Reconocimiento de patrones: Práctica 2

Parte 1: Generación del modelo

Preprocesado de los datos

En primer lugar, cargamos los datos y transformamos las etiquetas de caracteres a números. Posteriormente normalizamos los datos a la unidad. Una vez que tenemos las categorías en número y los datos normalizados, generamos los conjuntos de entrenamiento y validación.

Selección de características

Tras tener los datos preparados, aplicamos el algoritmo de Random Forest sobre los atributos de los datos de entrenamiento para obtener cuales de estos atributos son más importantes para el reconocimiento de las categorías. Dado que hay 178 características, decidimos quedarnos con las 20 más significativas guardando sus valores en un archivo .txt llamado *top_atribute.txt*. De esta forma pasamos de trabajar de 178 categorías a 20 reduciendo así la dimensionalidad del problema.

Generación y selección del modelo

Una vez decididas las características a utilizar, realizamos el proceso de aprendizaje para varios clasificadores: Random Forest, AdaBoost y Gradient Boosting.

Para elegir el mejor hiperparámetro para los clasificadores hemos realizado un barrido y para cada uno de los hiperparámetro validación cruzada eligiendo el mejor clasificador.

En el caso de Random Forest, hemos variado el rango de los números de los estimadores de 1 hasta 30, y el mínimo de impuridad de 0 a 0.2, ambos con 30 valores intermedios. Para AdaBoost y Gradient Boosting el número de estimadores se ha variado de 50 hasta 150 y su learning rate de 0.01 hasta 1, ambos con 20 valores intermedios. Se han reducido el número de valores intermedios para reducir el tiempo de cálculo.

Cabe decir que durante el barrido para cada pareja de hiperparámetro se ha utilizado la validación cruzada K-fold en el que se ha escogido el mejor clasificador entre los k clasificadores. En nuestro caso hemos utilizado una k de 5.

Una vez entrenado los modelos barriendo los hiperparámetros tal y como hemos descrito hemos obtenido los siguientes valores:

Random forest (n_estimators = 20.0, min_impurity=0.013793103448275864)
Score = 0.9705882352941176

AdaBoost (learning_rate=0.4268421052631579, n_estimators=50)
Score = 0.8823529411764706

Gradient Boosting (learning_rate=0.4789473684210527, n_estimators=66)
Score = 0.9411764705882353

Finalmente, escogemos el clasificador con mayor precisión, Random Forest y lo guardamos para llevar a cabo la segunda parte de la práctica.

Parte 2: Test del modelo

En primer lugar, cargamos el modelo que hemos generado. Posteriormente, le aplicamos el mismo preprocesado, tanto de estandarizar a la unidad como de reasignación de las etiquetas.

A continuación, seleccionamos las columnas que fueron detectadas como las más relevantes en el proceso de entrenamiento desde un archivo .txt con el nombre de *top_atribute.txt*.

Con todos los datos adecuados a nuestro modelo, realizamos la predicción de las etiquetas para los datos de test. Con las etiquetas reales y con las estimadas llevamos a cabo el cálculo de una matriz de confusión para ilustrar nuestra solución, así como el cálculo de la precisión del método.