Revista Pensamiento Americano
ISSN: 2027-2448. Vol 4 No. 7. Julio – Diciembre 2011
(Págs 75-80)

Revisión de los Árboles de Clasificación y Regresión (CART)

Juan Felipe Díaz Sepúlveda *
jsepulveda@coruniamericana.edu.co

Resumen
Dependiendo del problema, el propósito básico de un estudio de clasificación puede ser producir una correcta clasificación o descubrir la estructura predictiva del problema. Si nuestro objetivo es lo último, entonces estamos tratando de entender qué variables o interacciones de variables describen el fenómeno, esto es, dar caracterizaciones simples de las condiciones que determinan cuándo un objeto está en una clase más que en otra. Los Árboles de Clasificación y Regresión, en inglés Classification and Regression Trees (CART), deben su desarrollo a L. Breiman, J. Friedman, R. Olshen y C. Stone, autores del libro del mismo nombre, publicado en 1984 [Breiman y otros, 1984]. El objetivo de este artículo es dar a conocer desde el punto de vista teórico en qué consiste esta técnica de clasificación.

Palabras claves:
Clasificador, partición, árbol de clasificación, árbol de regresión, nodo, hoja, impureza, validación cruzada.

Abstract
Depending on the problem, the basic purpose of a classification study may be to produce a correct classification or predictive discovering the structure of the problem If our goal is the latter, then we are trying to understand what variables or interactions of variables describing the phenomenon, that is, give simple characterizations of the conditions that determine when an object is In a class more than another. The Classification and Regression Trees, Classification and Regression English Trees (CART), owes its development to L Breiman, J. Friedman, R. Olshen and C Stone, who wrote the book of the same name, published in 1984 [Breiman et al, 1984]. The aim of this paper is to report from the theoretical point of view it is this classification technique.

Key words:
Sorter, partition, classification tree, regression tree, node, leaf, impurity, cross validation.

Introducción:

Clasificadores como particiones

Definición 1.1. Un clasificador o regla de clasificación es una función d(x) definida en X tal que para todo x, d(x) es igual a uno de los números 1,2,...,J. Otra manera de mirar un clasificador es definir A_j como el subconjunto de X en el cual d(x)=j; esto es,

$$A_j=\{x \in X/d(x)=j\}$$

Los conjuntos $A_1,A_2,...,A_j$ son disjuntos y $X=U_j\ A_j$. Esto es, los A_j forman una partición de X.

Uso de datos en la construcción de clasificadores

Los clasificadores son construidos basándonos en experiencias pasadas.

En la construcción sistemática de clasificadores, estas experiencias pasadas son resumidas en una muestra de aprendizaje.

Definición 1.2. Una muestra de aprendizaje consiste de datos $(x_1,j_1),...,(x_N,j_N)$ sobre N casos donde

$$x_n \in X \text{ y } j_n \in \{1,...,J\},\ n=1,...,N.$$

* Candidato a M.Sc. en Ciencias-Estadística, Universidad Nacional de Colombia. Profesor de la Facultad de Ingeniería, Corporación Universitaria Americana, Medellín.
Artículo recibido: Diciembre 16/2011. Aceptado: Enero 28/2012.

La muestra de aprendizaje es denotada por L , es decir,

$L=\{(x_1,j_1),...,(x_N,j_N)\}$

Nosotros distinguimos dos tipos generales de variables que pueden aparecer en el vector de mediciones.

Definición 1.3. Una variable es llamada ordenada o numérica si sus valores de medida son números reales. Una variable es categórica si toma valores en un conjunto finito no teniendo un orden natural.

El propósito del análisis de clasificación

Dependiendo del problema, el propósito básico de un estudio de clasificación puede ser producir una correcta clasificación o descubrir la estructura predictiva del problema.

Si nuestro objetivo es lo último, entonces estamos tratando de entender qué variables o interacciones de variables describen el fenómeno, esto es, dar caracterizaciones simples de las condiciones que determinan cuándo un objeto está en una clase más que en otra.

Árboles de clasificación y regresión

Los Árboles de Clasificación y Regresión, en inglés Classification and Regression Trees (CART), deben su desarrollo a L. Breiman, J. Friedman, R. Olshen y C. Stone, autores del libro del mismo nombre, publicado en 1984 [Breiman y otros, 1984].

CART
Partimos de una muestra de entrenamiento
(1.1)
$(X_1,Y_1),...,(X_n,Y_n)$ $iid \sim \boldsymbol{\gamma}$,

donde cada $X_i=(X_i^1,....,X_i^p)$ es un vector con p variables aleatorias, que pueden ser todas continuas, todas discretas o mezclas de ambas. Las variables Y_i son unidimensionales, discretas o continuas. Con la muestra de entrenamiento, construimos una estructura del tipo árbol en dos etapas bien diferenciadas, en la primera, determinamos el llamado árbol maximal y en la segunda, aplicamos un procedimiento denominado de poda.

Veamos el siguiente ejemplo. Supongamos una muestra de entrenamiento como en (1.1), con p=2, y un árbol de 5 hojas como se representa en la Figura 1, donde las particiones se hacen en base a preguntas sobre los atributos X^1 y X^2 y los c_j son números reales.

En la Figura 2, se aprecia como el espacio $\mathbb{R}^2$ queda partido en regiones R_i, con i=1,....,5, donde cada R_i es un rectángulo de lados paralelos a los ejes.

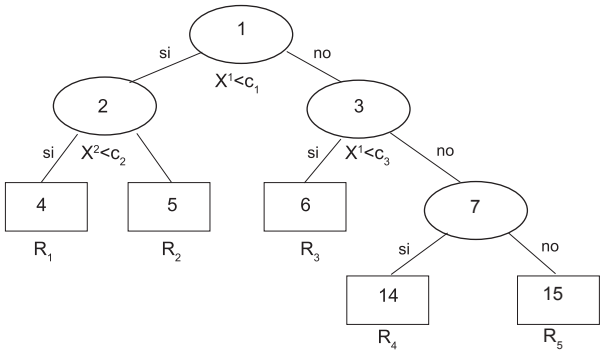


FIGURA 1. Ejemplo de un árbol de 5 hojas

Una vez construido el árbol maximal, el predictor le asigna a cada región (hoja) un determinado valor:

$$E[Y|(X^1,X^2)]=\sum_{j=1}^{5} \int_j (X^1,X^2) 1_{\{(X^1,X^2)\in R_j\}}$$

donde, si Y es continua (árbol de regresión) estimamos $\int_j$ por

$$\hat{\int}_j = \frac{\sum\{i:(X^1,X^2)\in R_j\} Y_i}{\text{card}\{i:(X^1,X^2)\in R_j}$$

(Promedio de los Y_i en la región R_j)

y si Y es discreta (árbol de clasificación)

$\hat{\int}_j$ = la clase más frecuente en R_j

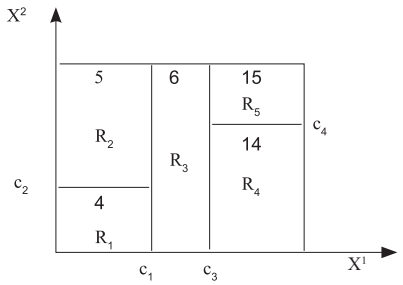


FIGURA 2. Partición del espacio $\mathbb{R}^2$, según árbol de la figura 1

Una de las características fundamentales de CART, es que luego de obtenido un árbol maximal se inicia una etapa de poda, en la cual se eliminan algunas de sus ramas. Este proceso, se explica más adelante.

Árboles de clasificación

La estructura general y gran parte de los algoritmos de CART son similares para regresión y clasificación, aquí solo abordaremos el desarrollo de los árboles de clasificación, que corresponden al caso en que Y

toma valores en $\{1,...,J\}\subset\mathbb{N}$

Reglas de partición

El objetivo fundamental, que nos planteamos al

construir una partición, es optimizar la homogeneidad de las regiones resultantes. En cada nodo del árbol, nos proponemos aumentar la pureza de los dos nodos obtenidos.

Las reglas de partición en un nodo, dependen exclusivamente de los atributos, por lo cual son las mismas tanto para clasificación como para regresión. Para el caso de atributos discretos, supongamos que

X^j toma valores en un conjunto finito $\{1,...,H\}\subset$ entonces las reglas son de la forma

$X^j \in C \text{ con } C\subset\{1,....,H\}$

Lo que hacemos en CART, es ordenar los valores que efectivamente toma cada atributo X^j, sobre la muestra de entrenamiento, elegir un punto intermedio m entre cada par de valores consecutivos y solamente considerar las reglas

$X^j \leq m$

Por lo tanto, el número de todas las posibles reglas es a lo sumo n-1.

Criterio de partición de un nodo

Entre todas las reglas posibles de partición de un nodo, debemos elegir la que mejor contribuya al aumento de la homogeneidad de sus dos hijos. Esto se logra, definiendo una medida de impureza sobre la variable de respuesta. Aquí sí, es fundamental tener en cuenta que estamos considerando que la variable Y es discreta.

Empecemos definiendo función de impureza, como una función ϕ definida sobre un conjunto de

$$\text{J-uplas } (p_1,\ldots,p_J) \in \ ^J, \text{ tales que}$$

$$p_j \geq 0, \quad \forall j=1,\ldots,J$$

$$\sum_{j=1}^{J} pj = 1$$

que verifica las siguientes propiedades:

ϕ tiene un único máximo en $(\frac{1}{J},\ldots,\frac{1}{J})$

ϕ tiene mínimo 0 y solamente lo alcanza en los puntos de la forma

(1,0,....,0),(0,1,....,0),....,(0,0,....,1),

ϕ es una función simétrica de $(p_1,...,p_J)$

Dada una función de impureza ϕ, podemos definir para cada nodo *t* de un árbol su medida de impureza *i(t)*, como

$i(t)=\phi[(p_1(t),...,p_J(t)]$

donde $p_j(t)$ es la probabilidad condicional de que un elemento pertenezca a la clase j, dado que pertenece al nodo t y puede estimarse en la práctica, como la proporción de elementos de clase j en el nodo t. Es claro, a partir de estas definiciones, que la máxima impureza en el nodo t se da cuando todas las clases están igualmente representadas y la mínima se obtiene cuando en t hay casos de un único tipo.

Algunos de los criterios más utilizados en CART, como medida de impureza de un nodo, son:

1) La medida de entropía

$i_{ent}(t)=-\sum_{j=1}^{J} p_j(t)\log(p_j(t))$, definiendo 0 *log (0)=0

El índice Gini

$$i_{Gini}(t)=-\sum_{\substack{j=1\\ i=j}}^{J} p_i(t)\,p_j(t)=1-\sum_{j=1}^{J}[p_j(t)]^2$$

En el libro de Breiman [Breiman y otros, 1984, pag 38], se afirma que la elección de la medida de impureza más adecuada depende del problema y que el predictor construido, no parece ser muy sensible a dicha elección.

Supongamos entonces, que mediante una regla hemos partido al nodo t en dos, t_I (nodo izquierdo) y t_D (nodo derecho). Sea p_I la proporción de elementos del nodo t que caen en el hijo izquierdo y p_D la del derecho.

Establecemos una medida de bondad de una partición s, para un nodo t, de la siguiente manera:

$$\triangle i(s,t)=i(t)-p^i_I(t_I)-p^i_D(t_D)\geq 0.$$

Es claro, que el aumento de la bondad depende de la disminución de la impureza, en los nodos hijos en relación al nodo padre. El criterio de selección de la mejor partición s^* en el nodo t, consiste en elegir aquella que proporciona la mayor bondad

$$\triangle i(s^*,t)=\max_{s\in\psi}\{\triangle i(s,t)\}$$

donde ψ es el conjunto de todas las particiones posibles del nodo t.

Impureza del árbol

Mediante las reglas de partición y comenzando desde el nodo raíz, se van partiendo sucesivamente los nodos. Una vez que termina el proceso de partición se obtiene un árbol, al cual nos va a interesar medirle la impureza, es decir cuantificar conjuntamente la impureza de todas sus hojas. Para eso, definimos la impureza del árbol A de la siguiente manera

$$I(A)=\sum_{t\in\tilde{A}} i(t)p(t),$$

donde Ã es el conjunto de hojas del árbol A y p(t) es la probabilidad de que un caso pertenezca al nodo t.

En Breiman y otros, 1984, pag 32-33 los autores demuestran un resultado fundamental, las selecciones sucesivas que maximizan $\triangle i(s,t)$ en cada nodo, son equivalentes a las que se realizarían con el fin de minimizar la impureza global del árbol. Esto significa, que la estrategia de selección de la mejor división en cada nodo, lleva a la solución óptima, en términos del árbol final.

Regla de asignación de clases

Una vez determinado que un nodo es terminal (hoja), corresponde asignarle una clase. La manera más habitual de hacerlo es por el método del voto mayoritario, que consiste en asignarle al nodo t la clase j* si

$$p_{j*}(t) = \text{máx}_{j=1,\ldots,J}\{p_j(t)\}$$

y en caso de empate hacer un sorteo.

Reglas de parada

Hasta ahora no hemos explicado cómo detener el proceso de partición, es decir cuándo declarar a un nodo lo suficientemente puro y que no se justifique partirlo.

De la forma como resolvamos este problema, dependerá el tamaño del árbol construido. Un árbol muy grande, generará sobre-ajuste al conjunto de entrenamiento y uno muy pequeño, puede contribuir a que se pierda parte importante de la estructura de los datos.

Un posible criterio de parada, podría ser el de utilizar las medidas de impureza definidas anteriormente, declarando que un nodo es terminal si la disminución de impureza no supera determinado umbral. Umbrales muy bajos, generan árboles muy grandes, con los inconvenientes ya comentados.

En cambio, umbrales mayores, pueden implicar que un nodo no se divida, cuando en realidad, una posterior partición de sus descendientes, sí está en condiciones de generar un buen decrecimiento de impureza. Otros criterio utilizados son, evitar partir un nodo si la cantidad de elementos es menor que un determinado umbral, por ejemplo menor que 5, o si algunos de los dos nodos, que resultan de la partición óptima, no supera un umbral.

Por lo expuesto, el tamaño del árbol es un parámetro de ajuste fundamental para el modelo, por lo que debería escogerse en función de los datos (aprendiendo de los datos).

La solución a la que llegan Breiman y demás autores, en [Breiman y otros, 1984], es primero construir un árbol llamado maximal, con la única condición de no permitir nodos con muy pocos elementos, para luego aplicarle un proceso denominado poda. Este consiste, en tomar el árbol maximal y sacarle aquellas ramas o sub-árboles que determinen beneficios muy pequeños, en lo respecta a la disminución de la impureza.

Con este procedimiento se obtiene un sub-árbol, que permite para determinados nodos, que una de sus ramas permanezca y la otra se pode, al contrario de los criterios de parada anteriormente mencionados, en los cuales el efecto es el equivalente a podar simultáneamente ambas ramas.

La construcción del árbol óptimo, la haremos a partir de un proceso de selección de subárboles, en la que interviene de manera fundamental el error asociado a cada uno de ellos.

Poda del árbol

El árbol obtenido es generalmente sobreajustado por tanto es podado, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño “adecuado” del árbol. Breiman et˜al. (1984) introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol.

Computacionalmente el procedimiento descrito es complejo. Una forma es buscar una serie de árboles anidados de tamaños decrecientes (De’ath & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño.

Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de costo complejidad, $R\alpha$ (T).

Para cada árbol T , la función costo - complejidad se define como (Deconinck et˜al., 2006):

$$R\alpha\,(T) = R(T) + \alpha|\tilde{T}| \qquad (1.9)$$

donde R(T) es el promedio de la suma de cuadrados entre los nodos, puede ser la tasa de mala clasificación total o la suma de cuadrados de residuales total dependiendo del tipo de árbol, |T | es la complejidad del árbol, definida como el número total de nodos del subárbol y α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero, Cuando α=0 se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol.

La función R_α (*T*) siempre será minimizado por el árbol más grande, por tanto se necesitan mejores estimaciones del error, para esto Breiman et˜al. (1984) proponen obtener estimadores “honestos” del error por “validación cruzada”.

Computacionalmente el procedimiento es exigente pero viable, pues sólo es necesario considerar un

árbol de cada tamaño, es decir, los árboles de la secuencia anidada.

Selección del árbol óptimo

De la secuencia de árboles anidados es necesario seleccionar el árbol áptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad (De'ath & Fabricius, 2000), por tanto se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada.

El objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones.

El procedimiento de validación cruzada puede implementarse de dos formas:

Si se cuenta con suficientes datos se parte la muestra, sacando la mitad o menos de los datos y se construye la secuencia de árboles utilizando los datos que permanecen, luego predecir, para cada árbol, la respuesta de los datos que se sacaron al iniciar el proceso; obtener el error de las predicciones; seleccionar el árbol con el menor error de predicción.

En general no se cuenta con suficientes datos como para utilizar el procedimiento anterior, de modo que otra forma sería:

Validación cruzada con partición en V, (v-fold cross validation, se menciona más adelante).

La idea básica de la "Validación cruzada" es sacar de la muestra de aprendizaje una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto sacado es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como "datos nuevos".

El desempeño entendido como el error de predicción, es acumulado para obtener el error medio absoluto del conjunto de prueba.

Como se mencionó anteriormente, para la metodología CART generalmente se utiliza Validación Cruzada con partición en V (v-fold cross validation), tomando V = 10 y el procedimiento es el siguiente:

Dividir la muestra en diez grupos mutuamente excluyentes y de aproximadamente igual tamaño.

Sacar un conjunto por vez y construir el árbol con los datos de los grupos restantes. El árbol es usado para predecir la respuesta del conjunto eliminado.

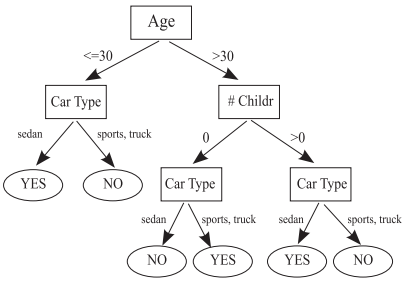
Calcular el error estimado para cada subconjunto. Repetir los "ítems" dos y tres para cada tamaño de árbol.

Seleccionar el árbol con la menor tasa de mala clasificación. Al llegar a este punto se procede a analizar el árbol obtenido.

Ejemplo:

Como ejemplo suponga el árbol y los datos en la Figura 1.1, donde se quiere determinar un conjunto de reglas que indiquen si un conductor vive o no en los suburbios.

Se concluye:



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

FIGURA 3. Ejemplo de árbol de clasificación

Si Age≤30 y CarType = Sedan entonces Si
Si Age≤30 y CarType = truck/Sports entonces No
Si Age > 30, Children = 0 y CarType = Sedan entonces No
Si Age > 30, Children = 0 y CarType = truck/Sports entonces Si
Si Age > 30, Children > 0 y CarType = Sedan entonces Si
Si Age > 30, Children > 0 y CarType = truck/Sports entonces No

Conclusión

La metodología CART es una técnica de clasificación no paramétrica, la cual se puede usar como alternativa a otras técnicas estadísticas o heurísticas en el análisis de clasificación. Uno de los trabajos que se ha venido adelantando en los últimos años es la comparación de la eficiencia de esta técnica bajo ciertas condiciones frente a otras comúnmente utilizadas.

Bibliografía

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. G. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California, USA.

Dobra, A. (2002). Classification and regression tree construction. Thesis proposal, Departament of Computer Science, Cornell University, Ithaca NY.

Serna, S. C. (2009). Comparación de Árboles de Regresión y Clasificación y Regresión Logística. Tesis de Maestría en Estadística, Facultad de ciencias, Universidad Nacional de Colombia, Medellín.

Roche, A. (2009). Árboles de decisión y Series de tiempo. Tesis de Maestría en Ingeniería Matemática, Facultad de Ingeniería, UDELAR.