NGHIÊN CỨU XÂY DỰNG MÔ HÌNH NHẬN DIỆN CẢM XÚC QUA GIỌNG NÓI

RESEARCH AND BUILD A MODEL SPEECH EMOTION RECOGNITION

Hà Huy Giáp[1,*], Nguyễn Quang Đại[2]

TÓM TẮT

Trong bối cảnh công nghệ số, dữ liệu lớn, đòi hỏi con người phải xử lý rất nhiều thông tin cùng một lúc. Bài toán đặt ra là các công ty cần phân loại, đánh giá các phản hồi khách hàng qua các đoạn tin nhắn thoại, nhận diện cảm xúc qua giọng nói cho các chatbot để có hướng xử lý tiếp theo, nhận diện cảm xúc qua giọng nói cho các robot con người, nhằm xử lý các hướng phản hồi tiếp. Bài báo đưa ra phương án tạo ra model để xác định, phân loại dữ liệu này. Nội dung của bài báo tập trung vào việc ứng dụng machine learning trong việc phân loại dữ liệu và sử dụng bài toán logistic regressiong để thiết lập Model, thiết lập Loss Function, tối ưu loss Function và dự đoán mô hình.

Từ khóa: *Logistic regression, hàm chi phí, học máy.*

ABSTRACT

In the context of digital technology, big data, it requires people to process a lot of information at the same time. The problem is that companies need to classify and evaluate customer feedback through voice messages, voice recognition for chatbots to have the next processing direction, and identify emotions through voice. speaking for the human robots, in order to handle the feedback directions. The article proposes a plan to create a model to identify and classify this data. The content of the paper focuses on the application of machine learning in data classification and using the logistic regressiong problem to set up the Model, set the Loss Function, optimize the loss function and predict the model.

Keywords: *Logistic regression, loss function, mechine learning.*

[1]Khoa Điện, Trường Đại học Kinh tế Kỹ thuật Công nghiệp

[2]Trường Đại học Công nghiệp Hà Nội

[*]Email: hhgiap@uneti.edu.vn

Ngày nhận bài: 06/4/2021

Ngày nhận bài sau phản biện: 06/5/2021

Ngày chấp nhận đăng: 25/6/2021

1. GIỚI THIỆU

Ngày nay, đã có những thay đổi rất lớn về cách thức con người trao đổi thông tin với hệ thống. Sự thay đổi này biểu hiện ở chỗ, các cách thức trao đổi thông tin đã được định dạng và có cấu trúc chặt chẽ được chuyển sang các cách thức linh hoạt và tự nhiên hơn. Trong đó, tiếng nói là cách thức trao đổi thông tin tự nhiên nhất, cho phép tương tác giữa con người với hệ thống nhanh và dễ dàng. Đối thoại dùng ngôn ngữ nói không chỉ đơn giản, thuận tiện và tiết kiệm thời gian mà còn góp phần đảm bảo khía cạnh an toàn trong những môi trường có tính rủi ro.

Để có thể thiết lập hệ thống tương tác có tính linh hoạt cao, kiến trúc của các hệ thống đối thoại người - máy cần được trang bị thêm các chức năng mới. Các chức năng này bao gồm nhận dạng cảm xúc tiếng nói, phát hiện các tham biến dựa trên tình huống cũng như trạng thái của người dùng và quản lý tình huống để đưa ra các mô hình dựa trên các tham biến đã được phát hiện làm cho quá trình đối thoại phù hợp. Chính vì vậy, trong nhiều năm qua, các nghiên cứu về cảm xúc tiếng nói đã thu hút mối quan tâm mạnh mẽ trong lĩnh vực tương tác người - máy và mong muốn tìm ra cách làm thế nào có thể tích hợp trạng thái cảm xúc của người nói vào hệ thống đối thoại người - máy dùng tiếng nói.

Với tính thiết thực của cảm xúc trong tiếng nói được áp dụng trong thực tế đang rất được quan tâm, mục tiêu chính của bài báo là nghiên cứu nhận dạng cảm xúc cho tiếng nói dựa trên phương diện xử lý tín hiệu tiếng nói. Bài báo trình bày nghiên cứu thử nghiệm và đề xuất mô hình nhận dạng cảm xúc cho tiếng nói dựa trên việc nghiên cứu đánh giá các tham số và so sánh một số mô hình nhận dạng. Bốn cảm xúc cơ bản sẽ được nghiên cứu bao gồm cảm xúc: vui, buồn, tức và bình thường.

2. ỨNG DỤNG BÀI TOÁN LOGISTIC REGRESSION TRONG NHẬN DIỆN CẢM XÚC QUA GIỌNG NÓI

2.1. Thiết lập Model

Gọi $x_1^{(i)}$, $x_2^{(i)}$, $x_3^{(i)}$ là các thông số đặc trưng cho giọng nói tương ứng là:

- $x_1^{(i)}$: Hệ số Coff của Cepstral với Mel filter MFCC (Mel Frequency Cepstral Coefficients):
- $x_2^{(i)}$: CHROMA: Sắc ký (12 classes)
- $x_3^{(i)}$: Mel (Melody): Giai điệu audio
- $\hat{y}_i$: Xác suất mà model dự đoán đúng cảm xúc của giọng nói thứ i

Thiết lập model cho bài toán nhận diện cảm xúc qua giọng nói. Sử dụng công thức của logistic regression ta được:

$$\hat{y}_i = \sigma(w_0 + w_1.x_1^{(i)} + w_2.x_2^{(i)} + w_3.x_3^{(i)}) = \frac{1}{1+e^{-(w_0+w_1.x_1^{(i)}+w_2.x_2^{(i)}+w_3.x_3^{(i)})}} \quad (1)$$

2.2. Thiết lập Loss Function

Ta cần một hàm để đánh giá mức độ chính xác của model. Nếu $\hat{y}_i$ càng gần với xác suất thực tế y thì càng tốt.

- Nếu giọng nói thứ i đúng với cảm xúc dự đoán tức là y = 1 thì ta mong muốn $\hat{y}_i$ càng gần 1 thì càng tốt.

- Nếu giọng nói thứ i không đúng với cảm xúc dự đoán tức là y = 0 thì ta mong muốn $\hat{y}_i$ càng gần 0 càng tốt.

Với mỗi giọng nói $(x^{(i)}, y_i)$ ta sử dụng hàm loss function binary_crossentropy để đánh giá hiệu qủa của model.

$$L = -(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i)) \quad (2)$$

Ta thấy rằng:

• Hàm L tăng dần từ 0 đến 1

• Khi model dự đoán $\hat{y}_i$ gần 0, tức giá trị dự đoán gần với giá trị thật y_i thì L nhỏ, xấp xỉ 0

• Khi model dự đoán $\hat{y}_i$ gần 1, tức giá trị dự đoán ngược lại giá trị thật y_i thì L rất lớn.

Hàm L nhỏ khi giá trị model dự đoán gần với giá trị thật và rất lớn khi model dự đoán sai, hay nói cách khác L càng nhỏ thì model dự đoán càng gần với giá trị thật → Bài toán tìm model trở thành tìm giá trị nhỏ nhất của L.

Hàm loss function trên toàn bộ N dữ liệu:

$$J = -\frac{1}{N}.\sum_{i=1}^{N}(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i)) \quad (3)$$

2.3. Tối ưu Loss Function

Để áp dụng thuật toán gradient descent tìm tối ưu loss function mình cần tính đạo hàm của loss function với w.

Với mỗi điểm $(x^{(i)}, y_i)$, gọi hàm loss function.

$$L = -(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i))$$

Trong đó:

$\hat{y}_i = \sigma(w_0 + w_1.x_1^{(i)} + w_2.x_2^{(i)} + w_3.x_3^{(i)}) = \sigma(z)$ là giá trị của model dự đoán.

y_i là giá trị thật của dữ liệu.

$$\frac{dL}{dw_0} = \frac{dL}{\hat{y}_i}.\frac{d\hat{y}_i}{w_0}$$

$$\frac{dL}{\hat{y}_i} = -\frac{d(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i))}{d\hat{y}_i} = \frac{\hat{y}_i - y_i}{\hat{y}_i.(1-\hat{y}_i)}$$

$$\frac{d\hat{y}_i}{w_0} = \frac{d(\sigma(w_0 + w_1.x_1^{(i)} + w_2.x_2^{(i)} + w_3.x_3^{(i)}))}{dw_0} = \hat{y}_i.(1-\hat{y}_i)$$

$$\frac{dL}{dw_0} = \frac{dL}{\hat{y}_i}.\frac{d\hat{y}_i}{w_0} = \hat{y}_i - y_i$$

Tương tự:

$$\frac{dL}{dw_1} = x_1^{(i)}.(\hat{y}_i - y_i)$$

$$\frac{dL}{dw_2} = x_2^{(i)}.(\hat{y}_i - y_i) \quad (4)$$

$$\frac{dL}{dw_3} = x_3^{(i)}.(\hat{y}_i - y_i)$$

Xét trên toàn bộ dữ liệu:

$$\frac{dL}{dw_0} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

$$\frac{dL}{dw_1} = \frac{1}{N}\sum_{i=1}^{N}x_1^{(i)}.(\hat{y}_i - y_i)$$

$$\frac{dL}{dw_2} = \frac{1}{N}\sum_{i=1}^{N}x_2^{(i)}.(\hat{y}_i - y_i)$$

$$\frac{dL}{dw_3} = \frac{1}{N}\sum_{i=1}^{N}x_3^{(i)}.(\hat{y}_i - y_i)$$

Biểu diễn dưới dạng ma trận

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} \\ . & . & . & . \\ 1 & x_1^{(n)} & x_2^{(n)} & x_3^{(n)} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}, w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} \quad (5)$$

$$J = -\frac{1}{N}.\sum_{i=1}^{N}(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i)) \quad (6)$$

$$\frac{dJ}{dw} = \frac{1}{N}.X^T.(\hat{y} - y) \quad (7)$$

Thuật toán **Gradient descent** là thuật toán tìm giá trị nhỏ nhất của hàm số f(x) dựa trên đạo hàm. Thuật toán thực hiện theo các bước:

Bước 1. Khởi tạo giá trị $w = w_0$ tùy ý.

Bước 2. Gán w = w - learning_rate * J'(w) (learning_rate là hằng số dương ví dụ learning_rate = 0.001).

Bước 3. Tính lại J(w): Nếu J(w) đủ nhỏ thì dừng lại, ngược lại tiếp tục bước 2.

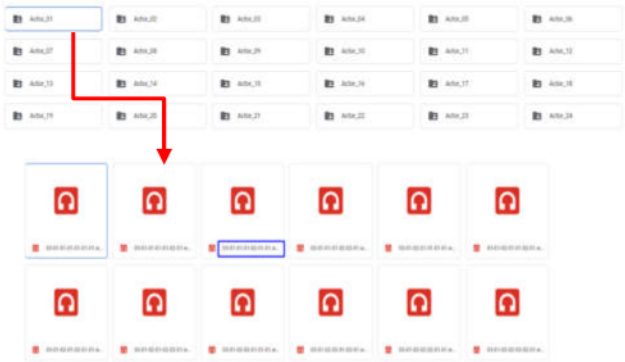
Sau khi thực hiện thuật toán gradient descent ta sẽ tìm được w_0, w_1, w_2, w_3. Với mỗi giọng nói ta sẽ tính được thông số cảm xúc của giọng nói đó $\hat{y}_i = \sigma(w_0 + w_1.x_1^{(i)} + w_2.x_2^{(i)} + w_3.x_3^{(i)})$ sau đó so sánh với thông số cảm xúc đặc trưng (Labels) từ đó ta sẽ biết được cảm xúc của going nói cần chuẩn đoán.

3. THỰC HIỆN THUẬT TOÁN VỚI ỨNG DỤNG GOOGLE COLLABORATORY

3.1. Xử lý dữ liệu (Dataset)

Dữ liệu ở dạng Audio*.wav có gán sẵn các labels như hình 1.

Dữ liệu gồm có: 24 Actors, 1536 files audio, 8 emotions (calm, happy, fearful, disgust, sad, angry, supprised).

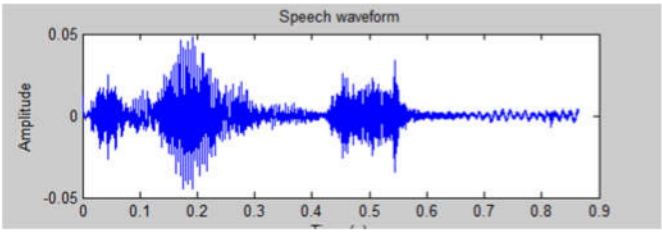


Hình 1. Dữ liệu ở dạng Audio*.wav có gán sẵn các labels

3.2. Số hóa và lấy ra các đặc trưng của dữ liệu

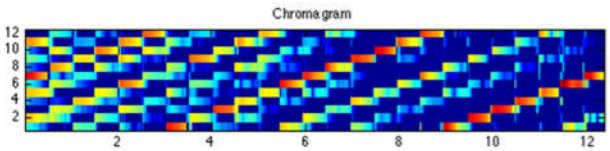
Gọi $x_1^{(i)}$, $x_2^{(i)}$, $x_3^{(i)}$ là các thông số đặc trưng cho giọng nói tương ứng là:

$x_1^{(i)}$: Hệ số Coff của Cepstral với Mel filter MFCC (Mel Frequency Cepstral Coefficients) (hình 2).



Hình 2. Đồ thị sóng âm

$x_2^{(i)}$: CHROMA: Sắc ký (12 classes) (hình 3).



Hình 3. Hình ảnh sắc ký âm thanh

$x_3^{(i)}$: Mel (Melody): Giai điệu audio

Kết quả là 1 vector đặc trưng shape = (180,)

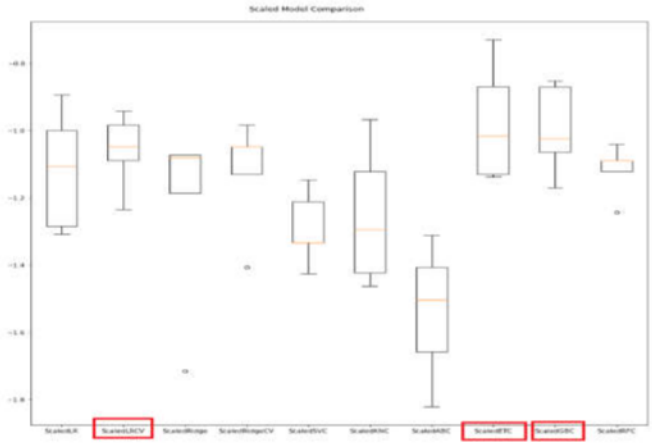
Chia Dataset: Train set (80%), Validation (10%), Test set (10%)

Số hóa Labels:

```
# All emotions in the dataset
emotions={
 '01':'neutral',
 '02':'calm',
 '03':'happy',
 '04':'sad',
 '05':'angry',
 '06':'fearful',
 '07':'disgust',
 '08':'surprised'
 }
# Observed_Emotion
observed_emotions=['calm', 'happy', 'fearful', 'disgust']
```

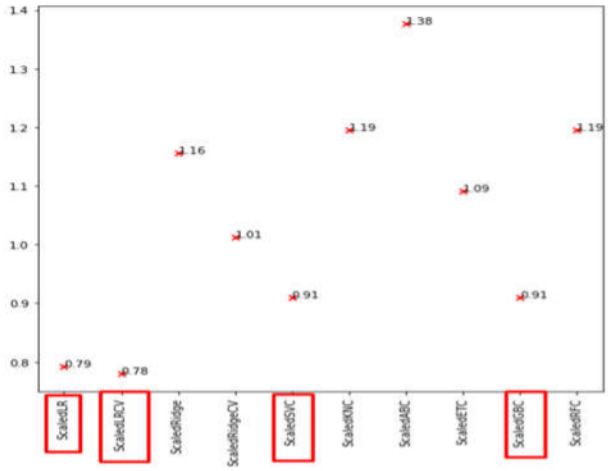
3.3. Sử dụng các model Logistic Regression trong thư viện Sklearn để train và evaluate data

Kết quả Train data như hình 4.



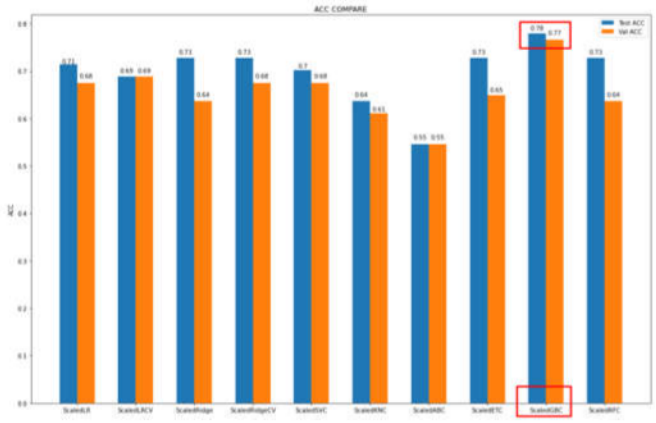
Hình 4. Kết quả với Train data

Kết quả với Validation data như hình 5.



Hình 5. Kết quả Train với Validation data

Kết quả so sánh các model như hình 6.



Hình 6. Kết quả so sánh các model

Chọn Model: GradientBoostingClassifier:

- Không Overfit
- Không Underfit

3.4. Điều chỉnh HYPER-PARAMETERS

Điều chỉnh hệ số "n_estimators":

Chương trình:

```
# SearchCV
from sklearn.ensemble import GradientBoostingClassifier
scaler = StandardScaler().fit(x_train)
rescaledX = scaler.transform(x_train)
para_grid = {'n_estimators' : [50,100,300,400,500]}
model = GradientBoostingClassifier(random_state= 8)
grid = GridSearchCV(estimator= model, param_grid= para_grid, scoring= 'neg_mean_squared_error')
grid_result = grid.fit(rescaledX, y_train)
```

Kết quả:

```
Best: -0.945862 using {'n_estimators': 300}
-1.102439 (0.151368) with: {'n_estimators': 50}
-0.994962 (0.126236) with: {'n_estimators': 100}
-0.945862 (0.230431) with: {'n_estimators': 300}
-0.949167 (0.220215) with: {'n_estimators': 400}
-0.949167 (0.220215) with: {'n_estimators': 500}
```

Chọn: *'n_estimators': 400*

Điều chỉnh hệ số "learning_rate":

Chương trình:

```
from sklearn.ensemble import GradientBoostingClassifier
scaler = StandardScaler().fit(x_train)
rescaledX = scaler.transform(x_train)
para_grid = {'learning_rate' : [0.1,0.01,0.001]}
model = GradientBoostingClassifier(random_state= 8)
grid = GridSearchCV(estimator= model, param_grid= para_grid, scoring= 'neg_mean_squared_error')
grid_result = grid.fit(rescaledX, y_train)
```

Kết quả:

```
Best: -0.994962 using {'learning_rate': 0.1}
-0.994962 (0.126236) with: {'learning_rate': 0.1}
-1.470585 (0.185699) with: {'learning_rate': 0.01}
-1.913621 (0.225609) with: {'learning_rate': 0.001}
```

Chọn: *'learning_rate': 0.1*

3.5. Build Model

Xây dựng Logistic Model với "Best Model" và "Best Hyper-para":

- GradientBoostingClassifier
- n_estimators: 400
- learning_rate: 0.1

4. ĐÁNH GIÁ MSE VÀ ACCURACY QUA BỘ TEST

Chương trình:

```
# prepare the model
scaler = StandardScaler().fit(x_train)
rescaledX = scaler.transform(x_train)
model = GradientBoostingClassifier(random_state=8, n_estimators=400, learning_rate = 0.1)
model.fit(rescaledX, y_train)
# Checking the accuracy with test data
rescaledX_test = scaler.transform(x_test)
predictions = model.predict(rescaledX_test)
```

Kết quả:

```
MSE: 0.42857142857142855
Accuracy: 0.8181818181818182
```

Đánh giá:

- Mô hình Logistic Regression hoạt động tốt với Acc = 81,81%.

- Kết quả chứng minh LR đôi khi hoạt động tốt hơn 1 mạng Neutral Network đơn giản, Ví dụ trong paper mẫu dùng mạng "MLPClassifier" mà Acc = 72,40%.

5. KẾT LUẬN

Bài báo trình bày về bài toán nhận diện cảm xúc qua giọng nói. Trong bài báo đã đặt ra vấn đề và tầm quan trọng của việc nhận diện được cảm xúc qua giọng nói từ đó đưa ra phương án giải quyết.Các bước để thực hiện được thể hiện rất rõ từ việc xử lý dữ liệu, thiết lập model, thiết lập loss function, tối ưu loss function và dự đoán mô hình. Kết quả của phương pháp cũng đạt được độ chính xác cao so với một số phương pháp đã làm trước đó.

Hướng phát triển tiếp theo của nghiên cứu là tối ưu thuật toán nhận diện cảm xúc qua giọng nói và có thể phát triển trên nhiều ứng dụng thực tế như chatbot, nhận diện cảm xúc qua giọng nói cho robot con người nhằm xử lý các hướng phản hồi tiếp theo, đánh giá các phản hồi của khách hàng qua các đoạn tin nhắn thoại. Từ đó mang lại hiệu quả ứng dụng và hiệu quả kinh tế cho người dùng.

TÀI LIỆU THAM KHẢO

[1]. Rao, K. Sreenivasa, Koolagudi, Shashidhar G., 2013. *Emotion Recognition using Speech Features*. Springer.

[2]. Cowie, Roddy, et al.,. 2001. *Emotion recognition in human-computer interaction.* IEEE Signal processing magazine 18.1, vol. 12, pp. 32–80.

[3]. Robert Plutchik, Henry Kellerman, 1989. *Emotion: Theory, research and experience.* New York, USA: Academic Press.

[4]. Ayadi M. E., Kamel M. S., Karray F., 2011. *Survey on speech emotion recognition: Features,classication schemes, and databases.* Pattern Recognition, vol. 44, pp. 572–587

[5]. Craig A. D., 2009. *Handbook of Emotion, ch. Interoception and emotion: A neuroanatomical perspective.* New York: September: The Guildford Press, ISBN 978-1-59385-650-2.

[6]. Ekman P., 1999. *Handbook of Cognition and Emotion:* ch. Basic Emotions, Sussex, UK: JohnWiley and Sons Ltd.

AUTHORS INFORMATION

Ha Huy Giap[1], Nguyen Quang Dai[2]

[1]Faculty of Electrical Engineering, University of Economics - Technology for Industries

[2]Hanoi University of Industry