Bosques aleatorios basado en proyecciones

Natalia da Silva[1], Dianne Cook[2], and Eun-Kyung Lee [3]

[1]IESTA-FCEA, UDELAR
[2]Monash University
[3]Ewha Womans University

Bosque aleatorio es un método de agregación supervisado basado en combinar modelos individuales de tipo árbol. Dos fuentes de aleatoriedad son introducidas, agregación bootstrap y selección aleatoria de variables en la partición del nodo. En este trabajo se presenta un método supervisado de agregación para problemas de clasificación basado en proyecciones que llamamos "projection pursuit random forest"(PPF). PPF usa el algoritmo PPtree introducido por Lee et.al (2013), donde los árboles individuales son construidos en base a combinaciones lineales de variables seleccionadas aleatoriamente en cada nodo.

"Projection pursuit" (PP) es utilizado para seleccionar una proyección de variables que mejor separe las clases. Utilizar combinaciones lineales de variables para separar las clases toma en cuenta la correlación entre variables y permite una mejora en la performance productiva de PPF respecto al bosque tradicional cuando la separación entre grupos ocurre en combinaciones de variables. Algunos trabajos previos usando árboles oblicuos en la construcción del bosque aleatorio han mostrado resultados positivos en términos de la performance predictiva pero solamente para problemas de dos clases.

El método que se presenta en este trabajo puede ser utilizado en problemas de múltiples clases y está implementado en un paquete de R llamado **PPforest** que se encuentra disponible en: `https://github.com/natydasilva/PPforest`.

Este proceso de combinar múltiples árboles de clasificación produce numerosos diagnósticos que mediante gráficos interactivos pueden darnos información sobre la estructura de clases en altas dimensiones. En este trabajo varios aspectos serán explorados como la evaluación de la complejidad del modelo, la contribución de los modelos individuales, la importancia de las variables y la incertidumbre en la predicción asociada con las observaciones individuales. Las ideas serán aplicadas a bosques aleatorios basados en proyecciones (PPF) pero puede ser extendido a otros métodos de conjuntos.