EXPERIMENTO 10. ANÁLISIS DE RED NEURONAL BASADA EN INFORMACIÓN CONTEXTUAL

CONTROL DE VERSIONES

Versión	Descripción	Fecha de última modificación	Validado
1.0	Primera versión estable	11/04/21	No

INTRODUCCIÓN

En este experimento se pretende crear una red neuronal basada en los datos contextuales adheridos al conjunto de datos objeto de este proyecto.

El análisis de imágenes es una buena herramienta, pero trabajar con datos contextuales puede aumentar la precisión del diagnóstico.

En este caso es importante evaluar el rendimiento del modelo neuronal, esto es decisivo de cara a decidir si proseguir o no con el problema. Como ya se ha constatado en experimentos anteriores uno de los factores determinantes para que un ensamblaje de modelos funcione es que ambos modelos tienen una precisión similar.

Basado en lo anterior en este estudio se plantea la siguiente hipótesis:

- *Una red neuronal sencilla basada en información contextual puede conseguir un desempeño óptimo y posteriormente usarse en un ensamblaje de modelos*

ANÁLISIS DE DATOS

En esta sección realizaremos un proceso de análisis y depurado de datos.

Las características contextuales disponibles son:

- ***Edad***
- ***Sexo***
- ***Método de evaluación***
- ***Lugar de la lesión***

En primer lugar, localizamos que entradas tienen entradas vacías, en este caso:

- ***Sexo***
- ***Edad***

Para el sexo hemos optado por fijarlo como mujer, esta decisión se ha tomado en base a dos razones:

1. El número de entradas vacías apenas eran 60 de las 10000 totales.
2. El número de hombres y mujeres en el conjunto de datos es ligeramente superior en el caso de las hombres. Con esta decisión se obtiene un conjunto de datos más estable.

En lo referente a la edad se ha decidido sustituir las entradas vacías por la media de edad del conjunto de datos. De nuevo la realidad es que hay muy pocos casos, por lo que realmente el efecto es escaso.

El siguiente paso es mapear la información para que pueda ser suministrada a la red. Se realizará de la siguiente manera:

- ***Edad***. {edad <= 16: 0, edad > 16 & edad <= 32: 1, edad > 32 & edad <= 48: 2, edad > 48 & edad <= 64: 3, edad > 64: 4}
- ***Sexo***. {"male": 1, "female":0}
- ***Método de evaluación***. {"histo": 1, "follow_up": 2, "consensus": 3, "confocal": 4}
- ***Lugar de la lesión***. {"acral": 0,"back": 1,"lower extremity": 2,"trunk": 3,"upper extremity":4,"abdomen":5,"face":6,"chest":7,"foot":8,"unknown":9,"neck":10,"scalp":11,"hand": 12,"ear":13,"genital":14}

DISEÑO DE LA RED Y PRUEBAS

Se propone una red muy sencilla. Con la siguiente arquitectura:

Este tamaño ha sido el óptimo encontrado entre algunas aproximaciones previas. Es importante recordar que solo contamos con cuatro variables de entrada a la red.

Los resultados obtenidos con esta red a lo largo de diez iteraciones han sido:

loss: 1.3881 - acc: 0.4257
loss: 1.3881 - acc: 0.4257
loss: 1.3882 - acc: 0.4257
loss: 1.3882 - acc: 0.4257
loss: 1.3883 - acc: 0.4257
loss: 1.3883 - acc: 0.4257
loss: 1.3883 - acc: 0.4257
loss: 1.3882 - acc: 0.4257
loss: 1.3882 - acc: 0.4257
loss: 1.3881 - acc: 0.4257

Como se puede ver una precisión media muy lejos de lo aceptable.

No obstante, antes de desistir se plantea aplicar la estrategia de selección de características.

Los resultados fueron los siguientes:

- Usando función d_classif de Sklearn: [513.69, 284.31, 16.40, 71.66]
- Usando ExtraTreesClassifier de Sklearn: [0.407, 0.155, 0.040, 0.398]
- Usando RFE de Sklearn (clasificador basado en regresiones logísticas): [1, 2, 4, 3]

Los tres métodos coinciden en que el sexo es la variable menos importante. Por ello mismo procedemos a eliminarla y probar de nuevo el modelo. Esta vez el modelo ha dado los siguientes resultados:

loss: 1.3568 - acc: 0.4536
loss: 1.3568 - acc: 0.4536
loss: 1.3567 - acc: 0.4536
loss: 1.3566 - acc: 0.4536
loss: 1.3564 - acc: 0.4536
loss: 1.3562 - acc: 0.4536
loss: 1.3562 - acc: 0.4536
loss: 1.3561 - acc: 0.4536
loss: 1.3560 - acc: 0.4536
loss: 1.3560 - acc: 0.4529

CONCLUSIONES

Este experimento nos deja dos conclusiones:

1. Los datos contextuales no permiten crear una red lo suficientemente buena como para dar resultados óptimos.
2. Los conjuntos de datos de carácter médico suelen tener una información contextual insuficiente debido a la necesidad de anonimizar estos datos. Como futura ruta de investigación se proponen la creación de datos sintácticos mediante GANs.