Comparación de métodos de selección de variables para la clasificación de pacientes con depresión mediante Random Forest

Luciana C. Chiapella, Leandro Grendas, Andrea E. Errasti, Eugenio A. Carrera Silva, Federico M. Daray

Palabras clave: Random Forest, clasificación, selección de variables

Introducción

Random Forest es uno de los algoritmos de aprendizaje automático más utilizados para la clasificación de unidades en función de ciertas variables o atributos. En muchos casos, se cuenta con una gran cantidad de variables para la construcción del modelo. Algunas de ellas pueden ser irrelevantes para la clasificación o redundantes en relación a otras variables. Además, la inclusión de demasiados atributos puede derivar en la construcción de un modelo complejo de difícil interpretación y en la pérdida de precisión en la clasificación.

El objetivo de las técnicas de selección de variables en el contexto del aprendizaje automático es encontrar un subconjunto de variables que permitan construir un modelo de clasificación optimizado para estudiar un fenómeno. Existen numerosos algoritmos diseñados de acuerdo a diversos criterios que pueden integrarse a Random Forest, a fin de seleccionar las variables que permitan mejores resultados en la clasificación mediante dicho algoritmo.

El objetivo de este trabajo es comparar los resultados obtenidos al aplicar cuatro métodos de selección de variables como paso previo a la aplicación de Random Forest para clasificar a un conjunto de individuos como depresivos o controles sanos, en función de biomarcadores inmunológicos e inflamatorios.

Métodos de selección de variables

En R, existen numerosos métodos para la selección de variables en problemas de clasificación. Uno de ellos es el Algoritmo Genético (*Genetic Algorithm*, GA), que se basa en los principios de la genética y la evolución biológica. En el GA, una población de soluciones candidatas evoluciona hacia mejores soluciones mediante la mutación y alteración aleatoria de sus propiedades. En el contexto de la selección de variables, la población está compuesta por combinaciones de variables de un conjunto de datos, y el objetivo es encontrar el subconjunto que maximice la diferenciación entre clases y minimice las diferencias dentro de la misma clase.

Otro método es el algoritmo de recocido simulado (*Simulated Annealing*, SA), que se inspira en el proceso de recocido de metales. El SA comienza seleccionando aleatoriamente un subconjunto de variables, construyendo un modelo y evaluando su rendimiento. Luego, modifica aleatoriamente un pequeño porcentaje de variables del subconjunto inicial y recalcula el rendimiento. Si el rendimiento mejora, el nuevo subconjunto se conserva; si no, se calcula una probabilidad de aceptación y se la compara con un número aleatorio uniforme. Si este es mayor que la probabilidad de aceptación, se rechaza el nuevo subconjunto y se utiliza el anterior. Si no se ha encontrado una nueva solución óptima dentro de I iteraciones, la búsqueda se restablece a la última solución óptima conocida y se repite el procedimiento.

La Eliminación Recursiva de Características (*Recursive Feature Elimination*, RFE) es otro método común para la selección de variables. RFE entrena un modelo utilizando todas las variables disponibles, evalúa la importancia de cada una y elimina la menos importante en cada iteración. Este proceso se repite hasta que se alcanza un criterio de parada, como un número predefinido de variables seleccionadas.

Un cuarto algoritmo es Boruta, que inicia duplicando las variables del conjunto de datos y asignándoles valores aleatorios para crear variables *sombra*. Luego, entrena el modelo de Random Forest con las variables originales y las *sombra*, y luego compara la importancia de cada variable original con su correspondiente *sombra*. Si la variable original es más importante que la *sombra*, se considera importante; de lo contrario, se elimina. Este proceso se repite hasta que todas las variables retenidas sean importantes.

Métodos

Se consideró un conjunto de 121 individuos, de los cuales 79 (65%) padecían depresión y el resto eran controles sanos. Para cada uno, se contaba con mediciones de 39 variables correspondientes a marcadores inmunológicos e inflamatorios. Los datos se dividieron aleatoriamente en un conjunto de entrenamiento y otro de prueba, en una proporción de 70/30. Los cuatro métodos de selección de variables se aplicaron al conjunto de entrenamiento, utilizando las librerías *caret* y *Boruta*. Con las variables seleccionadas en cada caso, se evaluó la precisión, sensibilidad, especificidad, valor predictivo negativo, valor predictivo positivo y coeficiente Kappa del algoritmo Random Forest para la clasificación de las unidades en el conjunto de prueba. Además, se registraron las variables retenidas por cada algoritmo. Este procedimiento se repitió iterativamente 100 veces. Finalmente, se compararon las medidas de eficiencia calculadas y la proporción de veces que cada algoritmo retuvo a cada una de las variables del conjunto de datos.

Resultados

En la Figura Nro. 1 se resumen los datos de las medidas de eficiencia para los cuatro métodos a partir de las 100 repeticiones del proceso. Se observa que SA presenta los resultados más desfavorables para todas ellas, mientras que Boruta y GA muestran los mejores valores. RFE tiene una performance similar a Boruta y GA en la mayoría de las métricas, a excepción del valor predictivo negativo y la sensibilidad, con resultados levemente inferiores. En cuanto a la cantidad de variables retenidas por cada método, GA presenta mayor valor medio y variabilidad, siendo Boruta y RFE los que tienden a seleccionar menor cantidad de atributos, con una mediana de 8 y 6 variables, respectivamente. Las variables con mayor frecuencia de selección son coincidentes con ambos métodos (datos no mostrados). Observando los resultados de manera global, Boruta resulta el método con mejor rendimiento.

Figura Nro. 1: Medidas de eficiencia para la clasificación de pacientes con depresión en función de las variables seleccionadas con cada método, a partir de 100 iteraciones.