

Sistema de recomendación de atención odontológica mediante arboles de decisión

Recommendation system for dental care through decision trees

Augusto Cortez Vásquez, Luis Alberto Cortez Vásquez¹

Resumen

El presente artículo describe los arboles de decisión como herramientas de desarrollo y muestra como aprenderlos a construir. Esta técnica puede considerarse entre las mas utilizadas dentro del área de aprendizaje automático para la obtención de modelos para problemas de clasificación que admiten atributos numéricos(continuos) y nominales(discretos). Se trata el caso de recomendación de atención odontológica mediante arboles de decisión utilizando un conjunto de entrenamiento. Utilizaremos el aprendizaje inductivo a partir de casos particulares para construir un modelo basado en árbol de decisión.

Palabras claves

Árbol de decisión, aprendizaje, aprendizaje inductivo, Sistema de recomendación odontológica

Abstract

This paper describes the decision trees as tools for development and shows how to learn. This technique can be considered among the most used within the area of machine learning to obtain models for classification problems that support numeric attributes (continuous) and nominal (discrete). It is recommended for attention by decision trees using a training set. Use inductive learning from individual cases to build a model based on decision tree.

Key words

Decision tree, learning, inductive learning, dental recommendation system

1. Introducción

La inducción de aboles de decisión₁, constituyen una de las técnicas mas sencillas para construir algoritmos de aprendizaje₂. Un árbol de decisión desarrolla una secuencia de pruebas(test) para poder alcanzar una decisión. Cada nodo interno del árbol corresponde con la prueba sobre el valor de una de las propiedades, y las ramas que salen del nodo están etiquetadas con los posibles valores de una de dicha propiedad. Cada nodo hoja representa el valor que ha de ser devuelto si dicho nodo es alcanzado. EL presente artículo presenta la construcción de un modelo basado en árbol de decisión a partir de un conjunto de entrenamiento de datos

2. Marco conceptual

La solución de muchos problemas pueden representarse mediante arboles, por lo que resolver tales problemas se convierte en un problema de tratamiento de arboles. En este articulo se tratara específicamente de los arboles de decisión.

Árbol de decisión

Un árbol de decisión AD es una representación grafica de un procedimiento para clasificar o evaluar un concepto. Por ejemplo decidir si se recomienda atención odontológica a partir de datos como el tiempo de espera, urgencia etc.

"Intuitivamente un árbol de decisión es una colección de condiciones organizada jerárquicamente. Formalmente, un árbol de decisión es un árbol donde cada nodo representa una condición o test sobre algún atributo y cada rama que parte de ese nodo corresponde a un posible valor de ese atributo" [8] PALMA, 2008 pp 727)

Los atributos descritos pueden ser discretos o continuos. Aprender de una función continua se denomina regresión, mientras que aprender de una función con valores discretos se denomina clasificación [10]

Expresividad de los arboles de decisión

Desde la época de Porfirio₃, se ha considerado la importancia de los arboles para propósitos de distribución y clasificación. La imagen del árbol no están importante como la configuración arbórea. El principio de esta es la ramificación. Los arboles de decisión pueden expresar exactamente lo mismo que los lenguajes de tipo proposicional; es decir cualquier función booleana puede ser descrita como un árbol de decisiones puede hacer corresponder a cada fila de la tabla de verdad una función correspondiente a cada camino del árbol. Esto llevaría a un árbol de decisión con un crecimiento exponencial, ya que las filas de cada tabla de verdad crecen de forma exponencial a medida que se introducen nuevos atributos[4,11].

Aprendizaje inductivo

Solomonof define aprendizaje desde una visión mas teórica y matemática como eliminación de la redundancia, vista como comprensión de información. Esto significa que el aprendizaje nos permite identificar regularidades en un conjunto de observaciones[6].

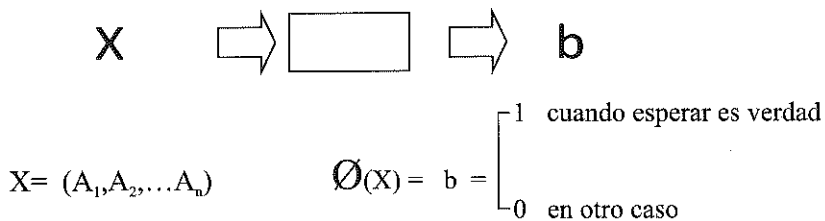
El aprendizaje inductivo es un tipo especial de aprendizaje que parte de casos particulares (ejemplos) y obtiene casos generales (reglas o modelos) que generalizan o abstraen la evidencia [9,10]. La idea es construir una función que tenga el comportamiento observado en los datos.

La tarea de aprendizaje para lo cual los arboles de decisión se adecuan mejor es la clasificación, de hecho clasificar es determinar de entre varias clases a que clase pertenece un objeto [Hernandez, pp 283]

Lo que se pretende es construir modelos "comprensibles" y "proposicionales", en donde lo primero se refiere al hecho de que el modelo se pueda expresar en forma simbólica, en forma de conjunto de condiciones, en este caso representados en los nodos de un árbol. En segundo termino "proposicional" se refiere a que los métodos se restringen a algoritmos que aprenden modelos sobre una única tabla de datos y que no establecen relaciones entre mas de una fila de la tabla a la vez ni sobre mas de un atributo a la vez, por tal motivo también se le denomina reglas atributo-valor, dado que las condiciones se expresan sobre el valor de un único atributo.

3. Construcción de arboles de decisión a partir de ejemplos

A partir de un conjunto de datos denominado conjunto de entrenamiento se puede construir un árbol de decisión. Un ejemplo para un árbol de decisión booleano consiste de un vector de atributos de entrada X , y un único valor booleano b .



Los ejemplos positivos son aquellos en los que la meta ESPERAR es verdadera; los ejemplos negativos son aquellos en la que es falsa [10].

Consideremos el caso siguiente:

Meta: Esperar. El propósito es aprender una definición para el predicado meta Esperar. Primero se tiene que establecer que atributos están disponibles para describir los ejemplos del dominio.

Supongamos la siguiente lista de atributos:

(Es_Urgente, Hay_Paciente, Tiene_Cita, Esperar, Alternativa)

A_1 : Es_Urgente: si es urgente o no

A_2 : Hay_Pacientes: cuantos pacientes están esperando atención (ninguno, algunos muchos)

A₃: **Tiene_Cita**: si ha hecho una reserva de cita o no

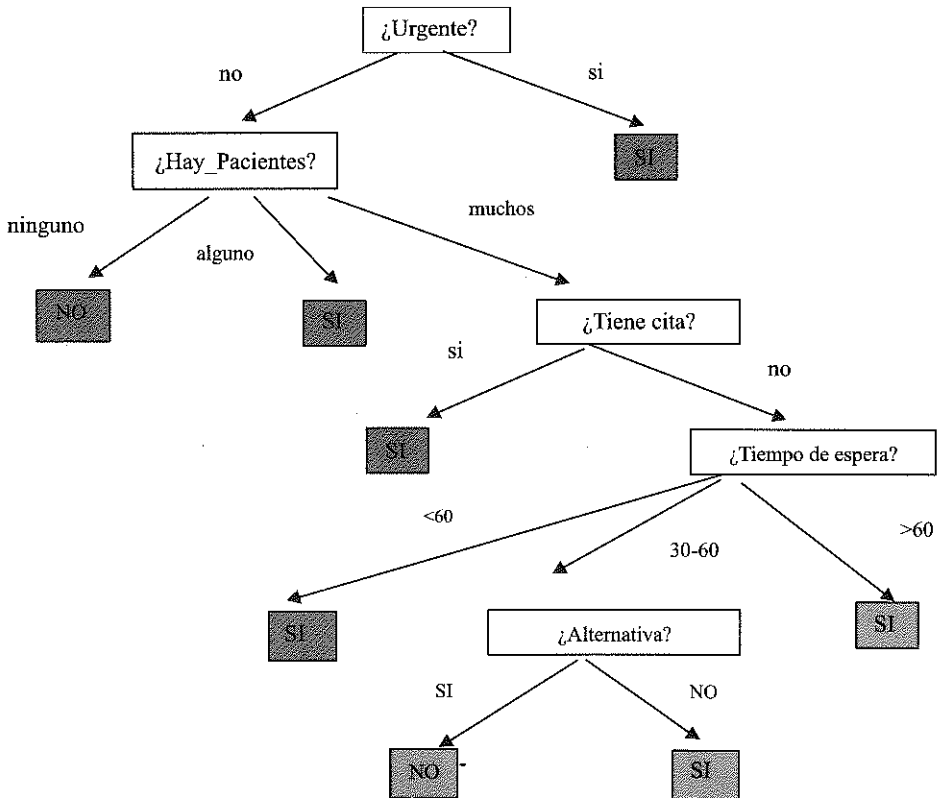
A₄: **Esperar**: Tiempo de espera estimado por recepción(menos de 30 minutos, entre 30 y 60 minutos, mayor a 60 minutos)

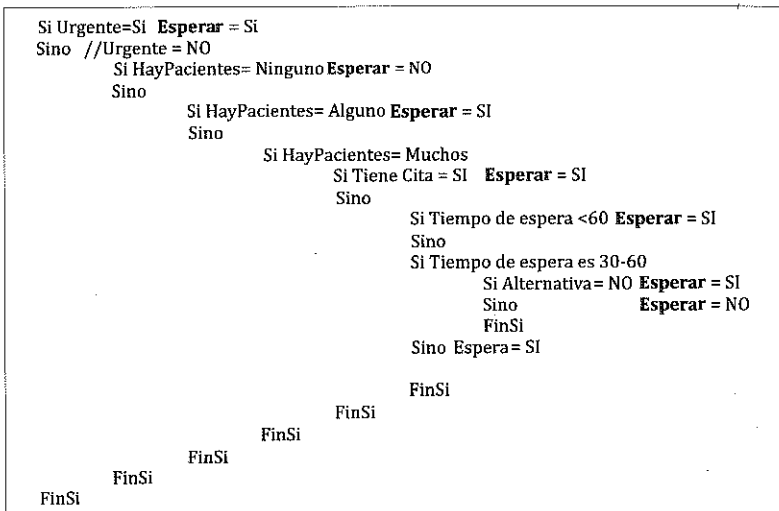
A₅: **Alternativa**: Si existe un consultorio alternativo que pueda prestar atención

Consideremos el siguiente conjunto de entrenamiento:

Ejemplo	Urgente	HayPacientes	TieneCita	TiempoEspera	Alternativa	Meta
E ₁	No	Ninguno	Si	0-30	No	No
E ₂	Si	Muchos	Si	30-60	Si	No
E ₃	No	Algunos	No	>60	Si	Si
E ₄	Si	Ninguno	No	30-60	No	Si
E ₅	No	Muchos	Si	>60	No	No
E ₆	No	Algunos	Si	30-60	No	Si
E ₇	No	Algunos	No	>60	No	No
E ₈	Si	Ninguno	No	0-30	Si	Si
E ₉	No	Algunos	Si	30-60	Si	Si
E ₁₀	No	Muchos	No	>60	No	No

A partir del conjunto de entrenamiento se puede construir el árbol de decisión siguiente:





Como puede verse el tipo de reglas resulta comprensible porque se construye usando las propiedades y atributos de los mismos datos que se han utilizado para inducir regla

Si (Urgente=SI) Espera=SI

Si (HayPacientes = muchos y Tiempo Espera >60) Espera = NO

Los algoritmos de arboles de decisión se pueden clasificar, como bien indica Palma en "Inteligencia artificial", atendiendo al principio utilizado en la generación de reglas:

A partir del conjunto de entrenamiento, podríamos construir un árbol de decisión que tenga un camino hasta cada hoja por cada ejemplo, donde el camino comprueba cada atributo y sigue el valor que tiene en el ejemplo hasta llegar a una hoja que contiene la clasificación del mismo. Cuando se proporciona el mismo ejemplo otra vez, el árbol de decisión propondrá la clasificación correcta. Desafortunadamente no tendrá mucho que decir sobre cualquier otro caso.

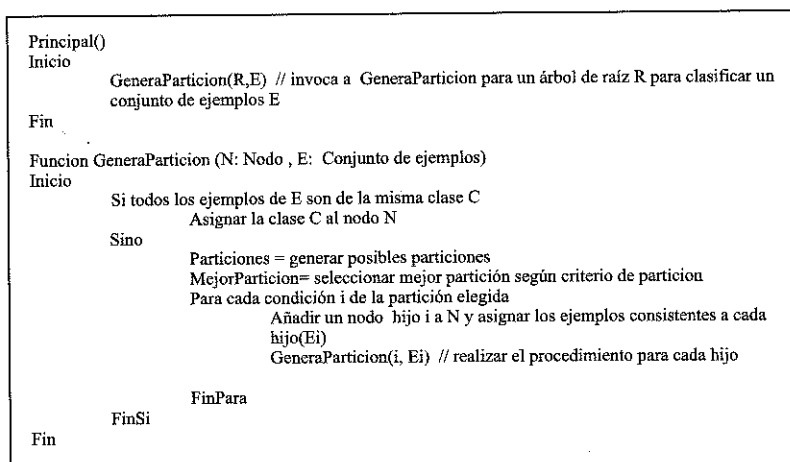
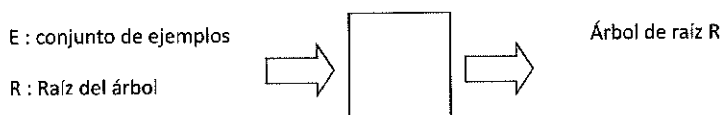
Se pone de relieve el problema suscitado con esta solución, en el sentido de que solo se memoriza las soluciones pero no extrae ningún patrón a partir de los ejemplos, de tal forma que no se podrá extrapolar a ejemplos que no le han sido proporcionados. Aplicando la navaja de Ockham⁴, deberíamos encontrar en vez de este árbol, el árbol de decisión mas pequeño que sea consistente con los ejemplos. Desafortunadamente, encontrar el árbol mas pequeño es un problema intratable. La idea básica de aprendizaje-arbol-decisión es realizar primero el test sobre el atributo mas importante. Se considera como "atributo mas importante" aquel que des crimina mas claramente los ejemplos. Nos hemos referido a la importancia de los atributos con el objetivo de obtener la clasificación correcta con un número de test pequeño, es decir, que todos los caminos en el árbol sean cortos y así el árbol completo será pequeño

Algoritmos basados en principio "Divide y Vencerás", también llamados "basados en criterios de partición": consiste en ir partiendo sucesivamente los datos en función del valor del atributo seleccionado cada vez. Esta es la estrategia que siguen los algoritmos de inducción de árboles de decisión que estudiaremos en este artículo.

Algoritmos basados en el principio de "Separa y Vencerás", también denominados "basados en criterio de cobertura): consiste en encontrar condiciones de las reglas que cubran la mayor cantidad de ejemplos de una clase y la menor del resto de las clases. Estas técnicas no se abordaran en este artículo.

Generación del árbol de decisión a partir de un conjunto de ejemplos, utilizando la técnica de partición

Cuadro A: Algoritmo de Creación del Árbol



Los árboles de decisión se adecuan mejor para tareas de clasificación, esto es determinar de entre varias clases a que clase pertenece un objeto; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema. El algoritmo va construyendo el árbol desde el árbol que contiene solo la raíz, añadiendo particiones y los hijos resultantes de cada partición. En cada partición, los ejemplos se van dividiendo entre los hijos. Finalmente, se llega a la situación en la que todos los ejemplos que caen en los nodos inferiores son de la misma clase y esa rama ya no sigue creciendo.

Elección de atributos de prueba

Advirtamos, que no obstante el árbol este bien construido, desde el punto de vista de aprendizaje interesa la forma de elegir los atributos que conformaran los test en cada uno de los nodos no hojas.. El esquema que se utiliza para seleccionar atributos esta diseñado para minimizar la profundidad del árbol final[8,9]. La idea es elegir el atributo que proporcione una clasificación lo mas exacta posible de los ejemplos.

4. Discusión

Vimos al principio de este artículo que frente a un conjunto de atributos en espacios de alta dimensionalidad y en problemas que mezclen datos categóricos y numéricos, es posible construir varios árboles de decisión. No obstante, los problemas de los árboles de decisión que adquieren más relieve al momento de construir el árbol son los relacionados a elegir el atributo más importante sobre el que se aplicará el test. Un árbol de decisión toma como entradas un objeto o una situación descrita a través de un conjunto de atributos y devuelve una decisión. Los árboles de decisión siguen una aproximación "divide y vencerás" para partir el espacio del problema en subconjuntos.

Existen muchos métodos de árboles de decisión que difieren entre sí en la forma de crear el árbol. Lo que más interesa no es utilizar y entender el árbol de decisión construido, sino construirlo a partir de ejemplos. La dificultad apuntada al momento de construir el árbol puede resolverse con un conjunto de entrenamiento bien constituido.

Palma ha desatacado que la fuerte popularidad de los árboles de decisión ha contribuido a que la aplicación más utilizada sea la clasificación, no obstante, existe variantes de árboles de decisión que permiten que se pueda emplear en otras tareas como son la regresión y el agrupamiento. En la aplicación de recomendación odontológica, es posible añadir más atributos en el árbol, siempre teniendo en cuenta que se debe considerar la importancia de los nuevos atributos sobre los considerados inicialmente.

CONCLUSIONES

La discusión precedente sugiere que un algoritmo de aprendizaje es bueno si produce hipótesis que hacen un buen trabajo al predecir clasificaciones de ejemplos que no han sido observados. Una predicción es buena si resulta cierta, por lo tanto, se puede calcular la calidad de una hipótesis contrastando sus predicciones con la clasificación correcta una vez que es conocida.

Una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente llegar a una sola acción o decisión a tomar.

En general, cada camino desde la raíz a una hoja corresponde a una conjunción de test sobre los atributos, mientras que todo el árbol corresponde a una disyunción de tales conjunciones. Esto es directamente expresable como un conjunto de reglas del tipo SI...ENTONCES SINO donde las condiciones son la conjunción de test sobre los atributos y las hojas son los consecuentes.

REFERENCIAS BIBLIOGRAFICAS

- [1] AHO ALFRED. (1998) Estructuras de datos y algoritmos
Edit Addison Wesley México ISBN 968-444-345-5
- [2] BRASSARD G. (1998) Fundamentos de Algoritmia
Edit Prentice Hall Madrid ISBN 0-13-335068-1
- [3] DROZDK, ADAM. (2007), Estructura de datos y algoritmos en JAVA.
Edit Thomson México 2007 ISBN 0-534-49252-5
- [4] GRASSMANN W. (1997) Matemática Discreta y Lógica

Edit Prentice Hall Madrid ISBN 0-13-501206-6

- [5] GRIMALDI R. (1997) Matemática Discreta y Combinatoria
Edit Adisson Wesley México ISBN 0-201-65376-1
- [6] HERNANDEZ JOSE (2004) Introducción a la minería de datos
Edit Pearson Educación Madrid 2004 ISBN 978-84-205-4091-7
- [7] JOHNSONBAUGH R. (1999) Matemáticas Discretas
Edit Prentice Hall México ISBN 0-13-518242-5
- [8] LEE R.C. (2007) Introducción al diseño y análisis de algoritmos
Edit McGraw-Hill Interamericana México ISBN 978-970-10-6124-4
- [9] PALMA JOSE (2008) Inteligencia artificial: métodos, técnicas y aplicaciones
Edit Mc Garw Hill Madrid 2008 ISBN 978-84-481-5618-3
- [10] RUSSELL S. (2007) Inteligencia artificial, un enfoque moderno
Edit Pearson Educacion Madrid 2004 ISBN 978-84-205-4003-0
- [11] WEISS MARK (2000) Estructura de datos en Java
Edit Adisson Wesley España ISBN 0-201-54991-3
- [12] WEISS MARK (1995), Estructura de datos y algoritmos.
Edit Adisson Wesley 1995 Wilmington E.U.A. ISBN 0-201-62571-7
- [13] <http://es.scribd.com/doc/41336427/OPT-08-4-Tecnicas-de-ramificacion-y-poda> consultado junio 2011

NOTAS DEL AUTOR

- 1 **Algoritmos de inducción** : El esquema básico que sigue la mayoría de los algoritmos de inducción de arboles de decisión consiste en construir el árbol de la forma top-down, desde la raíz hasta las hojas
- 2 **Aprendizaje**: El termino aprendizaje es utilizado en muchas áreas como psicología, Pedagogía, Zoología, Antropología y mas recientemente en Informática. En informática la visión mas genérica define al aprendizaje como la mejora del comportamiento a partir de la experiencia[Mitchell 1977]. Una visión mas extensa define el aprendizaje como la capacidad de predecir observaciones futuras con plausibilidad o explicar observaciones pasadas.
- 3 **Arbol de porfirio**: Se da este nombre al cuadro en el cual se representa la relación de subordinación(solo lógica según unos, lógica y ontológica según otros)
- 4 **William de Ockham**: Filósofo ingles del siglo XIV, quien plantea que es preferible la hipótesis consistente con los datos que sea mas sencilla. Ya que las hipótesis que no son tan sencillas como los datos van a fallar a la hora de extraer cualquier patrón de los mismos