Nhân diện Covid-19 qua tiếng ho

Nguyễn Thành Trung[a, c], Dương Văn Bình[b, d], Nguyễn Thanh Trung[a, c]

[a]Đại học Bách Khoa – Đại học Quốc gia Thành phố Hồ Chí Minh

[b]Đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh

[c]{trung.nguyendx, trung.nguyen1408}@hcmut.edu.vn

[d]18520505@gm.uit.edu.vn

Ngày 30 tháng 08 năm 2021

Tóm tắt

Cuộc thi "AICV-115M Challenge"[1] là một cuộc thi về nhận diện Covid19 qua tiếng ho với tổng giải thưởng 115 triệu VND với 168 cá nhân tham gia. Chúng tôi - nhóm "đi thi" đã đạt được Hạng 3 với điểm (tính bằng AUC) là 0.92 (chính xác hơn là 0.921527). Giải pháp mà chúng tôi triển khai đó là trích lọc những đặc trưng: tỷ lệ giao nhau bằng không, MFCC, Sắc độ và Quang phổ Mel của những tín hiệu âm thanh. Sau đó huấn luyện mô hình Light Gradient Boosting Machine để xử lý. Chúng tôi mong báo cáo kỹ thuật này sẽ cung cấp một cái nhìn rõ ràng hơn về giải pháp mà chúng tôi đã triển khai, cũng như là một sự tham khảo cho những người muốn quan tâm.

Toàn bộ mã nguồn cũng như các tệp tin liên quan có thể được tìm thấy tại: https://github.com/dee-ex/aicovidvn115m.

1. Đặt vấn đề

Tại thời điểm viết báo cáo này, dịch bệnh COVID-19 ở Việt Nam đang diễn biến rất phức tạp. Số ca nhiễm mỗi ngày luôn được tính ở đơn vị ngàn và vẫn chưa có dấu hiệu giảm. Một trong điều then chốt trong công cuộc phòng và chống dịch hiện nay đó là việc xét nghiệm cho người dân. Cụ thể, để được xét nghiệm nhanh kháng nguyên (hay còn gọi là kiểm tra nhanh COVID-19), người dân đến các bệnh viện và trung tâm ý tế ở quân, huyện, thành phố. Nếu muốn xét nghiệm khẳng định SARS-CoV-2 (hay còn gọi là phương pháp xét nghiệm RT-PCR) thì đến các phòng xét nghiệm của các đơn vị được Bộ Y tế cho phép xét nghiệm khẳng định SARS-CoV-2. Về mức giá xét nghiệm, Sở Y tế cho biết, xét nghiệm bằng phương pháp RT-PCR bằng mức giá thanh toán bảo hiểm y tế theo quy định tại Công văn 4356/2020 của Bộ Y tế về việc hướng dẫn mức giá thanh toán chi phí thực hiện xét nghiệm COVID-19. Xét nghiệm nhanh kháng nguyên: bằng giá dịch vụ được quy định tại Phụ lục đính kèm Thông tư số 13/2019 và Thông tư số 14/2014 của Bộ Y tế là 238.000 đồng/mẫu [1]

[1] AICV-115 Challenge, đường dẫn: https://aihub.vn/competitions/22#learn_the_details.

Với tình hình hiện tại, việc xét nghiệm không chỉ khó khăn về tài chính mà còn là sự thiếu hụt về dụng cụ. Liệu có cách nào để thực hiện việc này hiệu quả hơn? Các nhà nghiên cứu trên thế giới đã chỉ ra rằng có thể phát hiện COVID-19 qua tiếng ho, qua điện thoại di động bằng trí tuệ nhân tạo (AI) với độ chính xác trên 90% [2]. Trên những cơ sở đó, dự án cộng đồng AICovidVN tổ chức cuộc thi ``AICV-115M Challenge'' kêu gọi cộng đồng tham gia xây dựng các giải pháp trí tuệ nhân tạo để thu âm tiếng ho của người dân qua tổng đài điện thoại, rồi đưa ra chẩn đoán COVID-19 sơ bộ. Chúng tôi nhận thấy, đây là một cơ hội tốt để góp một chút sức lực nhỏ cho xã hội, và cũng là để học hỏi thêm những kiến thức về mảng trí tuệ nhân tạo nên đã đăng ký tham dự cuộc thi.

2. Tổng quan lý thuyết

Trong bài báo được xuất bản trên IEEE Journal of Engineering in Medicine and Biology [3], những nhà nghiên cứu đã huấn luyện một mô hình với khoảng một vạn mẫu tiếng ho. Mô hình này khi đem đi thử nghiệm với những tiếng ho của người đã được xác nhận là dương tính với COVID-19 đã cho kết quả nhạy đến 98.5% và độ đặc hiệu 94.2%. Nhóm nghiên cứu trên cũng đã và đang làm việc để kết hợp mô hình này vào một ứng dụng thân thiện với người dùng. Nếu được FDA chấp thuận và áp dụng trên quy mô lớn thì có thể là một công cụ sàng lọc trước miễn phí, tiện lợi để xác định những người có khả năng dương tính với COVID-19.

Theo như chúng tôi tìm hiểu, chưa thực sự có một cơ sở khoa học quá rõ ràng nào cho điều này, nhưng những kết quả đã đạt được từ [3] cho thấy phương pháp nhận diện COVID-19 qua tiếng ho là chuyện rất khả thi. Báo cáo này sẽ đóng góp thêm những kết quả về việc áp dụng học máy trong việc nhận diện người nhiễm COVID-19 qua tiếng ho ở Việt Nam.

3. Phương pháp

3.1 Phân tích dữ liệu

Trong vòng đề đích mà chúng tôi tham gia, ban tổ chức (BTC) đã thu thập và công bố một tập dữ liệu đã được dán nhán gồm 4,504 mẫu dữ liệu, một tập dữ liệu thêm gồm 1,195 mẫu và số mẫu được dán nhãn là 30, kèm thêm một tập dữ liệu gồm 1627 mẫu chưa được dán nhãn để nộp tính điểm cho cuộc thi.

Để miêu tả về tập dữ liệu, mỗi mẫu dữ liệu trong đó là một tệp tin âm thanh có 1 cho tới 4 tiếng ho. Dù BTC đã rất cố gắng, nhưng vẫn tồn tại những một số mẫu không có tiếng hoặc tiếng không rõ. Sau khi loại bỏ những mẫu không đạt yêu cầu nhờ thông tin được cuộc thi cung cấp, thêm vào đó là lấy thêm 10 mẫu có nhãn dương tính ở tập dữ liệu thêm, số lượng mẫu mà chúng tôi có được là 4,068. Trong đó, có *3,399* mẫu được dán nhãn là âm tính và *669* được dán nhán là dương tính, tỷ lệ giữa hai nhãn khoảng 8/2 (chính xác hơn là 8.4/1.6).

Cũng giống như nhiều tập dữ liệu về y tế, tập dữ liệu ta có là một tập không cân bằng giữa hai nhãn, dẫn tới việc xây dựng và đánh giá mô hình sẽ khó khăn hơn đôi chút so với những tập dữ liệu cân bằng [4]. Vì thế, chúng tôi đã cân bằng dữ liệu bằng phương pháp SMOTE [5] (cụ thể là SVM SMOTE) để có thể giúp cho số lượng mẫu dương tính bằng với mẫu âm tính trong quá trình học của mô hình.

3.2 Trích xuất đặc trưng

So với dữ liệu hình ảnh - một loại dữ liệu quen thuộc với đại đa số những người làm về học máy/học sâu, âm thanh là một kiểu dữ liệu khá đặc biệt. Nếu mỗi điểm ảnh của một bức hình được coi là một đặc trưng, ta có thể gom tất cả các điểm ảnh có được thành một véc-tơ đầu vào và trong rất nhiều trường hợp chuyện này rất đỗi bình thường và cũng không kém phần hiệu quả. Còn âm thanh thì không đơn giản như vậy, dữ liệu ta có là một mảng dài với mỗi phần tử là biên độ của sóng được lấy mẫu tại một thời điểm nhất định. Tần số lấy mẫu, tức nghịch đảo của khoảng thời gian để ta ghi lại biên độ giữa lần thứ t và $t+1$ thường sẽ là 44,1 KHz (chất lượng tốt) [6], còn dữ liệu mà chúng tôi có được là 22,05 KHz (bằng một nửa của 44,1 KHz). Việc đưa tất cả những giá trị có được của một tệp âm thanh vào một mô hình học máy/học sâu trực tiếp để nhận diện cấu trúc thường không cho ra những kết quả khả quan. Thay vào đó, một số những phương pháp trích xuất đặc trưng quan trọng của âm thanh đã được áp dụng.

Chúng tôi đã thử nghiệm qua một số những đặc trưng thường được sử dụng, tuy vậy không phải đặc trưng nào cũng đóng góp được nhiều và tích cực cho bài toán này. Có 3 đặc trưng ảnh hưởng lớn tới mô hình của chúng tôi, đó là: **MFCCs** (**M**el-**f**requency **c**epstral **c**oefficient**s**) (13), **Sắc độ** (Chroma), và **Quang phổ Mel** (Mel spectrogram) (128). Trong những giờ cuối của cuộc thi, chúng tôi cũng đã thử thêm đặc trưng **tỷ lệ giao nhau bằng không** (zero-crossing rate), tuy vậy không cải thiện được nhiều (xấp xỉ 0.1% AUC). Về phần trích lọc các đặc trưng này trong quá trình lập trình, thư viện chúng tôi sử dụng để xử lý tác vụ này là librosa[2] - một thư viện mạnh mẽ để xử lý tín hiệu âm thanh.

3.3 Tiền xử lý dữ liệu

Mỗi mẫu dữ liệu đầu vào sẽ được trích lọc các đặc trưng: MFCCs - $\mathbf{X}_{\text{MFCCs}}$, Sắc độ - $\mathbf{X}_{\text{Chroma}}$, và quang phổ Mel - $\mathbf{X}_{\text{Mel}}$.

Mỗi đặc trưng này đều là một mảng 2D với số chiều khác nhau, và mỗi dữ liệu lại có thể cho cùng một đặc trưng với số chiều cũng khác nhau. Chúng tôi đã tiến hành lấy trung bình mỗi cột trong mảng 2D đưa về thành một véc-tơ (dạng dữ liệu 1D). Ví dụ như với MFCCs, thì ta sẽ có $\mathbf{x} = \mathbb{E}\{(\mathbf{X}_{\text{MFCCs}})_{,*}\}$. Những đặc trưng khác được tiến hành tương tự như đặc trưng quang phổ MFCCs.

Lý do để chúng tôi lấy trung bình các cột của các đặc trưng là để kết hợp những đặc trưng này lại với nhau. Một lí do nữa là vì số chiều 2D của các đặc trưng khá khác biệt

[2] Thư viện librosa, đường dẫn: https://librosa.org/doc/latest/index.html.

để có thể điều chỉnh và lựa chọn. Dẫu biết phương pháp này sẽ có khả năng làm mất đi thông tin rất nhiều nhưng may mắn thay, kết quả chúng tôi có được không phải là quá tệ.

Tóm lại, mỗi mẫu dữ liệu đầu vào sẽ cho một vector đặc trưng:

$$\mathbf{x} = [\mu_{\text{ZCR}}; \mathbf{X}_{\text{MFCCs}}; \mathbf{X}_{\text{Chroma}}; \mathbf{X}_{\text{Mel}}] \in \mathbb{R}^{154}$$

Trong đó, μ_{ZCR} là trung bình của đặc trưng tỷ lệ giao nhau bằng không.

Độ lệch chuẩn của các đặc trưng cũng cách biệt nhau khá lớn nên chúng tôi đã **chuyển khoảng giá trị** (khoảng $[0,1]$) theo phương trình sau:

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

3.4 Mô hình phân loại

Mô hình được sử dụng là Light GBM - **Light G**radient **B**oosting **M**achine, là một mô hình Gradient Boosting phân tán dành cho học máy do Microsoft phát triển. Nhóm không tập trung nhiều vào tối ưu các tham số của mô hình, nên các cấu hình khá đơn gian, và cụ thể là:

```
"objective": "binary", "boosting_type": "gbdt", "metric" :
"auc", "learning_rate": 0.03, "subsample": 0.68, "tree_learner":
"serial", "colsample_bytree": 0.28, "early_stopping_rounds":
100, "subsample_freq": 1, "reg_lambda": 2, "reg_alpha": 1,
"num_leaves": 500 + seed * 25
```

Chúng tôi sử dụng xác thực chéo (k-fold cross validation) và chọn $k = 10$ dựa vào kinh nghiệm những nhà nghiên cứu đi trước. Kết quả đầu ra của phiên xác thực chéo sẽ được tính theo kiểu bầu chọn mềm (soft voting) [7], tức là giá trị đầu ra sau mỗi phiên xác thực chéo là $y = \sum_{i=1}^{10} y^{(i)}$, trong đó $y^{(i)}$ là kết quả của lần xác thực thứ i.

Không dừng lại ở đó, chúng tôi còn lồng các phiên xác thực chéo lại với nhau (nested cross validation) để cho ra kết quả cuối cùng. Trong thực nghiệm, chúng tôi triển khai 10 phiên xác thực chéo đánh số thứ tự 0 tới 9, ở mỗi phiên này, giá trị cài đặt `seed` sẽ bằng số thứ tự phiên bội với 25.

Nói một cách rõ ràng hơn, giá trị cài đặt `num_leaves` sẽ lần lượt nhận các giá trị 500, 525, ..., 700, 725 qua các phiên. Chúng tôi cũng tiếp tục sử dụng bầu chọn mềm giữa các phiên, do đó kết quả cuối cùng sẽ là trung bình của 10 phiên.

Quá trình huấn luyện được chúng tôi cài đặt và thực thi trên môi trường Google Colab[3] bản miễn phí với thời gian khoảng hơn *16 phút* (không tính phần trích xuất và chuẩn hoá đầu vào).

4. Đánh giá kết quả

Khi chưa sử dụng đặc trưng tỷ lệ giao nhau bằng không, AUC chúng tôi nhận được từ BTC là **0.92093**. Sau khi sử dụng đặc trưng trên, kết quả có cải thiện lên thành **0.921527**, và đây là kết quả mà chúng tôi đã nộp để đạt được **hạng 3** chung cuộc.

Chúng tôi cũng đã cố gắng tinh chỉnh tham số mô hình, nhưng vì chưa có nhiều kinh nghiệm với mô hình Light GBM nên kết quả không thay đổi.

Nhìn chung, mô hình chúng tôi không quá phức tạp, nên việc chạy dự đoán (tính cả phần trích lọc đặc trưng) cho duy nhất một tệp âm thanh chỉ trong vòng vài giây, với phần cứng CPU ở mức trung bình mặt bằng chung hiện nay.

Về phần cải thiện, nhóm chúng tôi đề xuất nên trích lọc đặc trưng một cách kĩ càng hơn khi dữ liệu còn ở dạng 2D bằng cách bộ lọc tích chập, thay vì chỉ gộp lại bằng cách tính toán kỳ vọng.

5. Kết luận

Thông qua báo cáo kỹ thuật này, nhóm "đi thi" chúng tôi mong rằng đã trình bày một cách kỹ càng hơn về giải pháp giúp đạt hạng 3 với AUC 0.92 trong vòng Về đích của cuộc thi. Mong rằng giải pháp của nhóm có thể khả thi để áp dụng trong môi trường thực tế.

Tài liệu tham khảo

[1] K. Vân, "Địa chỉ và mức giá xét nghiệm COVID-19 cho người dân ra khỏi TP.HCM." Bộ Y tế, 07 July 2021 https://ncov.moh.gov.vn/en/-/6847426-5408 (2021).

[2] Chu, J., "Artificial intelligence model detects asymptomatic Covid-19 infections through cellphone-recorded coughs." MIT News, 29 October 2020 https://news.mit.edu/2020/covid-19-cough-cellphone-detection-1029 (2020).

[3] Laguarta, J., Hueto, F., and Subirana, B., "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology* **1**, 275-281 (2020).

[4] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F., [*Learning from imbalanced data sets*], vol. 10, Springer (2018).

[3] Google Colab, đường dẫn: https://colab.research.google.com/.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., “Smote: Synthetic minority oversampling technique,” *Journal of Artificial Intelligence Research* **16**, 321-357 (Jun 2002).

[6] Devi, R. and Pugazhenthi, D., “Ideal sampling rate to reduce distortion in audio steganography,” *ProcediaComputer Science* **85**, 418-424 (2016). International Conference on Computational Modelling and Security (CMS 2016).

[7] Islam, R. and Shahjalal, M., “Soft voting-based ensemble approach to predict early stage drc violations,” 1081-1084 (August 2019).