Maestría en Explotación de Datos y Descubrimiento de Conocimiento
Aprendizaje Automático

1er cuatrimestre 2021

Trabajo Práctico nro 2

Integrantes:

López Malizia, Álvaro
Padula, Eliana Inés
Rossi, Fabiana Alejandra

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

July 14, 2021

Resumen

El presente documento pretende explorar el uso de algoritmos de ensambles. A tal fin, se trabajó con una muestra de archivos de audio con el objeto de predecir qué emociones se expresaron en los mismos.

Se estudió el efecto de la separación de los datos mediante *k fold cross validation* o *leave one group out* en los modelos de ensamble *Random Forest* y *AdaBoost*. Luego se realizó una búsqueda de hiperparámetros para los modelos elegidos con *leave one group out*, y se comparó su desempeño. Se evaluó cuáles eran las emociones más frecuentemente confundidas por el modelo de *Random Forest*, utilizando una matriz de confusión que concatenó el resultado de las predicciones de cada *fold*. Adicionalmente, se analizó el efecto de la normalización de los datos por actor.

Nuestros resultados indicaron que los modelos estudiados eran robustos ante los distintos modos de separación de los datos. *Random Forest* resultó mejor para modelar las emociones a partir de audios, con un mejor *accuracy* para los datos cantados respecto aquellos hablados, al utilizar los datos originales. La normalización de los datos de entrada mejoró la *performance* del modelo para ambos tipos de categorías del discurso. Finalmente, aunque con menor desempeño, se construyó un modelo de red neuronal multicapa para la clasificación de las emociones.

El código de cálculo del algoritmo y gráficos se encuentran en: https://colab.research.google.com/drive/12BMplIe2OmnkA2ULaTnnlhC2zPwToPta?usp=sharing.

Introducción

La inferencia del estado emocional del hablante a partir del audio de su discurso ha despertado un creciente interés en los últimos años. Particularmente, en el área de inteligencia artificial y robótica, dicho estado emocional cumpliría un rol esencial en posibles interacciones humano-robot [1]. En este sentido, el objetivo de este trabajo práctico fue analizar el uso de algoritmos de ensambles aplicados en el conjunto de datos *Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS)* [2], con el fin de predecir qué emociones se expresaron a partir de archivos de audio. Se extrajeron de los mismos un conjunto de atributos determinados a partir de *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)* [3], y se construyó una base de datos única.

En primer lugar se realizó una partición del *dataset* utilizando *12 k fold cross validation* o dejando fuera dos actores en cada *split*, y se analizó el efecto de dicha partición mediante los modelos de ensamble *Random Forest* y *AdaBoost*. En ambos casos se observó que el modelo de *Random Forest* presentaba mejores métricas de *performance* que *AdaBoost*, y que la estrategia de separación no parecía afectar la *performance* de los modelos. Por tal motivo, y en segundo lugar, se realizó una búsqueda de hiperparámetros para el modelo de *Random Forest* y de *AdaBoost* utilizando la separación de datos que deja un par fuera, sobre el conjunto de datos. Se evaluó cada modelo en validación, y se observaron mayores valores de *performance* para el modelo de *Random Forest*. A continuación se utilizó el mejor modelo de *Random Forest* para evaluar los datos de discurso hablado y cantado por separado. Se generó una matriz de confusión concatenando los resultados de la predicción de cada *fold*. La misma permitió estudiar cuáles eran las emociones de peor discriminación, es decir, las emociones más frecuentemente confundidas por el modelo. Se repitió el análisis para el modelo de *Random Forest*, pero utilizando datos previamente normalizados por actor. Se observó un aumento en la métrica de *performance* estudiada para los dos tipos de audios, cantado y hablado. Por último, se realizó una red neuronal multicapa que no superó las *performances* obtenidas por los modelos de árboles de decisión.

Datos

La base de datos multimodal *RAVDESS* contiene archivos de audio y video de actores profesionales (12 mujeres y 12 hombres) que vocalizan dos frases en el idioma inglés, con acento neutral de Norte América y en dos modalidades distintas: habla o canto. Las frases habladas incluyen emociones como calma, felicidad, tristeza, enojo, miedo, sorpresa y desagrado. Las frases cantadas incluyen calma, felicidad, tristeza, enojo y miedo. Cada expresión se generó con dos niveles de intensidad emocional (normal o fuerte), y con una expresión neutral extra. En este trabajo se utilizaron los datos correspondientes a los archivos de audio, que incluyen expresiones habladas y cantadas para todos los actores, salvo para el número 18, del cuál sólo se dispone de registros hablados.

Metodología

Datos

Modelo de datos: Se extrajeron los atributos de habla determinados en *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)* [3] de los archivos de audio utilizados.

Preprocesamiento: Se construyó una base de datos única con estadísticas de resumen para cada atributo y para cada actor, derivados de las secuencias temporales o atributos de "bajo nivel" de los audios. Se compararon los resultados de nuestro análisis con datos que fueron previamente normalizados por actor. Dicha normalización se realizó utilizando la función *zscore* del módulo *stats* de la librería *scipy*.

Separación de datos: Se realizó una primera separación del *dataset* mediante la técnica de *k fold cross validation* o dejando un par de actores fuera en cada partición (cada pareja incluyó un actor hombre y una mujer). No se observaron diferencias significativas entre ambos métodos. Se decidió entonces utilizar la técnica que deja un par afuera para optimizar los hiperparámetros de los modelos a testear, ya que asegura un menor sesgo al momento de emplear el modelo en personas que no se encuentran incluidas en el conjunto de entrenamiento.

El entrenamiento de los modelos se realizó sobre la totalidad de los datos, utilizando las categorías habladas y cantadas. Dado que las clases de la variable objetivo estaban desbalanceadas en las distintas categorías de discurso (hablado o cantado), se evaluó cada categoría por separado. Las emociones disgustado y sorprendido no están representadas en los audios cantados. En la **Figura 1** se indican las frecuencias relativas de cada clase de la variable objetivo en las categorías de discurso hablado o cantado.

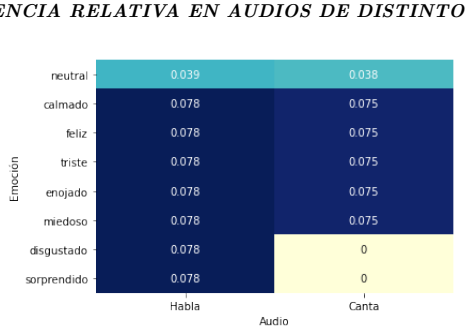

Figura 1: *Frecuencias relativas de cada clase de la variable objetivo del discurso hablado o cantado.*

Valores faltantes: Se poseen únicamente registros de audio hablados para el actor 18. Se decidió no eliminar los mismos, dado que implicaría modificar la estrategia de comparación de *12 k folds versus* dejando un par de actores afuera en cada *split*. Para esto, sería necesario remover los registros del actor 18 y algún otro actor, para obtener un número par de individuos grabados. Las implicancias de esta decisión van a verse reflejadas en la estrategia de división de datos que dejan un par de actores afuera por vez. Los datos de entrenamiento serán más completos para aquellas particiones en las que el actor 18 quede en el conjunto de validación.

Algoritmos

Selección de los hiperparámetros definidos:

Random Forest: Se realizó una búsqueda de hiperparámetros mediante la metodología de *Random Search*. Se varió la cantidad de árboles generados (entre 50 y 100, tomando de a 5 unidades) y el impacto de un remuestreo *boostrap*. Los hiperparámetros seleccionados fueron: Cantidad de árboles: 75. Remuestreo *boostrap*: No.

AdaBoost: Mediante *Random Search* se evaluó el uso de la metodología de SAMME o SAMME.R para la convergencia al error, evaluando entre entre 50 y 100 modelos (tomando de a 5 unidades). La selección de hiperparámetros encontró que aquellos con mejor *performance* eran los siguientes: Cantidad de árboles: 95. Metodología de convergencia al error: SAMME.

Matriz de confusión:

Se construyeron modelos para la predicción de las emociones a partir de los *dataset* de audios cantados o hablados. Bajo los hiperparámetros previamente definidos, se creó una matriz de confusión para los datos predichos por *Random Forest*, utilizando las predicciones de cada *fold*, y se determinaron las métricas de *precision*, *recall*, *f1-score* y *support*. Se utilizó la métrica *accuracy* para comparar la *performance* promedio del mejor modelo ensayado sobre los distintos grupos de datos.

Red Neuronal Multicapa:

Se construyó una red neuronal con una capa oculta de 128 neuronas, una función de activación del tipo "Relu" y una capa de salida de 8 neuronas (emociones a predecir). Se utilizó la optimización del tipo "adam".

Resultados

Evaluación del sesgo de los datos:

La separación de los datos utilizando la metodología de *12 k fold*s o dejando dos actores afuera no modificó significativamente la *performance* de los modelos (**Figura 2**). No obstante, se observó un aumento en la variabilidad registrada para el modelo de *Random Forest* cuando los datos se evaluaban con la metodología que deja un par afuera.

Este aumento en la variabilidad se debe a que la metodología de *12 k folds* separa los datos independientemente del actor que los originó, y por ende el modelo podría aprender las particularidades propias de cada individuo, lo cuál se ve reflejado en el rendimiento en la validación. Por el contrario, en la división que deja un par afuera por vez, el modelo tiene mayor dificultad para generalizar e independizarse de las variabilidades que tiene cada actor, y el rendimiento depende del par particular que se valide en cada *fold*. Asimismo, esta última estrategia se ve particularmente afectada por la ausencia de datos para el actor 18, ya que los datos de entrenamiento de la partición en la cual el actor 18 quede en el conjunto de validación, serán más completos respecto las demás particiones. No obstante, y aún cuando la mediana de las métricas de *performance* estudiadas son semejantes para la separación con *12 k folds* o dejando un par de actores fuera, resulta importante destacar que cuando el modelo se encuentre en producción, el mismo seguramente será utilizado en personas no incluidas en el conjunto de datos de entrenamiento. Por ende, la estrategia de dejar un par por vez es más adecuada.

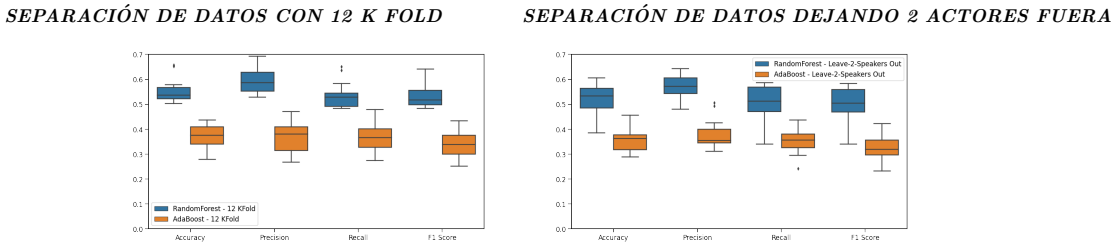

Figura 2: *Evaluación de métricas de performance utilizando 12 k folds (izq) o dejando un par afuera (der).*

Predicción de emociones

A continuación, se realizó un análisis de predicción de emociones utilizando el método de *Random Forest* y *AdaBoost*, con hiperparámetros definidos mediante la técnica que deja un par de actores por vez, y utilizando datos de audio cantados y hablados por separado. Dado que la *performance* de *Random Forest* superó aquella del modelo de *AdaBoost*, se construyó una matriz de confusión para las distintas emociones predichas de este modelo. En dicha matriz se incluyen las emociones predichas en cada uno de los *folds*. Asimismo, se comparó dicho análisis con datos que habían sido previamente normalizados por actor (**Figura 3**).

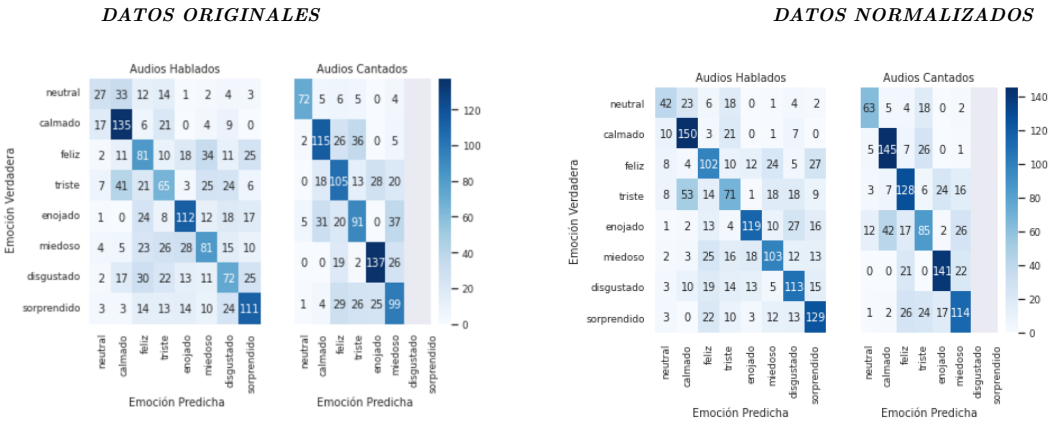

Figura 3: *Matriz de confusión correspondiente a los audios cantados y hablados (izquierda), y matriz de confusión con datos normalizados (derecha). Se indican los resultados concatenados de las predicciones en cada una de las 12 particiones del dataset, dejando fuera para validación a dos actores por vez.*

Nuestros resultados indicaron que la normalización de los datos mejoró la métrica de *accuracy* tanto para los audios hablados como cantados (**Tabla 1**).

Accuracy	originales	normalizados
audios hablados	0.475 +/- 0.064	0.576 +/- 0.094
audios cantados	0.614 +/- 0.104	0.67 +/- 0.064

Tabla 1: *Evaluación de modelo de Random Forest en el conjunto de datos de test utilizando accuracy. Los datos fueron evaluados de acuerdo a la modalidad del discurso, y se analizó el efecto de la normalización por actor. En la tabla se indican los valores medios del valor de accuracy para cada uno de los 12 splits y su desvío estándar.*

Un análisis más detallado de las matrices de confusión permitió arribar a las siguientes conclusiones: la emoción neutral es mayoritariamente predicha como tal en el discurso cantado, pero el modelo también la predice como calmado, triste y feliz en los audios hablados. La normalización de los datos mejora las predicciones de dicha emoción en audios hablados. La emoción calmado está muy bien identificada por los modelos. El modelo de *Random Forest* confunde la emoción feliz y triste con las emociones miedoso/sorprendido y calmado/miedoso/disgustado, respectivamente, y muestra mejores valores predictivos para los audios cantados. La normalización de los datos mejora la predicción para estas emociones. En el caso de los audios etiquetados con emoción enojado, el modelo confunde esta emoción con feliz, disgustado y sorprendido en los audios hablados, y muestra mejor precisión para los audios cantados. La normalización de los datos mejora las predicciones. La emoción predicha miedoso no identifica correctamente los verdaderos positivos, independientemente de la modalidad de discurso o la normalización de los datos. Para dicha etiqueta, el modelo también predice la emoción enojado, triste, feliz y calmado. La emoción disgustado en los datos hablados es muy mal predicha por el modelo, ya que la clasificación predice además a las emociones sorprendido, feliz y triste, entre aquellas de mayoritaria proporción. La normalización de los datos mejora las predicciones sobre esta emoción. Finalmente, la emoción sorprendido se encuentra, en gran medida, bien predicha por el modelo. La normalización de los datos mejoró levemente su predicción. Otro factor que impactó positivamente en las métricas, fue la menor cantidad de emociones presentes en los audios cantados. Para todas las emociones el nivel de predicción fue más acertada.

Por último, se realizó un análisis con redes neuronales, utilizando una capa oculta de 128 neuronas. Dicho modelo mostró un valor de *accuracy* de de 0.287 para los datos de entrenamiento, mientras que un 0.297 de precisión sobre el conjunto de testeo, utilizando tanto los datos hablados como cantados. La arquitectura de la red se muestra en el **apéndice**.

Conclusión

El presente trabajo práctico nos permitió familiarizarnos con distintas modelos de ensamble. La utilización de un *dataset* cuyos atributos correspondían a características de audios resultó novedosa. Utilizando estos atributos complejos, se construyeron modelos que permitieron predecir las emociones de discursos hablados o cantados.

El estudio de distintos modos de separación de datos (*12 k fold cross validation* o dejando un par de actores fuera) indicó que la mediana de las distintas métricas utilizadas en los modelos de *Random Forest* o *AdaBoost* no cambió significativamente. No obstante, el modelo de *Random Forest* mostró mayor variabilidad en sus métricas cuando los datos fueron particionados utilizando la estrategia de dejar dos actores por fuera del conjunto de entrenamiento. Esto se debe a que la eliminación de datos por par de actores disminuye la homogeneidad de los datos del entrenamiento, y se pierde información en cada *split*. Dicha pérdida, en última instancia, impacta generando modelos de precisión más variables frente a los atributos de validación. Si bien esto podría interpretarse como una cualidad indeseable para un modelo en producción, es relevante resaltar que el mismo posiblemente esté destinado a ser utilizado con datos nuevos con los cuales el modelo no haya interaccionado. En este sentido, resulta más apropiado generar un modelo con una separación de datos que discretice los mismos por individuo, de modo tal que se adecue más fielmente a las posibles aplicaciones futuras.

Luego de la selección de hiperparámetros utilizando la técnica que deja un par fuera, se seleccionaron los mejores modelos para *Random Forest* y *AdaBoost*, y se determinó su *accuracy* en el conjunto de validación. El desempeño del modelo de *Random Forest* fue mejor, tal como se esperaba que sucediera dado que los arboles de decisión son modelos de bajo sesgo y alta varianza. Por otra parte, se evidenció un valor de *accuracy* mayor del modelo para los audios cantados en comparación con los hablados. Esto podemos atribuirlo a la menor cantidad de emociones registradas en los audios cantados, que podrían ser un factor que optimice la generación de los modelos. Una hipótesis alternativa es un incremento en la expresividad generada durante la entonación en una canción (de forma consciente o inconsciente). La calidad de los datos asociados en esta condición podría permitir la construcción de árboles más eficientes. De manera complementaria, se estudió el efecto de la normalización de los datos por actor en un modelo de *Random Forest*. El desempeño del modelo en estas condiciones fue superior a la condición en la que se utilizaron los datos originales. Esta normalización permite eliminar la variación debida a los actores. No obstante, la normalización de datos empeoró la precisión y *recall* para algunas de las emociones predichas.

En términos generales, el modelo de *Random Forest* predijo las emociones neutro, calmado, enojado y sorprendido correctamente. Las emociones feliz, triste fueron confundidas con otras emociones, y la peor predicción se observó para miedoso y disgustado.

Finalmente, el modelo de red neuronal demostró una baja capacidad de predicción. Probablemente dicho modelo podría optimizarse con el agregado de capas ocultas de neuronas.

Bibliografía

[1] Noroozi, Fatemeh; Sapinski, Tomasz; Kaminska, Dorota; Anbarjafari, Gholamreza (2017). Vocal-based emotion recognition using *Random Forest* and decision tree. International Journal of Speech Technology, 20 (2), 239246. DOI: 10.1007/s10772-017-9396-2.

[2] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[3] F. Eyben, et al.,"*The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)* for Voice Research and Affective Computing" in IEEE Transactions on Affective Computing, vol. 7, no. 02, pp. 190-202, 2016. doi: 10.1109/TAFFC.2015.2457417

Apéndice

En esta sección se incorporan la estructura de la red neuronal generada. En la **Figura Suplementaria 1** se muestran las capas, el tamaño del vector de salida y la cantidad de parámetros en cada una.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_26 (Dense)             (None, 128)               11392
_________________________________________________________________
dense_27 (Dense)             (None, 8)                 1032
=================================================================
Total params: 12,424
Trainable params: 12,424
Non-trainable params: 0
_________________________________________________________________
```

Figura Suplementaria 1: *Detalles del modelo de redes neuronales utilizado.*