BOSQUES ALEATORIOS

APRENDIZAJE AUTOMÁTICO SUPERVISADO

Guillermo de Jorge Botana

Dpto. Psicobiología y Metodología en Ciencias del Comportamiento

Facultad de Psicología. Universidad Complutense de Madrid.

NOTA:

El contenido de este texto corresponde a uno de los temas de una asignatura del Máster de Metodología de las Ciencias del Comportamiento y de la Salud. Está elaborado para tener un texto base de lo que es explicado en clase. Aunque el texto es seguido y coherente, puede ser susceptible de algunas mejoras y ampliaciones. No obstante, es lo suficientemente autocontenido para llevar a cabo un estudio independiente sobre él. Se recomienda también haber leído el texto referente a los Árboles de Decisión, publicado también en este formato:

https://docta.ucm.es/entities/publication/a5d7f8b9-b77f-43fc-befb-9c8b5895d4a2

He decidido publicar este texto fuera del ámbito de la asignatura por si puede resultar de utilidad para otros estudiantes o por si a otros docentes les puede facilitar la tarea.

Tabla de contenido

Idea general

Anteriormente hemos visto los árboles de decisión (en caso contrario, acudir a texto base sobre árboles de decisión). Se explicaba entonces que dicha técnica trataba de encontrar una secuencia de particiones en las variables independientes que fuese reduciendo el ruido en la variable dependiente. El ruido era la entropía. Por ello se iba calculando en cada punto a expandir la entropía de la Variable Dependiente antes y después de la partición de cada variable independiente. La partición que más entropía retirase era la que se materializaba. Recordemos que el algoritmo era "voraz", y que de las tentativas había que ejecutar sólo la mejor. No se permiten pasos atrás.

Pues bien, puede decirse que los Bosques Aleatorios son una expansión de lo ya visto con los Árboles de Decisión, pero esta vez tomando de la muestra total de datos diferentes submuestras de menor tamaño (muestras con reemplazo). Esto se hace en aras a conseguir una mejor generalización del modelo.

Con cada submuestra se generará un árbol de decisión y de cada árbol se obtendrá una predicción. Cada árbol esta sintonizado a una parte de la realidad total representada en los datos. Cada árbol ve solo algunos aspectos. Juntando todas estas predicciones, se determinará la predicción definitiva, generalmente la mayoritaria o por promedio.

Los Árboles de Decisión son herramientas potentes para predecir, pero son deterministas, por lo que son inflexibles en su diagnóstico. Un conjunto de datos lleva a una misma predicción. Además, por esta inflexibilidad, suelen presentar una alta varianza en los resultados si se cambian o aumentan levemente los casos de la fuente de datos.

Sin embargo, empleando la misma técnica sobre diferentes aristas de la muestra y tomando éstas de manera aleatoria, podemos conseguir mejores adaptaciones en la predicción. Haciendo diversos árboles sobre esas submuestras conseguimos mayor estabilidad y generalización del modelo.

Ese es justo el valor añadido de los Bosques Aleatorios: hacer muchos árboles a partir de submuestras de la fuente de datos original y poder así apoyar la decisión en muchas decisiones paralelas (muchos aprendices débiles hacen una decisión fuerte). De esa manera, es previsible que, si varía la fuente de datos original, los nuevos casos se repartirán entre las diferentes muestras aleatorias y la predicción no se verá tan afectada.

Al final, teniendo muchos árboles, la decisión será mancomunada. De cada árbol se desprenderá una predicción, y la predicción definitiva será la mayoritaria (el promedio o la de mayor frecuencia). Es fácil imaginarse vía meronimia el porqué del nombre de “Bosque Aleatorio”.

Aspectos técnicos

Obtención de submuestras

En la introducción se ha aludido a que la técnica de Bosques Aleatorios generaba múltiples árboles de decisión a partir de tomar de la muestra principal muestras de menor tamaño. Con cada una de estas submuestras se calculará un árbol, es decir, un modelo. Es importante decir que este muestreo se hará con reemplazo. Esto quiere decir que la submuestra extraída no se retirará de la muestra total, por lo que el siguiente muestreo podrá de nuevo obtener casos de la submuestra anterior.

El hecho de que la submuestra obtenida sea de menor tamaño es dependiente de dos factores:

- **El número de casos en la submuestra**: de la muestra total se seleccionará aleatoriamente un porcentaje de casos que pasarán a formar parte de la submuestra (generalmente 60%-80%).
- **El número de Variables Independientes incluidas:** del conjunto de Variables Independientes de la muestra total, se tomará sólo un porcentaje de ellas para realizar las particiones.

Al proceso repetido de tomar submuestras y construir un modelo, es decir, un árbol, se le suele llamar “Bootstrap”. El concepto de “Bootstrap” alude a la generación recurrente de modelos con submuestras obtenidas de la muestra total.

Generación de los árboles

El algoritmo de generación de árboles es recurrente. Toma una submuestra y genera secuencialmente las mejores particiones siguiendo el criterio de la Ganancia de Información en la Variable Dependiente. De esta forma se generarán cada uno de los árboles. Un hiperparámetro importante es el número de árboles a generar porque superado un número de árboles, los resultados en la predicción serán asintóticos y el tiempo de proceso superfluo. Se pueden encontrar estimadores para identificar un número razonable.

Además, al igual que se dijo para los Árboles de Decisión, existen parámetros para controlar la profundidad de las ramas y el riesgo que se asume en cuanto a incertidumbre. En conjuntos de datos ingentes puede ser una buena opción limitar la profundidad si las ramas nos minimizan suficientemente la entropía. Las librerías traen estos valores para configurar, aunque también suelen venir por defecto.

Otra de las opciones con las que podemos contar al generar cada uno de los árboles es la posibilidad de reservar parte de la submuestra para medir el rendimiento de forma individual de cada uno. Este conjunto reservado no participa en la generación de las particiones. Solo se empleará para poner a prueba su predicción individual y contrastarla con la real. Esto no es estrictamente necesario, pero si puede ayudar a calibrar cada árbol por separado e incluso rastrear posibles parámetros de la calidad de las submuestras. En los paquetes de software se suele aludir a este conjunto reservado de cada submuestra como OOB (Out Of Bag).

Secuencia de predicción

Generados ya los árboles estamos en disposición de dar una predicción total. Para ello se ha de contar con caso (un evento o un ejemplar) al cuál se le asignará la predicción. Este caso ha de tener las mismas Variables Independientes (propiedades) que los casos en la muestra total de entrenamiento. Aunque cada árbol haya sido entrenado con un subconjunto de estas variables, siempre habrá un árbol que requiera las que los otros no contemplan, por lo que al final todas las Variables Independientes participarán en la predicción.

El caso susceptible de predicción será introducción a cada árbol del bosque, y de esta manera obtendremos tantas predicciones como árboles en el bosque. La predicción final será o bien la

predicción mayoritaria, en el caso de variables discretas, o bien el promedio, en caso de variables continúas.

Para probar el modelo podemos contar con un conjunto de casos nuevos previamente identificados en la Variable Dependiente (un histórico) o bien una parte de la muestra total reservada desde el principio para la validación final.

El esquema general de la secuencia se muestra en la figura 1. En esta figura se muestra un conjunto de entrenamiento del cual se extraen tres muestras aleatorias en cuanto a Variables Independientes y casos (en gris). Cada una de esas muestras genera un árbol de decisión (en azul). Aquí terminaría la fase de entrenamiento de los árboles. En la fase de validación, cada árbol predeciría la Variable Dependiente a partir de un conjunto de datos nuevos cuya Variable Dependiente está identificada también con antelación. Este conjunto de datos nuevos puede ser un conjunto reservado del conjunto de datos total. La predicción final sería la predicción mayoritaria de todos los árboles, un caso positivo en el caso de la figura. Si la predicción final del modelo coincide con la previamente establecida se contabilizaría un éxito. En caso contrario, fracaso. La proporción entre éxitos y fracasos determinará la eficiencia del modelo.

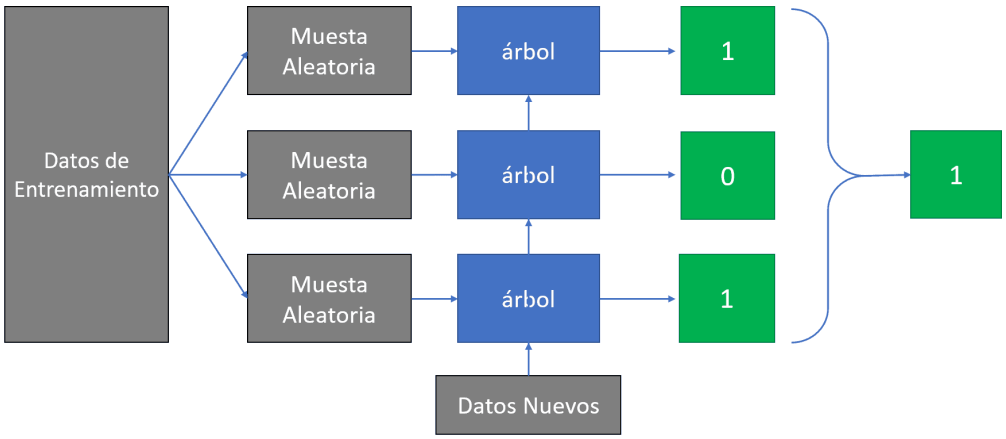

Figura 1. Tomada de https://www.iartificial.net/random-forest-bosque-aleatorio/

Código en R

A continuación, se muestra código R para aplicar Árboles Aleatorios con el conjunto de datos de IRIS:

```
#install.packages("randomForest")
library(RWeka)
library(caret)
library(datasets)
library(randomForest)

#se toma la muestra de iris
data<-iris
str(data)

# se indica que la especia es una variable discreta
data$Species <- as.factor(data$Species)
table(data$Species)

#se prepara el conjunto de entrenamiento y el de prueba
set.seed(222)
ind <- sample(2, nrow(data), replace = TRUE, prob = c(0.7, 0.3))
train <- data[ind==1,]
test <- data[ind==2,]

#se crea el bosque aleatorio con valores por defecto
rf <- randomForest(Species~., data=train, proximity=TRUE)
print(rf)

#se predice de nuevo con la muestra de entrenamiento.
p1 <- predict(rf, train)
confusionMatrix(p1, train$ Species)

#ahora sí se predice con la de prueba. Esta es la importante
p1 <- predict(rf, test)
confusionMatrix(p1, test$ Species)


#gráfico de error por especie y número de árboles

plot(rf)
```