Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Tipos de Árboles de decisión

1

Profesora: *Dra. Fabiola Ocampo Botello*

Algoritmo ID3

Rokach & Maimon (2015) y Bhumika, Aditya, Akshay, Arpit & Naresh (2017) establecen que el ID3 tiene las siguientes características:

- Es un algoritmo desarrollado por Ross Quinlan.
- Sólo acepta atributos categóricos
- Usa la ganancia de información como criterio de división.
- Deja de crecer cuando:
  - Todas las instancias pertenecen a un solo valor de una característica objetivo o
  - Cuando la mejor ganancia de información no es mayor que cero.
- No aplica ningún procedimiento de poda.
- No maneja atributos numéricos o valores faltantes.

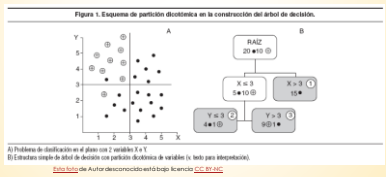
2

Continuación ID3

Rokach & Maimon (2015) mencionan que el ID3 tiene varias desventajas:

- No garantiza una solución óptima.
- Puede sobreajustar los datos de entrenamiento.
- Está diseñado para atributos nominales.

ID3 significa algoritmo iterativo de dicotomizador 3



3

Continuación ID3

Dunham (2002) presenta las siguientes características del algoritmo ID3:

- La técnica ID3 para construir un árbol de decisión se basa en la teoría de la información y los intentos de minimizar el número esperado de comparaciones.
- La idea básica del algoritmo de inducción es hacer preguntas cuyas respuestas brinden la mayor cantidad de información.
- La estrategia básica utilizada por ID3 es elegir primero la división de atributos con la mayor ganancia de información.
- La cantidad de información asociada con un valor de atributo está relacionada con la probabilidad de ocurrencia.
- El concepto utilizado para cuantificar la información se llama entropía. La entropía se usa para medir la cantidad de incertidumbre o sorpresa o aleatoriedad en un conjunto de datos.

4

Continuación ID3

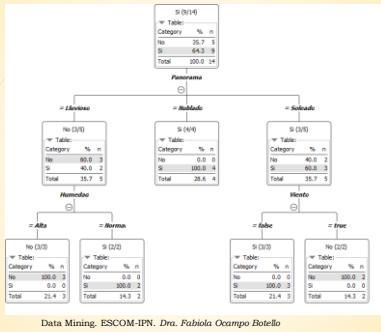
Ejemplo de un árbol ID3:

Tomando como referencia las características climáticas, un modelo que generalice las condiciones necesarias para determinar la posibilidad de ir o no a jugar golf.

Continuación ID3

JuegaGolf	Panorama	Temperatura	Humedad	Viento
No	Lluvioso	Caliente	Alta	FALSO
No	Lluvioso	Caliente	Alta	VERDADERO
Si	Nublado	Caliente	Alta	FALSO
Si	Soleado	Templado	Alta	FALSO
Si	Soleado	Frío	Normal	FALSO
No	Soleado	Frío	Normal	VERDADERO
Si	Nublado	Frío	Normal	VERDADERO
No	Lluvioso	Templado	Alta	FALSO
Si	Lluvioso	Frío	Normal	FALSO
Si	Soleado	Templado	Normal	FALSO
Si	Lluvioso	Templado	Normal	VERDADERO
Si	Nublado	Templado	Alta	VERDADERO
Si	Nublado	Caliente	Normal	FALSO
No	Soleado	Templado	Alta	VERDADERO

ación ID3



Algoritmo C4.5

Rokach & Maimon (2015) y Bhumika et al. (2017) establecen que el algoritmo C4.5 tiene las siguientes características:

- Es una evolución de ID3.
- Fue desarrollado por Ross Quinlan.
- Puede manejar atributos numéricos.
- Utiliza la relación de ganancia como criterio de división.
- Es n-ario con valores discretos y binario con datos continuos.
- La poda basada en errores se realiza después de la fase de crecimiento.

- La división termina cuando el número de instancias a dividir está por debajo de un cierto umbral.

9

El algoritmo C4.5 proporciona varias mejoras para ID3. Las mejoras más importantes según Rokach & Maimon (2015) son:

(1) C4.5 utiliza un procedimiento de poda que elimina las ramas que no contribuyen a la precisión y las reemplaza con nodos foliares.

(2) C4.5 permite que falten valores de atributos (marcados como ?).

(3) C4.5 maneja atributos continuos dividiendo el rango de valores del atributo en dos subconjuntos (división binaria). Específicamente, busca el mejor umbral que maximice el criterio de relación de ganancia. Todos los valores por encima del umbral constituyen el primer subconjunto y todos los demás valores constituyen el segundo subconjunto.

Algoritmo C5.0

Rokach & Maimon (2015) expresan que el algoritmo C5.0 es una versión comercial actualizada de **C4.5** que ofrece una serie de mejoras: se afirma que C5.0 es mucho más eficiente que C4.5 en términos de memoria y tiempo de cálculo. Además, es compatible con el procedimiento de refuerzo que puede mejorar el rendimiento predictivo.

Algoritmo J48

10

Rokach & Maimon (2015) mencionan que el algoritmo J48 es una implementación Java de código abierto del algoritmo **C4.5** en la herramienta de minería de datos Weka. Debido a que el algoritmo J48 es simplemente una reimplementación de C4.5, se espera que funcione de manera similar a C4.5. Sin embargo, un estudio comparativo reciente que compara C4.5 con J48 y C5.0 [Moore et al. (2009)] indica que C4.5 tiene un rendimiento consistentemente mejor (en términos de precisión) que C5.0 y J48 en particular en conjuntos de datos pequeños.

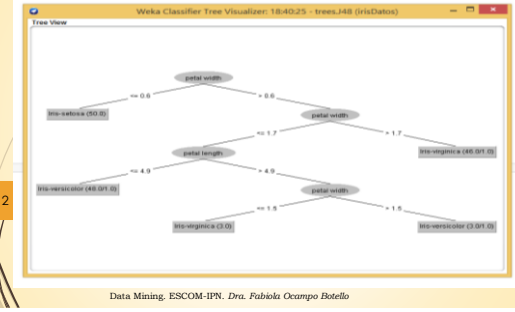
Ejemplo de un árbol J48 (C4.5):

11

Considerando las características descriptivas que tiene una flor de iris, crear un modelo que generalice la identificación de tres tipos de iris: iris versicolor, iris setosa e iris virginica.

Ejemplo de un árbol J48 (C4.5):

12



ID3 y C4.5

Dunham (2002) establece que **el algoritmo del árbol de decisión C4.5 mejora al ID3 en los siguientes aspectos** (se presentan algunos):

- **Datos faltantes**: cuando se crea el árbol de decisión, los datos faltantes simplemente se ignoran. Es decir, la relación de ganancia se calcula considerando sólo los otros registros que tienen un valor para ese atributo. Para clasificar un registro con un valor de atributo faltante, el valor para ese elemento puede predecirse en función de lo que se sabe sobre los valores de atributo para los otros registros.
- **Datos continuos**: la idea básica es dividir los datos en rangos basados en los valores de atributo para ese elemento que se encuentran en la muestra de entrenamiento.

ID3 y C4.5

- **Poda**: hay dos estrategias principales de poda propuestas en C4.5 (Dunham, 2002):
  - Con el reemplazo del subárbol, un subárbol se reemplaza por un nodo hoja si este reemplazo da como resultado una tasa de error cercana a la del árbol original.
  - Otra estrategia de poda, llamada elevación de subárbol, reemplaza un subárbol por su subárbol más utilizado.

Larose & Larose (2015) establecen que el algoritmo C4.5 es la extensión de Quinlan de su propio algoritmo iterativo de dicotomizador 3 (ID3) para generar árboles de decisión. Al igual que con CART, el algoritmo C4.5 visita recursivamente cada nodo de decisión, seleccionando la división óptima, hasta que no se produzcan más divisiones posibles.

ID3 y C4.5

Sin embargo, existen las siguientes diferencias interesantes entre CART y C4.5 (Larose, T. Daniel & Larose, D. Chantal, 2015) :

- A diferencia de CART, el algoritmo C4.5 no está restringido a divisiones binarias. Mientras que CART siempre produce un árbol binario, C4.5 produce un árbol de forma más variable.
- Para los atributos categóricos, C4.5 por defecto produce una rama separada para cada valor del atributo categórico. Esto puede resultar en más "arbustos" de lo deseado, porque algunos valores pueden tener baja frecuencia o pueden estar asociados naturalmente con otros valores.
- El método C4.5 para medir la homogeneidad de los nodos es bastante diferente del método CART y se examina en detalle a continuación.

El algoritmo C4.5 utiliza el concepto de ganancia de información o reducción de entropía para seleccionar la división óptima.

Árbol CART

El árbol CART (Bhumika et al. (2017).

- Significa árboles de clasificación y regresión (*Classification And Regression Trees*).
- Fue presentado por Breiman en 1984.
- El algoritmo CART construye árboles de clasificación y regresión.
- CART construye el árbol de clasificación mediante la división **binaria** del atributo.
- El índice de Gini se usa para seleccionar el atributo de división
- Permite datos de atributos continuos y nominales.

C4.5 y CART

Sin embargo, existen las siguientes diferencias interesantes entre CART y C4.5 (Larose & Larose, 2015) :

- A diferencia de CART, el algoritmo C4.5 **no** está restringido a divisiones binarias. Mientras que **CART siempre produce un árbol binario**, C4.5 produce un árbol de forma más variable.
- Para los atributos categóricos, **C4.5 por defecto produce una rama separada para cada valor del atributo categórico**. Esto puede resultar en más "arbustos" de lo deseado, porque algunos valores pueden tener baja frecuencia o pueden estar asociados naturalmente con otros valores.

17

Ejemplo de un árbol CART:

Ejercicio adaptado con fines educativos de:
Portal IBM. SPSS Statistics 23.0.0. Casos de estudio. Disponible en:
https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/tutorials/trees_scoring_intro1.html

Descripción del enunciado:

Considerando un conjunto de datos que contiene información demográfica y el precio de compra del vehículo. Construir un modelo que se puede usar para predecir cuánto es probable que las personas con características demográficas similares gasten en un automóvil nuevo. El modelo creado podrá ser aplicado a otros archivos de datos donde la información demográfica está disponible, pero no la información sobre compras anteriores de vehículos.

18

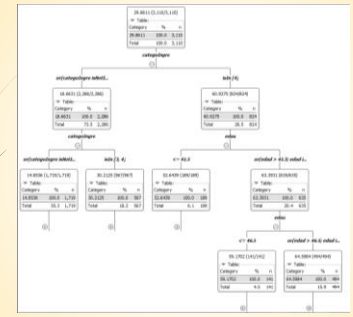
CART

Diccionario de datos:

Nombre	Descripción	Tipo	Dominio
coche	Precio del vehículo principal	Numérico	
edad	Edad en años	Numérico	
sexo	sexo	Nominal	f = femenino m = masculino
cating	Categoría de ingresos en miles	Ordinal	1.00 = "Menos de $25" 2.00 = "$25 - $49" 3.00 = "$50 - $74" 4.00 = "$75+"
educ	Nivel de estudios	Ordinal	1 = "No completó el bachillerato" 2 = "Bachillerato" 3 = "Estudios universitarios" 4 = "Licenciado" 5 = "Estudios de post-grado"
e_civil	Estado civil	Nominal	0 = "Sin casar" 1 = "Casado"

19

CART



20

Referencias bibliográficas

Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications (0975–8887). Volume 163 – No 8, April 2017.

Dunham, M. H. (2002). Data mining: introductory and advanced topics. Prentice Hall.

Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.