ARTÍCULO ORIGINAL

Árboles de decisión como herramienta en el diagnóstico médico

Decision trees as a tool in the medical diagnosis

Rocío Erandi Barrientos Martínez[1], Nicandro Cruz Ramírez[1], Héctor Gabriel Acosta Mesa[1], Ivonne Rabatte Suárez[2], María del Carmen Gogeascoechea Trejo[2], Patricia Pavón León[2], Sobeida L. Blázquez Morales[2].

Recibido: 07/09/2009 - Aceptado: 18/09/2009

RESUMEN

En este trabajo se evalúa el desempeño de tres de los algoritmos más representativos para la construcción de árboles de decisión. Los árboles de decisión son un modelo de clasificación utilizado en la inteligencia artificial, cuya principal característica es su aporte visual a la toma de decisiones. Para poner a prueba el rendimiento en el proceso de clasificación de los árboles de decisión, se utilizarán dos bases de datos que contienen datos médicos de pacientes reales. Estos datos corresponden a la sintomatología que un médico especialista considera para el diagnóstico de cáncer de seno. Una de las bases de datos contiene 692 casos recopilados de las observaciones de un solo médico y la otra, contiene 322 casos recopilados de la observación de 19 especialistas. En suma, se busca determinar la pertinencia de los árboles de decisión, es decir, si pueden ser una herramienta de apoyo para el diagnóstico médico.

Palabras clave: árboles de decisión, cáncer de mama, algoritmo, clasificación.

ABSTRACT

In this paper, we evaluate the performance of three of the most representative algorithms for constructing decision trees. Decision trees are a classification model used to in Artificial Intelligence, whose main characteristic is its contribution to visual decision making. In order to test performance of the classification process of decision trees, we use two databases, that contain medical data of real patients. These data correspond to the symptoms that a doctor takes into account for the diagnosis of breast cancer. One of the databases contains 692 cases collected from the observation of one single doctor and another contains 322 cases collected from the observation of 19 specialists. The purpose is to determine whether the decision trees can be a support tool for medical diagnosis.

Keywords: decision trees, breast cancer, algorithm, classification.

[1] Facultad de Física e Inteligencia Artificial.
[2] Instituto de Ciencias de la Salud, Universidad Veracruzana, Xalapa, Veracruz, México.

INTRODUCCIÓN

A través del tiempo se han desarrollado una gran cantidad de métodos para el análisis de datos, los cuales principalmente están basados en técnicas estadísticas. Sin embargo, a medida de que la información almacenada crece considerablemente, los métodos estadísticos tradicionales han empezado a enfrentar problemas de eficiencia y escalabilidad. Debido a que la mayor parte de esta información es histórica y procede de fuentes diversas, parece clara la inminente necesidad de buscar métodos alternativos para el análisis de este tipo de datos y a partir de ellos, poder obtener información relevante y no explícita.

En la mayoría de los casos, el análisis e interpretación de los datos se hace de forma manual, es decir, el especialista analiza y elabora un informe o hipótesis que refleja las tendencias o pautas de los mismos, para poder presentar sus conclusiones y a partir de ellas poder tomar decisiones importantes y significativas. Como se puede observar, este proceso es lento, caro y altamente subjetivo, de hecho, el análisis manual es impracticable en situaciones en las que el volumen de los datos crece exponencialmente, ya que la gran cantidad de datos sobrepasa la capacidad humana para comprenderlos sin la ayuda de una herramienta adecuada. Por lo tanto, en la mayoría de los casos las decisiones importantes se toman no a partir de los datos, sino de la intuición y experiencia de los expertos, puesto que carecen de herramientas idóneas que los apoyen[1].

Para el caso de la medicina, es posible aplicar métodos alternativos, debido a la gran cantidad de padecimientos involucrados, las sintomatologías y los pacientes. Lo ideal sería que los médicos pudieran contar con el apoyo de una herramienta que les permita analizar los datos sintomatológicos de cada uno de sus pacientes para poder determinar con base en casos anteriores, el diagnóstico más acertado así como el tratamiento óptimo a seguir, lo cual representaría un soporte y ayuda para el médico. Una herramienta alternativa para la predicción y clasificación de grandes cantidades de datos que es utilizada ampliamente en el área de la inteligencia artificial son los árboles de decisión.

Árboles de decisión

Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema. Constituyen probablemente el modelo de clasificación más utilizado y popular. El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo[2]. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver (Ver Figura 1).

Este modelo se construye a partir de la descripción narrativa de un problema, ya que provee una visión gráfica de la toma de decisión, especificando las variables que son evaluadas, las acciones que deben ser tomadas y el orden en el que la toma de decisión será efectuada. Cada vez que se ejecuta este tipo de modelo, sólo un camino será seguido dependiendo del valor actual de la variable evaluada. Los valores que pueden tomar las variables para este tipo de modelos pueden ser discretos o continuos[3].

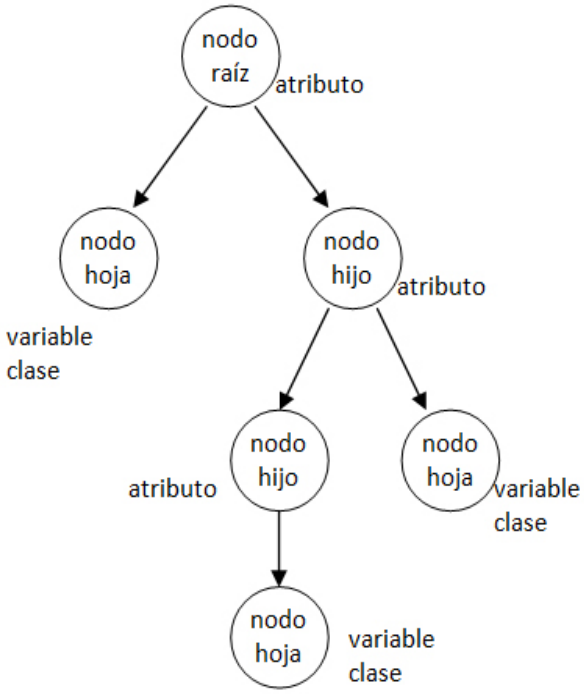

Figura 1: Estructura de un árbol de decisión.

Un algoritmo de generación de árboles de decisión consta de 2 etapas: la primera corresponde a la inducción del árbol y la segunda a la clasificación. En la primera etapa se construye el árbol de decisión a partir del conjunto de entrenamiento; comúnmente cada nodo interno del árbol se compone de un atributo de prueba y la porción del conjunto de entrenamiento presente en el nodo es dividida de acuerdo con los valores que pueda tomar ese atributo. La construcción

del árbol inicia generando su nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de entrenamiento en dos o más subconjuntos; para cada partición se genera un nuevo nodo y así sucesivamente. Cuando en un nodo se tienen objetos de más de una clase se genera un nodo interno; cuando contiene objetos de una clase solamente, se forma una hoja a la que se le asigna la etiqueta de la clase. En la segunda etapa del algoritmo cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja, a partir de la que se determina la membresía del objeto a alguna clase. El camino a seguir en el árbol lo determinan las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente en él.

OBJETIVO

Evaluar el desempeño de clasificación de datos médicos a partir de los resultados obtenidos en la aplicación de algoritmos basados en árboles de decisión, para poder determinar si esta técnica de clasificación puede ser una herramienta de soporte y ayuda eficaz en el tratamiento y diagnóstico médico.

MATERIALES Y MÉTODOS

Para llevar a cabo este trabajo de investigación utilizamos dos bases de datos [4] que contienen información acerca de tumores (malignos o benignos) para la detección de cáncer de mama. Estas bases de datos tienen las siguientes características:

- La primera está integrada por un conjunto de datos recopilados por la experiencia de un solo patólogo. Esta base de datos contiene 692 casos tomados del Departamento de Patología del Hospital "Royal Hallamshire" en Sheffield, Reino Unido del año 1992 al 1993.
- La segunda corresponde a un conjunto de datos recopilados por la experiencia de 19 patólogos diferentes, quienes cuentan de 5 a 20 años de experiencia en la detección de cáncer de mama; esta base de datos contiene 322 casos tomados del departamento de patología arriba citado, pero del año 1996 al 1997.

Las bases de datos anteriores contienen las mismas variables, las cuales corresponden a características que los patólogos toman en cuenta para poder emitir un diagnóstico sobre el cáncer de mama. En la tabla 1 se describen estas variables así como los valores que los patólogos asignaron a cada una de ellas para su descripción (interpretación y codificación); cabe mencionar que para poder llegar a un diagnóstico final, es decir para confirmar si el tumor detectado en las pacientes es maligno o no, fue necesaria una biopsia y una mamografía.

Puesto que el objetivo de este trabajo es verificar si los árboles de decisión son una herramienta para el diagnóstico

Tabla 1. Variables de entrada para las bases de datos sobre cáncer de mama.

Característica observada (variable)	Valores que puede tomar la variable	Definición
Edad	uno, dos o tres	Toma valor de "uno" si la paciente tiene menos de 50 años, "dos" si está entre 50 y 70 años y "tres" si tiene más de 70 años.
Dimensión celular	verdadero o falso	Toma valor de "verdadero" si la mayoría de las células epiteliales de la paciente se encuentran dentro de grupos adhesivos y valor de "falso" si la mayoría están dentro de grupos cohesivos.
Lumina intracitoplasmática	verdadero o falso	Toma valor de "verdadero" si en las células epiteliales de la paciente está presente esta característica y "falso" si está ausente.
Agrupaciones de células epiteliales tridimensionales	verdadero o falso	Toma valor de "verdadero" si algunos grupos de células epiteliales no son planas y valor de "falso" si todos los grupos de las células son planos.
Núcleos bipolares	verdadero o falso	Toma valor de "verdadero" si el núcleo bipolar está presente en las células y valor de "falso" si está ausente.
Macrófago espumoso	verdadero o falso	Toma valor de "verdadero" si el macrófago espumoso está presente y valor de "falso" si está ausente.
Nucleolos	verdadero o falso	Toma valor de "verdadero" si más de tres nucleolos visibles están presentes en las células de la paciente y valor de "falso" si tres o menos están presentes.
Pleomorfismo nuclear	verdadero o falso	Toma valor de "verdadero" si algunas de las células epiteliales de la paciente tienen diámetros nucleares dos veces superior al de otros núcleos de las células y valor de "falso" si no tienen dicha dimensión el diámetro de sus células.
Tamaño nuclear	verdadero o falso	Toma valor de "verdadero" si alguno de los núcleos de las células epiteliales de la paciente tienen un diámetro dos veces mayor al diámetro de los glóbulos rojos, y valor de "falso" si los núcleos de las células tiene un diámetro inferior al doble de los glóbulos rojos.
Células epiteliales necróticas	verdadero o falso	Toma valor de "verdadero" si las células epiteliales necróticas están presentes en la paciente y valor de "falso" si están ausentes.
Cambios apócrinos	verdadero o falso	Toma valor de "verdadero" si la mayoría de las células epiteliales del núcleo de la paciente muestran cambios apócrinos y valor de "falso" si no presentan dichos cambios la mayoría de sus células.
Resultados (variable clase)	maligno o benigno	Toma valor de "maligno" o "benigno" dependiendo del diagnóstico que el doctor haya detectado al tumor.

médico, partiremos del hecho de que no podemos llevar a cabo los procedimientos médicos apropiados para detectar un tumor a una paciente ni determinar si éste es benigno o maligno. Sólo contamos con la información cualitativa (valor de las variables) proporcionada por los patólogos, por lo que la evaluación de los árboles la llevaremos a cabo con los resultados que arroje la tarea de clasificación, a partir de las bases de datos descritas anteriormente. Los algoritmos de clasificación que utilizamos para evaluar a los árboles son los siguientes:

- **ID3:** Algoritmo que aprende a partir de la diferencia que existe entre los datos para analizar, esto es, un procedimiento de divide y vencerás, que maximiza la información obtenida, la cual se utiliza como una métrica para seleccionar el mejor atributo que divida los datos en clases homogéneas [5].
- **J48:** Este algoritmo construye un árbol a partir de datos. Se construye iterativamente al ir agregando nodos o ramas que minimicen la diferencia entre los datos. Este algoritmo es un descendiente del ID3 y se extiende en el sentido de su capacidad de utilizar atributos numéricos y vacíos para generar reglas del árbol. Con el propósito de clasificación de una nueva instancia, J48 prueba cada uno de los valores del atributo de acuerdo con su estructura hasta que encuentra una hoja, la cual contiene los valores de la clase para cada instancia [6].
- **Naive Bayes:** Algoritmo que genera un árbol de decisión a partir del clasificador bayesiano Naive Bayes, que es el modelo más simple de clasificación ya que asume independencia entre todos los atributos dada una clase. Por lo tanto, corresponde a un modelo de atributos independientes. En este caso, la estructura de la red es fija y sólo es necesario aprender los parámetros. El fundamento principal de este clasificador es la suposición de que todos los atributos son independientes del valor de la variable clase [7].

La elección de los algoritmos anteriores obedece a que son los más utilizados debido a su sencillez, precisión y bajo costo de cómputo en su ejecución.

METODOLOGÍA Y RESULTADOS

Para evaluar la capacidad de clasificación de los árboles de decisión se llevaron a cabo los siguientes experimentos:

- El primero consistió en dividir aleatoriamente la primera base datos en dos conjuntos. Uno de 462 casos (dos terceras partes del total) para poder entrenar el árbol de decisión a partir de estos datos. Un segundo conjunto de 230 casos (una tercera parte del total) para probar que el árbol también clasifica estos datos. A partir de éstos y los algoritmos descritos en la sección anterior se construyeron los árboles de decisión. Los resultados de dichos experimentos se muestran en la tabla 2 y en la figura 2 con uno de los árboles de decisión que se obtuvieron.

Tabla 2. Porcentajes de clasificación para la primera base de datos con árboles de decisión.

	ID3	J48	Naive Bayes Tree
Porcentaje de casos que clasificó correctamente el algoritmo.	93.04%	91.73%	**94.35%**

- El segundo experimento consistió en considerar a la segunda base de datos como un conjunto de prueba (322 casos), mientras que para el conjunto de entrenamiento se consideró el mismo conjunto de 462 casos del primer experimento. El experimento se hizo con el propósito de analizar qué conjunto de datos es más significativo, a partir del porcentaje de casos que clasifica correctamente. A partir de estos datos, se construyeron los árboles de decisión que corresponden a cada uno de los algoritmos descritos en la sección anterior. Los resultados de estos experimentos se muestran en la tabla 3 y en la figura 3 con uno de los árboles de decisión.

Tabla 3. Porcentajes de clasificación para la segunda base de datos con árboles de decisión.

	ID3	J48	Naive Bayes Tree
Porcentaje de casos que clasificó correctamente el algoritmo	82.60%	81.98%	**85.71%**

CONCLUSIONES

Los resultados obtenidos en los experimentos descritos en la sección anterior dan evidencia que es posible construir con precisión árboles de decisión a partir de datos médicos, ya que los porcentajes de clasificación, es decir el número de casos que clasificó correctamente, tienen un margen de error mínimo y es posible que pueda mejorar su eficiencia con la ayuda del experto, ajustando los datos mismos, esto es, agregando variables o cambiando sus parámetros.

También es importante mencionar que los resultados obtenidos con los datos de la primera base de datos, la cual corresponde a las observaciones obtenidas de un solo especialista, son mejores que los resultados obtenidos de los datos provenientes de las observaciones de 19 especialistas (datos de la segunda base de datos), ya que la cantidad de casos que clasificó correctamente en la primera base de datos es mayor que la cantidad en la segunda. Este resultado refleja que existe mayor discrepancia en la apreciación (lo cual se

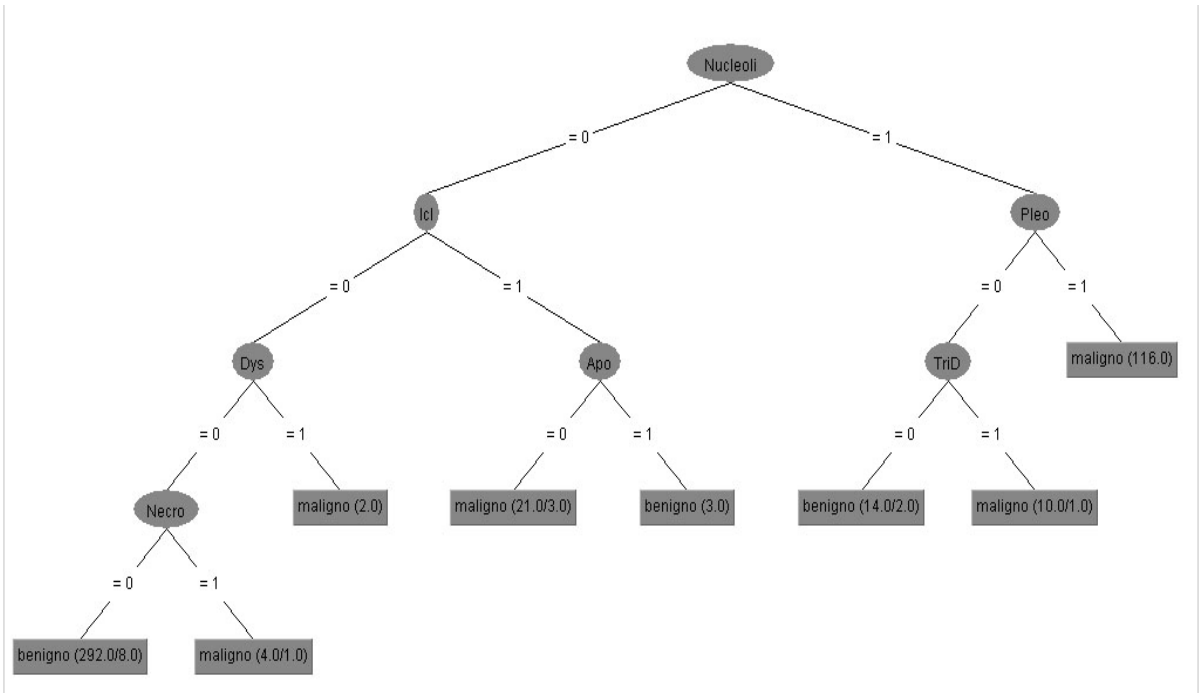

Figura 2. Árbol de decisión construido con el algoritmo ID3 para la primera base de datos.

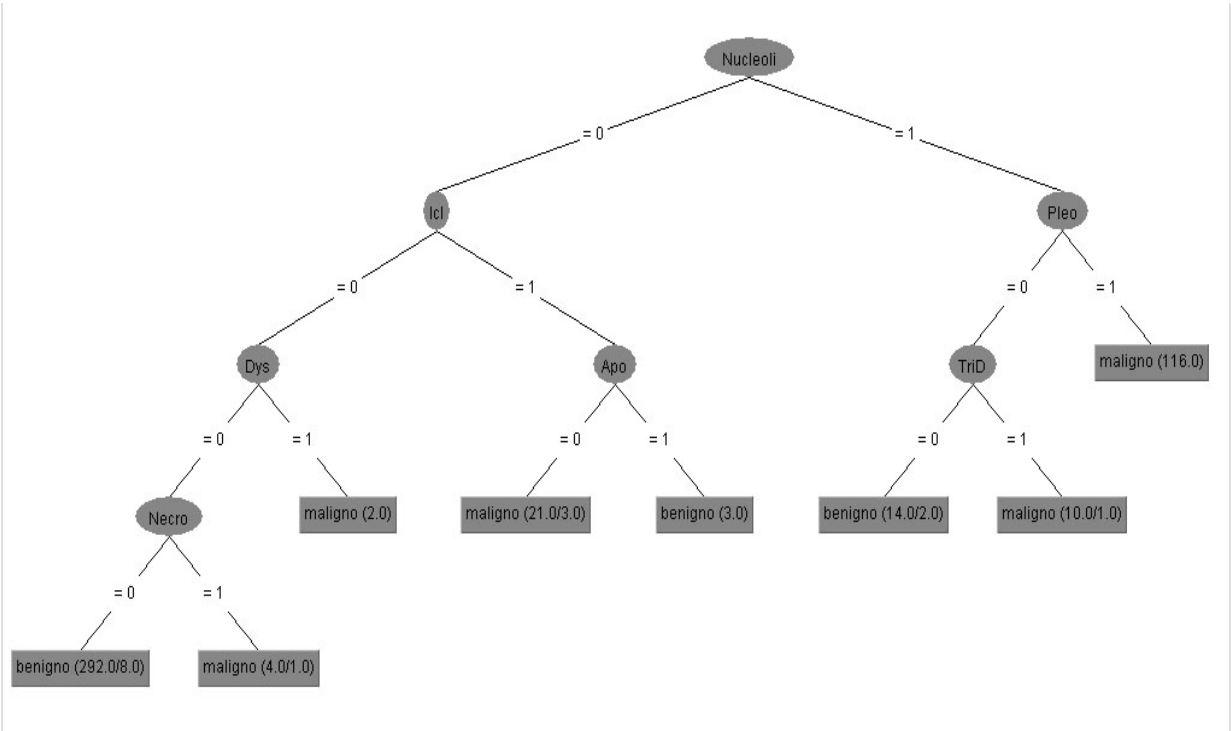

Figura 3. Árbol de decisión construido con el algoritmo ID3 para la segunda base de datos.

traduce en problemas para la clasificación) cuando intervienen más observadores, debido a que se aportan diferentes puntos de vista, incrementándose la subjetividad en los valores de las variables (lo cual corresponde a la sintomatología).

Con base en este planteamiento podemos concluir que a partir de un conjunto de datos aportados por un especialista en una disciplina es posible tener en los árboles de decisión una herramienta de apoyo y ayuda confiable para el diagnóstico médico, aun cuando es importante destacar que lo más importante es contar con un conjunto de datos consistente y confiable, ya que este tipo de herramientas están supeditadas al conocimiento del experto que aportará la información. Por ello es necesario continuar realizando pruebas en otras especialidades médicas para encontrar el conjunto óptimo para la construcción de este tipo de herramientas.

Agradecimientos

Agradecemos al Dr. Simon S. Cruz, Profesor Clínico de la Unidad Académica de Patología de la Universidad de Sheffield en Reino Unido, quien amablemente nos proporcionó las bases de datos para este trabajo.

BIBLIOGRAFÍA

1. Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Barrientos-Martínez RE. Comparison of the Performance of Seven Classifiers as Effective Decision Support Tools for the Cytodiagnosis of Breast Cancer: A Case Study. Analysis and Design of Intelligent Systems using Soft Computing Techniques. Advances in soft computing; 41: 79 - 87.
2. Russell, S. and P. Norvig, Artificial Intelligence: A Modern Approach. Second ed. Upper Saddle River (N J): Prentice Hall/ Pearson Education; 2003.
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees, Wadsworth (New York); 1994.
4. Cross SS y cols. Which Decision Support Technologies Are Appropriate for the Cytodiagnosis of Breast Cancer? Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis, A. Jain, et al., Editors. World Scientific 2000; 265-295.
5. Quinlan JR. Learning Decision Tree Classifiers. ACM Computing Surveys 1996; 28(1): 71-72.
6. Quinlan JR. Programs for Machine Learning. The Morgan Kaufmann Series in Machine Learning. San Mateo (California): Morgan Kaufmann Publishers; 1993.
7. Dunham MH. Data Mining. Introductory and Advanced Topics. Upper Saddle River (N J): Prentice Hall; 2003.