Evaluación de técnicas de modelos de aprendizaje supervisado para la clasificación del mejor tipo de cultivo según características del suelo

Valentina Forero Hurtado
Alfredo Cuellar Valencia
Pontificia Universidad Javeriana
valentina_foreroh@javeriana.edu.co
alcuellar@javeriana.edu.co

---

1. Introducción

El presente proyecto tiene por objetivo entrenar un modelo que logre clasificar un cultivo a partir de características numéricas del suelo. El conjunto de datos que se utilizó proviene de una recolección realizada por Agrosavia. Esta consiste en una serie de características categóricas y numéricas de varios cultivos que se siembran alrededor de Colombia. Para nuestro caso, solo se extrajeron las características numéricas, tomando como etiquetas los nombres de los cultivos.

2. Desarrollo

Para el desarrollo del proyecto, se inicia con la limpieza de los datos, eliminando todas las muestras que tengan valores NaN en la columna de características, para luego seleccionar solo los datos numéricos, en donde los datos restantes que tienen pocos valores NaN, se les imputa el valor utilizando KNNImputer; generando adicionalmente un proceso de reducción dimensional por PCA para encontrar el número mínimo de componentes. Posteriormente se generan dos algoritmos de clasificación, en donde en máquinas de soporte vectorial se varía la semilla, y en redes neuronales se implementa una búsqueda por grid search para la obtener el número de capas ocultas del método, adicionalmente se prueban tres tipos de funciones de activación, implementando en los dos algoritmos un proceso de cross validación con 5 folds, poniendo a prueba diferentes métodos para optimizar el modelo.

3. Resultados

A través de la implementación de SVM, se obtuvo como resultado un F1 score de entre 0.3235 a 0.3300 y un coeficiente de Matthews de 0.2888 a 0.2977, en donde al trabajar con diferentes semillas la variación en los resultados es mínima. Por otro lado, para implementar ANN se inició obteniendo el valor adecuado para el número de capas ocultas usando la búsqueda por rejilla, encontrando qué en un rango de 10 a 50, 45 capas ocultas es el valor adecuado; además se pusieron a prueba tres funciones de activación en donde con *relu* el F1 score es igual a 0.5467 y el coeficiente de Matthews es de 0.49, para *tanh* el F1 score: 0.5385 y el MCC es igual a 0.4726 y para *logistic* el F1 score es igual a 0.5363 y el coeficiente de Matthews: 0.4776.

4. Conclusiones

- El modelo que mejor se ajusta a la búsqueda de los datos corresponde a redes neuronales, trabajando con 47 capas ocultas, la función de activación *relu* y un máximo de iteraciones de 3000 priorizando de está forma que la función estocástica converja de manera exitosa, adicionalmente al entrenar el modelo con cross validación se evalúa el rendimiento del modelo de forma confiable y robusta, trabajando con 5 pliegues todo el dataset.

- El modelo planteado a partir de ANN cuenta con un 25 % más de correlación con las etiquetas en comparación con SVM, encontrando como la capacidad de aprendizaje no lineal, la flexibilidad, adaptabilidad y el tratamiento de características de alta dimensionalidad, le brindan al modelo de redes neuronales ventaja sobre los demás.
- Teniendo en cuenta que el valor máximo de F1 score obtenido es de 0.5467 y MCC es igual a 0.4900, se puede concluir que las características brindadas por el dataset no permiten capturar adecuadamente las diferencias entre las clases, dando paso a que existan características irrelevantes que están introduciendo ruido afectando así el rendimiento del modelo. Además, es posible que los datos sobre los cuales se están trabajando cuenten con un desbalance de clases, en donde el algoritmo tiene dificultades para aprender correctamente la clase minoritaria generando de manera consecutiva el resultado sobre las métricas de evaluación, esta hipótesis puede ser respaldada en como al entrenar y evaluar los modelos de clasificación, uno de los warning expresaba que algunas de las etiquetas son tenían las muestras suficientes para separar los datos en 5 pliegues.