cảm xúc của con người một cách chính xác vẫn là mục tiêu của nghiên cứu khoa học. Nhận biết biểu cảm được cho là đóng một vai trò quan trọng trong việc học tập và phát triển, là một phần quan trọng trong việc đánh giá sự khác biệt giữa sự phát triển bình thường của trẻ em so với trẻ tự kỷ, những người thường bị suy giảm khả năng nhận biết cảm xúc. Nhận diện cảm xúc đã được áp dụng trong nhiều lĩnh vực như lái xe an toàn, theo dõi sức khỏe con người,…

Để nhận biết được cảm xúc, trước hết các trạng thái cảm xúc cần được xác định một cách rõ ràng. Định nghĩa về những cảm xúc cơ bản lần đầu tiên được đề xuất cách đây nhiều thập kỷ. Các nhà tâm lý học có xu hướng mô hình hóa cảm xúc theo hai cách khác nhau. Một là chia cảm xúc thành các loại rời rạc. Cách khác là sử dụng nhiều chiều để gắn nhãn cảm xúc.



Các phương pháp nhận biết cảm xúc có thể được phân thành hai loại chính:

Sử dụng các tín hiệu bên ngoài của con người như nét mặt, giọng nói, cử chỉ, tư thế,… Các tín hệu này có ưu điểm là dễ thu thập và đã được nghiên cứu trong nhiều năm. Tuy nhiên, độ tin cậy không thể được đảm bảo vì mọi người tương đối dễ dàng kiểm soát các tín hiệu vật lý như nét mặt hoặc giọng nói để che giấu cảm xúc thực của họ, đặc biệt là trong giao tiếp xã hội. Một người có thể tỏ ra vui vẻ ngay cả đang ở trong trạng thái cảm xúc tiêu cực.

Phương pháp thứ hai đang sử dụng các tín hiệu bên trong - các tín hiệu sinh lý, bao gồm điện não đồ, nhiệt độ, điện tâm đồ, điện cơ đồ, hô hấp,... Hệ thống tế bào thần. Các thông số đo dựa vào tín hiệu sinh lý sẽ thay đổi theo một cách nhất định khi con người đối mặt với một số tình huống cụ thể. Một trong những lợi thế chính của phương pháp thứ hai là các thông số đo được phần lớn được kích hoạt không tự nguyện và do đó con người không thể dễ dàng kiểm soát được. Đã có một số nghiên cứu trong lĩnh vực nhận biết cảm xúc bằng cách sử dụng các tín hiệu sinh lý. Đã có nhiều nỗ lực nhằm thiết lập một tiêu chuẩn và mối quan hệ cố định giữa những thay đổi cảm xúc và tín hiệu sinh lý về các loại tín hiệu, tính năng và bộ phân loại. Tuy nhiên, nghiên cứu cho thấy rằng tương đối khó để phản ánh chính xác những thay đổi cảm xúc bằng cách sử dụng một tín hiệu sinh lý duy nhất. Do đó, nhận dạng cảm xúc bằng cách sử dụng nhiều tín hiệu sinh lý thể hiện ý nghĩa của nó trong cả nghiên cứu và ứng dụng thực tế.

### Nhận dạng cảm xúc thông qua tín hiệu giọng nói

### Nhận dạng cảm xúc thông qua hình ảnh khuôn mặt



### Nhận dạng cảm xúc thông qua tín hiệu sinh học

## Bộ dữ liệu sinh học của MIT

Bộ dữ liệu sinh học của MIT được thiết kế để phân loại tối đa lên đến 8 loại cảm xúc khác nhau bao gồm:

Trung tính: cảm giác thông thường, tẻ nhạt

Tức giận: giận dữ, mong muốn được giải tỏa bằng các hình thức bạo lực

Ghét bỏ: trạng thái bị động của cảm xúc tức giận

Đau buồn: sự mất mát, buồn chán, thất vọng

Đồng cảm: vui vẻ, cảm thấy yên bình

Lãng mạn: cảm thấy thu hút, thích thú

Vui vẻ: cảm giác hạnh phúc

Tôn kính: trạng thái thanh thản, ngưỡng mộ



Các nhà khoa học thử nghiệm bằng cách đeo 4 loại cảm biến lên người thử nghiệm: điện cơ đồ, độ điện dẫn da, xung thể tích máu, cảm biến hô hấp. Thí nghiệm được đo trong nhiều ngày, tuy nhiên có những ngày đo giá trị cảm biến bị lỗi do những lý do khách quan như môi trường hay thiết bị đeo bị lỏng trong quá trình đo, do vậy sau khi loại bỏ đi những ngày giá trị bị lỗi, dữ liệu còn lại bao gồm 20 ngày đo, mỗi ngày bao gồm các giá trị ứng với 8 loại cảm xúc bên trên. Bên dưới là hình ảnh biểu diễn các giá trị thu được từ 4 loại cảm biến ứng với 8 loại cảm xúc khác nhau.



Cảm xúc trung tính



Cảm xúc tức giận



Cảm xúc ghét bỏ



Cảm xúc đau buồn



Cảm xúc đồng cảm



Cảm xúc lãng mạn



Cảm xúc vui vẻ



Cảm xúc tôn kính

Có thể nhận thấy các giá trị thu được có sự khác biệt tương đối giữa các loại cảm xúc, điều đó cho thấy tính khả thi trong việc phân loại cảm xúc sử dụng bô dữ liệu của MIT.

# HỌC MÁY VÀ ỨNG DỤNG TRONG NHẬN DẠNG CẢM XÚC

## Học máy và các kỹ thuật sử dụng trong bài toán phân loại (classification)

Học máy (Machine learning) là một phần của trí tuệ nhân tạo. Học máy tập trung vào việc xây dựng các ứng dụng học từ dữ liệu và cải thiện độ chính xác của chúng theo thời gian. Quá trình học bắt đầu từ dữ liệu, mục đích chính là cho phép các máy tính tự động học mà không cần sự can thiệp hay trợ giúp của con người và tự điều chỉnh các hành động tương ứng. Trong học máy, các thuật toán được huấn luyện để tìm ra các mẫu và đặc trưng trong một lượng lớn dữ liệu để đưa ra quyết định và dự đoán dựa trên dữ liệu mới. Thuật toán càng tốt thì các quyết định và dự đoán càng chính xác khi nó xử lý nhiều dữ liệu hơn.

Ngày nay, học máy được ứng dụng rất nhiều trong đời sống như hệ thống gợi ý sản phẩm dựa trên thói quen người dùng, xử lý ngôn ngữ tự nhiên, phát hiện thư rác giúp ngăn các email không mong muốn, hệ thống phân tích hình ảnh y tế giúp phát hiện khối u, ...

Có bốn bước cơ bản để xây dựng một mô hình học máy:

Chuẩn bị tập dữ liệu huấn luyện: Dữ liệu huấn luyện là tập dữ liệu đại diện cho dữ liệu mà mô hình học máy sẽ đưa vào để giải quyết vấn đề. Trong một số trường hợp, dữ liệu huấn luyện được gắn nhãn để nhận biết đặc trưng dữ liệu và phân loại mô hình cần sử dụng. Dữ liệu khác không được gắn nhãn và mô hình sẽ cần trích xuất các đặc trưng đó và xác định định phân loại của riêng nó. Dữ liệu được chia thành hai tập: tập huấn luyện, được sử dụng để huấn luyện cho mô hình và tập đánh giá, được sử dụng để kiểm tra và tinh chỉnh.

Lựa chọn thuật toán để chạy trên tập dữ liệu huấn luyện: Thuật toán là một tập hợp các bước xử lý thống kê. Thuật toán được sử dụng phụ thuộc vào loại dữ liệu (được gắn nhãn hoặc không được gắn nhãn); lượng dữ liệu trong tập dữ liệu huấn luyện và vấn đề cần giải quyết. Các thuật toán học máy phổ biến được sử dụng cho dữ liệu được gắn nhãn như là thuật toán hồi quy tuyến hình hay cây quyết định, … hay các thuật toán học máy phổ biến được sử dụng cho dữ liệu không được gắn nhãn như thuật toán phân cụm K-means, mạng nơ-ron, …

Huấn luyện thuật toán để tạo ra mô hình: Là một quá trình lặp đi lặp lại việc chạy các biến thông qua thuật toán, so sánh kết quả đầu ra với kết quả mà nó đáng lẽ phải tạo ra, điều chỉnh trọng số và độ lệch trong thuật toán để có thể mang lại kết quả chính xác hơn và chạy lại các biến cho đến khi thuật toán trả về kết quả chính xác trong hầu hết thời gian.

Ứng dụng và cải tiến mô hình: Sử dụng mô hình với dữ liệu mới trong trường hợp tốt nhất để mô hình cải thiện độ chính xác và hiệu quả theo thời gian. Dữ liệu mới đến từ đâu sẽ phụ thuộc vào vấn đề đang được giải quyết.

Học máy được chia thành ba loại chính:

**Học máy có giám sát:** Mô hình sẽ được huấn luyện trên tập dữ liệu được gắn nhãn. Tức là, dữ liệu gắn nhãn được sử dụng để huấn luyện mô hình học máy đưa ra các dự đoán phân loại, và khi dự đoán sai thì mô hình sẽ được tinh chỉnh lại cho chính xác hơn. Học máy có giám sát yêu cầu ít dữ liệu huấn luyện hơn so với các phương pháp học máy khác và giúp huấn luyện dễ dàng hơn vì kết quả của mô hình có thể được so sánh với kết quả được dán nhãn thực tế. Tuy nhiên, dữ liệu được gắn nhãn cần được chuẩn bị kỹ lưỡng và trong trường hợp không lý tưởng sẽ tạo ra một mô hình quá chặt chẽ với bộ dữ liệu huấn luyện và làm mô hình không thể xử lý chính xác với các dữ liệu mới.

**Học máy không giám sát:** Học máy không giám sát đầu vào là lượng lớn dữ liệu không được gắn nhãn và sử dụng các thuật toán để trích xuất các đặc trưng có ý nghĩa để gắn nhãn, sắp xếp và phân loại dữ liệu trong thời gian thực mà không cần sự can thiệp của con người. Học không giám sát không phải là việc tự động hóa các quyết định và dự đoán, mà thiên về việc xác định các mẫu và mối quan hệ trong dữ liệu mà con người sẽ bỏ lỡ. Ví dụ như tính năng phát hiện thư rác — con người tạo ra nhiều email hơn một nhóm các nhà khoa học có thể gắn nhãn hoặc phân loại trong vòng đời của họ. Một thuật toán học tập không giám sát có thể phân tích khối lượng lớn email và phát hiện ra các tính năng và mẫu biểu thị spam (và tiếp tục gắn cờ spam tốt hơn theo thời gian).

**Học tập bán giám sát:** Là sự kết hợp giữa học tập có giám sát và học tập không giám sát. Trong quá trình huấn luyện, thuật toán sử dụng một tập dữ liệu nhỏ có gán nhãn để phân loại và trích xuất tính năng từ một tập dữ liệu lớn hơn, không được gắn nhãn. Học bán giám sát có thể giải quyết vấn đề không có đủ dữ liệu được gắn nhãn (hoặc không đủ khả năng gắn nhãn đủ dữ liệu) để huấn luyện thuật toán học có giám sát.

### Lựa chọn đặc trưng

Phương pháp lựa chọn đặc trưng dùng để giảm số lượng đặc trưng đầu vào trước khi đưa vào mô hình dự đoán của học máy. Các mô hình dự đoán có quá nhiều đầu vào sẽ làm tiêu tốn bộ nhớ cũng như thời gian huấn luyện, đồng thời các đặc trưng không cần thiết cũng làm nhiễu khả năng dự đoán của mô hình. Việc lựa chọn đặc trưng này nhắm đến việc giảm thiểu thời gian tính toán của mô hình dự đoán, và trong một số trường hợp sẽ làm tăng chất lượng dự đoán của toàn bộ mô hình.



Các phương pháp lựa chọn dựa trên thống kê sẽ đánh giá sự ảnh hưởng của các đặc trưng đến giá trị dự đoán, sau đó lựa chọn những đặc trưng có ảnh hưởng lớn nhất. Tuy nhiên việc lựa chọn phương pháp phù hợp cũng cần nghiên cứu kỹ lưỡng cho từng yêu cầu bài toán đặt ra. Có 3 cách lựa chọn đặc trưng chính bao gồm:

* Tìm kiếm tập hợp các đặc trưng con tốt nhất từ tập các đặc trưng ban đầu. Thuật toán điển hình của phương pháp này là RFE (Recursive Feature Elimination).
* Tính toán sự tương đồng cũng như ảnh hưởng của từng đặc trưng với giá trị cần đạt được, sau đó lựa chọn ra những đặc trưng có sự tương đồng cũng như ảnh hưởng lớn nhất. Các thuật toán đặc trưng của phương pháp này có thể kể đến như Chi-Squared hay Pearson Correlation.
* Sử dụng các thuật toán có sẵn phương pháp lựa chọn đặc trưng như Lasso hay RandomForrest

#### Recursive Feature Elimination (RFE)

RFE là thuật toán lựa chọn đặc trưng kiểu đóng gói, bên trong bao gồm một giải thuật được sử dụng để đánh giá. Thuật toán RFE hoạt động với tất cả các đặc trưng của dữ liệu, đánh giá mức độ quan trọng củ từng đặc trưng, sau đó dần loại bỏ các đặc trưng ít quan trọng nhất, lặp lái quá trình trên cho tới khi đạt được số lượng đặc trưng mong muốn.

#### Pearson Correlation và Chi-Squared

Đây là 2 phương pháp đặc trưng của cách lựa chọn thứ 2, là tính toán sự ảnh hưởng của từng đặc trưng đến kết quả cần đạt bằng công thức của từng phương pháp

Sau khi tính toán, thuật toán sẽ lựa chọn những đặc trưng có mức độ quan trọng nhất đến kết quả cần đạt được.

### Phân lớp và ra quyết định nhận dạng

#### Support Vector Machine (SVM)

#### Random Forest (RF)

#### Linear Discriminant Analysis (LDA)

#### Decision Tree (DTs)

## Áp dụng với bộ dữ liệu MIT

# TỔ CHỨC DỮ LIỆU ĐẦU VÀO NÂNG CAO CHẤT LƯỢNG NHẬN DẠNG CẢM XÚC

## Thử nghiệm một số kỹ thuật nhận dạng lên bộ dữ liệu MIT

## Thay đổi tỉ lệ chồng chập

## Phân tích kết quả đạt được