Tarea 3 – Raimundo Herrera

En esta tarea el objetivo es experimentar con diferentes *sets* de datos utilizando técnicas de selección y transformación de características, además de técnicas de clasificación. La discriminación a realizar es categórica entre las distintas clases de tortillas, caras y géneros, efectuando experimentos para cada una.

Dado que la extracción de características viene dada, la experimentación consiste en decidir como utilizar las técnicas de SFS para selección, PCA para transformación, LDA, QDA o KNN para clasificación, o cualquier combinación de ellas, acorde al problema y el rendimiento en los mismos.

Para todos los casos, al igual que en la tarea anterior, se realiza una limpieza y normalización de los *datasets* con los algoritmos disponibles en la librería *pybalu.* Adicionalmente en los 3 experimentos se realiza una división arbitraria entre conjuntos de entrenamiento y de pruebas correspondiente a una división 80%/20% para cada clase, respectivamente.

Para todos los experimentos se realizaron diversas pruebas para cada técnica utilizada, esto es, diferentes valores para la cantidad de características seleccionadas, diferentes valores para los parámetros de KNN[1], LDA o de QDA, etc. Todo lo anterior cuidando la restricción del enunciado de no superar las 10 características. En el caso de las tortillas no es necesaria ninguna transformación de características ni tampoco una clasificación distinta a la de vecinos cercanos. Lo que se realizó fue aplicar SFS de diversas cantidades de características obteniendo que con 10 y con ajustar vecinos cercanos con k = 1, el resultado era de 100% en entranamiento y testing. Se tuvo especial cuidado de excluir las características de posición, tras el consejo del profesor.

En el caso de las caras, se realizaron diversos experimentos para determinar una cantidad de características adecuada, que resultó ser 9. Únicamente con esa selección los resultados obtenidos eran bastante cercanos a la perfección, pero se observó que al aplicar la transformación de PCA posterior a la selección, reduciendo por ejemplo a 6 componentes el resultado no disminuía su precisión significativemente, pero sí la cantidad de características. Sin embargo, un análisis posterior mostró que utilizando tanto LDA como QDA con método de mínimos cuadrados no se volvía a mejorar el rendimiento en entrenamiento. Eso hizo que la elección final fuera KNN tras SFS sin otro método, ya que en training no mejoraba el rendimiento, sin embargo se ve que en testing era mejor QDA, pero no es adecuado usar esa información para elegir. Se tuvo un rendimiento final en testing del 94.3%.

Por último, el conjunto de datos más complejo resultó ser el de género. Se realizaron diversos experimentos similares a los anteriores, pero esta vez se volvió a utilizar una técnica que resultó ser muy útil en la tarea anterior: usar subconjuntos de las características escogidas por el SFS. Con eso, al realizar un SFS de 12 características y elegir una combinación de 7 de ellas se llegó a un rendimiento de un 99.8% en entrenamiento utilizando KNN de 2 vecinos, superior al de solo usar SFS. Se hizo así y también en las caras porque si se usaba k=1 el rendimiento en entrenamiento siempre era 100%. Así, es sorprendente que diversas aplicaciones de LDA, PCA o QDA, no entregaron mejores resultados, indicando que para este problema particular, los vecinos cercanos tras SFS otorgaban suficiente información. Se tuvo un rendimiento final en testing del 86.1%

El trabajo realizado es sumamente ilustrativo para mostrar que el dominio del problema es fundamental para decidir qué combinación de algoritmos de selección, transformación y/o clasificación utilizar. Se observa que no siempre mejoran los resultados al transformar, pero algunas veces lo más complejo ayuda. Se ve también que es importante no utilizar el set de *testing* para decidir qué modelo usar porque puede sesgar mucho y no es información que se tenga a priori, por lo que hay que utilizar solo *training*. No obstante, es muy importante comentar que los métodos utilizados tienen un grado de robustez alto debido a que se ajustaron a la restricción de no utilizar más de 10 características. Al hacer eso, se evita sobreajustar y es más probable estar eligiendo el conjunto de características o transformaciones de ellas que mejor sirve para discriminar entre las clases. Lo anterior da luces de que muchas veces es importante tener lineamientos generales, pero que también es clave considerar los aspectos propios de cada problema a la hora de decidir el camino a realizar para la clasificación, esto es, el conjunto de técnicas a utilizar.

[1] Se experimentó con distintos tipos de distancias, como euclideana, de minkowski, de chebyshev, etc.