**AI CUP 2023 春季賽**

附件

**多模態病理嗓音分類競賽報告**

隊伍：TEAM\_3680

隊員：江前昱(隊長)、蘇義新

Private leaderboard：0.621502 / Rank 3

**壹、環境**

作業系統：Windows 10

語言：Anaconda 4.10.1、Python 3.8.8

開發環境(IDE)：Jupyter notebook

套件：pandas 1.2.4、numpy 1.22.4、keras 2.12.0、scikit-learn 0.24.1、xgboost 1.6.2、catboost 1.0.6、librosa 0.10.0.post2、tensorflow 2.12.0、matplotlib 3.3.4、seaborn 0.11.1

**貳、演算方法與模型架構**

本任務目標為透過嗓音訊號結合病史紀錄偵測喉部病徵並分類。我們先將動態聲音及靜態文字分開處理，分別獲取資料特徵，再將兩者資料特徵合併做為訓練模型所需之X,y(如下圖所示):

病史紀錄

特徵工程

缺失值處理

嗓音訊號

梅爾倒頻譜

(MFCC)

音訊長度處理

建立模型及訓練

合併特徵做為X輸入

輸入訓練資料

分類結果y

XGBClassifier

CatBoostClassifier

SVC

Ensemble(Voting)

產出預測結果

模型部份我們個別訓練了決策樹、隨機森林、多層感知機(MLP) 、OneVsRestClassifier(linearSVC)、XGBClassifier、CatBoostClassifier、SVC多種模型，其中前四者(決策樹、隨機森林、多層感知機、OneVsRestClassifier)之訓練Recall值介於0.3~0.5，XGBClassifier、CatBoostClassifier、SVC之Recall值則可達0.5~0.6。同時為提升模型泛用性，我們將Recall值較高之三個模型(XGBClassifier、CatBoostClassifier、SVC)進行Ensemble Averaging，採投票(voting)方式產出預測結果，此模型也為我們最終使用之模型。

**參、創新性**

過程困難點及採取做法：

* 資料類別不平衡之問題：採用調整資料個別權重的做法(套用sklearn中compute\_sample\_weight)，
* 模型泛用性：我們沒有選擇表現最佳之模型CatBoostClassifier，而是採用多模型Ensemble Averaging，此方式雖於訓練當下分數是下降的，但於Private、Public資料中經實際測試其對分數是有所幫助的。
* 文字資料特徵處理：將欄位coding按程度高低進行排列，程度越高數字越大(ex:欄位[Occupational vocal demand]中數值4代表不需要，程度為最輕，經轉換重新定義為1; 原本數值1代表總是需要，程度為最高，經轉換為4)
* 音訊特徵工程處理：我們有嘗試針對音訊資料做autoencoder並取出中間層神經元作為訓練特徵。也有嘗試用多層感知機(MLP)預訓練音訊資料並將最後一層神經元取出當作音訊特徵。雖然經上線測試後分數不高，有overfitting的現象最終沒有採用，但可作為日後做法之參考

**肆、資料處理**

* 嗓音訊號

1. 音訊長度不同，用固定間隔採樣方式將訊號長度統一
2. 進行梅爾頻率倒譜分析，取得各時間區段內之mfcc
3. 將各時間區段內之mfcc分別進行加總，此加總數值做為音訊特徵值

* 病史紀錄

1. 缺失值處理：欄位[PPD]缺失值占總筆數80%以上故刪除此欄位
2. 缺失值處理：欄位[Voice handicap index]，用平均數進行補值
3. 欄位[Drinking]和[frequency]有高度相關性(相關係數>0.75)，處理方式如下:
   * 1. 挑選出[frequency]>1之row
     2. 將挑選出row中之[Drinking]欄位數值設為3
     3. [Drinking]欄位之新coding為[0/1/2/3]對應說明為[從未喝酒/已戒酒/有喝酒/常常喝酒]
4. 考量電子菸之影響程度無法與一般抽菸做比較，故不區分電子菸及一般抽菸將[Smoking]欄位中之數值3轉為數值2
5. 調整欄位[Sex]數值，1->0、2->1
6. 刪除欄位[Onset of dysphonia ]、[Diurnal pattern]
7. 調整欄位[Occupational vocal demand]，需求越高對應數值越高(1轉為4,2轉為3,3轉為2,4轉為1)

* 將上述兩者取得之特徵合併，進行StandardScaler標準化處理
* 因各類別數量差異大有資料不平衡之情況，故呼叫sklearn中compute\_sample\_weight方法，對於不同類別給予不同權重再進行訓練

**伍、訓練方式**

訓練方式為先廣泛地針對各種模型進行訓練，獲取初步表現較好之模型，再分別進行細部調參，最終確認模型間的搭配方式

step1.逐一針對不同模型進行訓練，cross validation設為10，用recall值來進行模型篩選，最終XGBClassifier、CatBoostClassifier、SVC模型表現較好，分別進行細部調參

step2.採用GridSearchCV進行細部調參，呼叫best\_params\_即可查看最佳參數，結果如下：

* XGBClassifier最佳參數
  + learning\_rate: 0.02
  + max\_depth: 2
  + n\_estimators: 300
* CatBoostClassifier最佳參數:
  + depth: 3
  + iterations: 170
  + l2\_leaf\_reg: 5
  + learning\_rate: 0.08
* SVC最佳參數:
  + C: 1.3
  + kernel: rbf

step3.各模型針對cross validation(cv)進行調參，呼叫best\_estimator\_取得最佳模型並儲存

**陸、分析與結論**

* 下圖為採用不同模型之分析結果，可觀察到CatBoostClassifier模型分數最高，但實際上線在public、private資料中，ensemble模型表現反而較好，表示ensemble之組合方式有對模型泛用性產生正面影響。

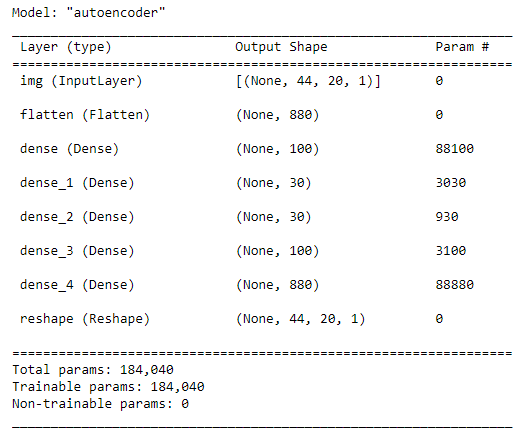
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Validation | Public | Private |
| MLP | 0.424 | 0.4413 | x |
| CatBoostClassifier | 0.535 | 0.5848 | 0.5872 |
| Ensemble(XGBClassifier、CatBoostClassifier、SVC) | 0.509(為三者平均) | 0.6195 | 0.6215 |

* 針對嗓音訊號之梅爾頻率倒譜分析設置不同n\_fft, hop\_length，測試不同的幀長及幀移對預測結果是否有影響，下圖為測試結果:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model**  **(n\_fft, hop\_length)** | **Validation** | **Public分數** | **Private** |
| Ensemble  (4096,1024) | 0.509(為三者平均) | 0.6195 | 0.6215 |
| Ensemble  (4096,2048) | 0.528(為三者平均) | 0.5884 | 0.5907 |
| Ensemble  (2048,1024) | 0.53(為三者平均) | 0.5884 | 0.6187 |

觀察到不同的幀長及幀移會對預測結果產生影響，但測試量體小無觀察到其上升下降趨勢，可做為日後研究方向

* 另有嘗試先將嗓音訊號之資料特徵(mfcc)進行模型預訓練，進行auto encoder(如下圖所示)，並擷取其中30個神經元代表訊號特徵，與病史紀錄特徵合併作為輸入進行XGBClassifier模型訓練， 測試結果Recall值為0.41，無明顯進步故放棄此做法。



最後，我們也嘗試盡可能提取嗓音訊號特徵，進行預訓練取出最後一層作為音訊特徵，並合併病史紀錄特徵進行訓練，其Recall值可達0.8，但放上public分數只獲得0.318，可見其受Overfitting的影響，但是否有機會藉dropout或調整其他特徵來降低此影響，可做為日後思考方向。

**柒、程式碼**

Google Drive：

<https://drive.google.com/drive/folders/1eOyFA5nv7C0YIwjNI3YAWIxC834QuJU2?usp=share_link>

**捌、使用的外部資源與參考文獻**

音訊處理套件:https://librosa.org/doc/main/tutorial.html

**報告作者聯絡資料表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 隊伍名稱 | TEAM\_3680 | Private Leaderboard 成績 | 0.621502 | Private Leaderboard 名次 | 3 |
| 身分  （隊長/隊員） | 姓名 (中英皆需填寫)  (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號) | 學校＋系所中文全稱  (請填寫完整全名，勿縮寫) | 學校＋系所英文中文全稱  (請填寫完整全名，勿縮寫) | 電話 | E-mail |
| 隊長 | 江前昱  Chien-Yu, Chiang | 友達光電股份有限公司 | AUO | 0926654388 | leo24237260@gmail.com |
| 隊員1 | 蘇義新  I-Hsin, Su | 友達光電股份有限公司 | AUO | 0922583106 | spsedward@gmail.com |
| 隊員2 |  |  |  |  |  |
| 隊員3 |  |  |  |  |  |
| 隊員4 |  |  |  |  |  |
| 指導教授資料 | | | | | |
| 每隊伍至多可填寫兩名 | 指導教授  中文姓名 | 指導教授  英文姓名  (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號) | 任職學校＋系所  中文全稱  (請填寫完整全名，勿縮寫) | 任職學校＋系所  英文全稱  (請填寫完整全名，勿縮寫) | E-mail |
| 教授 1 |  |  |  |  |  |
| 教授 2 |  |  |  |  |  |

★註1：請確認上述資料與AI CUP報名系統中填寫之內容相同。自2023年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信moe.ai.ncu@gmail.com。，報告繳交截止時間後將不予修改。

★註2：繳交程式碼檔案與報告，請Email至：aicenter@g.yzu.edu.tw，並同時副本至：

t\_brain@trendmicro.com與moe.ai.ncu@gmail.com。缺一不可。