

Tarea 5 – Raimundo Herrera

En esta tarea el objetivo es experimentar con distintos clasificadores para un mismo problema, pero tras realizar todo el proceso de extracción y selección. La intención es comparar los rendimientos de los clasificadores no únicamente respecto al *accuracy* obtenido por ellos, sino que también respecto a medidas relevantes, como lo son el *precision* y el *recall*. El primero indica la cantidad de verdaderos positivos sobre la cantidad de verdaderos más falsos positivos, y el segundo indica la cantidad de verdaderos positivos sobre la cantidad de verdaderos positivos más falsos negativos. Ambas dan intuiciones más allá de meramente la precisión.

Para todos los experimentos se realizó el mismo preprocesamiento, extracción y selección de características. En primer se pasaron las imágenes a escala de grises para evitar falsas correlaciones de color, con la hipótesis de que la información viene de la forma y posición y no del color. En segundo lugar, para la extracción se utilizaron las características de *local binary patterns* (LBP) e *histogram of gradients* (HoG). Para ambos se utilizaron divisiones y parámetros similares. Por ejemplo se utilizó el *mapping* uniforme y variante a rotación para LBP de modo que si toma en cuenta la orientación y maneja correctamente variaciones de brillo en la escala de grises. Para HoG se utilizaron 8 *bins* para las 8 direcciones. En tercer lugar, se hizo limpieza de características, para evitar aquellas que no otorgan información, y normalización de las mismas. Finalmente se realizó una selección de acuerdo al método *sequential feature selection* (SFS), de 50 características, para reducir el espacio original de 603 encontrando las que otorgan mayor separación.

Una vez realizado lo anterior se exploraron 3 opciones para la clasificación. La primera consta de un simple clasificador de vecinos cercanos (KNN) con 2 vecinos, el cual reporta, tras evaluar los puntajes con validación cruzada, los siguientes valores promedio: un *accuracy* de 0.95 ± 0.02 , un *precision* de 0.97 ± 0.04 y un *recall* de 0.94 ± 0.03 . Dichos valores resultan ser bastante elevados para la simpleza del clasificador.

La segunda opción utiliza *Support Vector Machines* (SVM) y se realiza para este clasificador una búsqueda exhaustiva de los parámetros C y Gamma, la cual se realiza con validación cruzada de 10 *folds*. Para este caso la validación cruzada no se realiza únicamente para evaluar sino que también para decidir los parámetros del modelo, que, considerando que no hay *set* de pruebas es una buena opción para elegir los mejores parámetros sin *overfitting*. Con SVM los resultados promedio que se consiguen tras obtener los mejores parámetros son: *accuracy* de 0.9925 ± 0.0114 , *precision* de 0.9905 ± 0.019 y *recall* de 0.995 ± 0.015 . Dichos valores son muy cercanos al 100% y además presentan desviaciones muy pequeñas entre los conjuntos elegidos durante la validación cruzada, con lo que se ve que es una solución robusta.

La tercera opción utiliza redes neuronales, y se eligen parámetros con un procedimiento similar al anterior, es decir, búsqueda exhaustiva con validación cruzada, de aquellos que se consideran más importantes, como lo son la tasa de aprendizaje, el tamaño de las capas ocultas y la función de activación. Con esa búsqueda se llega a parámetros que otorgan los mejores desempeños con los cuales se obtienen los siguientes resultados: *accuracy* de 0.9925 ± 0.0114 , *precision* de 0.995 ± 0.014 y *recall* de 0.989 ± 0.02 . Como se observa, valores ligeramente superiores a los anteriores en ciertos casos pero con una fuerte dependencia del *random state* inicial, es decir de cómo se inicializan los valores.

Es importante recalcar que en este experimento no hubo *set* de *testing* y que los resultados se obtienen puramente de la validación cruzada en entrenamiento, por lo que los resultados no dicen que el clasificador sea casi perfecto, ya que no se ha enfrentado a nueva información que no haya visto en entrenamiento.

Para apreciar la comparativa de mejor manera, en los siguientes tres gráficos se aprecian los puntajes con sus respectivos márgenes de error para cada modelo.

Se puede observar que los mejores resultados están en las redes neuronales y SVM.

Lo que da luces de la simplicidad del problema y que un *approach* relativamente simple da resultados satisfactorios, sin necesidad de sobre-complejizarlo. Además muestra que un procedimiento casi estándar para extracción y selección es lo suficientemente expresivo para luego con los clasificadores adecuados obtener resultados satisfactorios.

