



ARTÍCULO ORIGINAL

Evaluación del Potencial de Redes Bayesianas en la Clasificación en Datos Médicos

Evaluation of the Potential of Bayesian Networks on the Classification of Medical Data

Rocío Erandi Barrientos Martínez¹, Nicandro Cruz Ramírez¹, Héctor Gabriel Acosta Mesa¹, Ivonne Rabatte Suárez², Patricia Pavón León², María del Carmen Gogearcoechea Trejo², Ma. Sobeida L. Blázquez Morales².

¹ Facultad de Física e Inteligencia Artificial

² Instituto de Ciencias de la Salud.

Universidad Veracruzana, Xalapa, Veracruz, México.

Barrientos-Martínez RE, Cruz-Ramírez N, Acosta-Mesa HG, Rabatte-Suárez I, Pavón-León P, Gogearcoechea-Trejo MC, Blázquez-Morales MSL.

Evaluación del Potencial de Redes Bayesianas en la Clasificación en Datos Médicos

Rev Med UV 2008; 8(1): 33-37.

RESUMEN

En este trabajo, se presenta una evaluación del potencial de Redes Bayesianas para la clasificación de datos médicos. Por su doble naturaleza, cualitativa y cuantitativa, éstas permiten visualizar fácilmente las relaciones probabilistas entre variables de interés así como realizar inferencias como predicción, diagnóstico y toma de decisiones. Razones por las cuales, la medicina ha recurrido a ellas para el análisis y procesamiento de datos. La evaluación que se llevó a cabo de las Redes Bayesianas consistió en el procesamiento de bases de datos relacionadas con padecimientos médicos tales como Cáncer de Mama, Tumores Cancerígenos, Diabetes y Hepatitis. Este procesamiento se realizó utilizando algoritmos cuyo proceso de clasificación está basado en Redes Bayesianas. Lo anterior, se realizó con la finalidad de que los resultados obtenidos en el proceso de clasificación, permitan determinar si las Redes Bayesianas son una herramienta eficaz y confiable en la toma de decisiones en lo que al ámbito médico se refiere y poder convertirse en un auxiliar en el diagnóstico y tratamiento médicos.

Palabras Clave: Redes Bayesianas, Bases de Datos, Medicina, Clasificadores.

ABSTRACT

In this paper, we present the evaluation of Bayesian Networks on the classification of medical data. Their qualitative and quantitative nature permits representing the probabilistic relationships among variables as well as carrying out inferences such as prediction, diagnosis and decision-making. The medical area has used them for analysis and processing of data. Here, we evaluate the performance of Bayesian Networks on medical databases related to diseases such as Breast Cancer, Tumors, Diabetes and Hepatitis. In order to carry out such a task, we tested different Bayesian Network classifiers so that we can determine whether Bayesian Networks are a powerful and reliable tool for diagnosis and decision-making in this area.

Keywords: Bayesian Networks, Databases, Medicine, Classifiers.

Recibido 06/06/2007 - Aceptado 25/06/2008

INTRODUCCIÓN

Una Red Bayesiana es un modelo gráfico que representa las relaciones de probabilidad entre variables de interés. Este modelo tiene dos componentes: una estructura o topología y un conjunto de tablas de probabilidad condicional. La estructura es una representación gráfica descrita por nodos y arcos. Los nodos corresponden a las variables del problema y los arcos son los que muestran las relaciones entre las variables. La orientación de los arcos muestra la influencia entre las variables, lo cual no necesariamente es de naturaleza causal.⁶ La estructura de la red captura las relaciones cualitativas entre las variables ($x_1, x_2, \dots, x_n$), es decir, dos nodos se conectan directamente si uno afecta o causa el otro (Ver figura 1).

Un arco indica la dirección del efecto. En la estructura de este tipo de redes, se emplean los términos de padre, hijos, antecesor y descendiente. Un nodo es padre de un nodo hijo si existe un arco del primero al segundo. También se emplean tablas locales de distribución de probabilidad condicional, las cuales cuantifican las relaciones entre los nodos conectados.⁷

El proceso de clasificación de las Redes Bayesianas está basado en el teorema de Bayes, que es un simple producto de probabilidades condicionales. Sea x y y dos eventos con $p(x) > 0$, entonces la probabilidad del evento y dado que sucedió x está dada por:

$$p(Y/X) = \frac{p(Y)p(X/Y)}{p(X)}$$

Cabe mencionar que el uso principal de este teorema es mostrar cómo la probabilidad y/x está relacionada con la de x/y . Para el método de clasificación el teorema se establece como: Suponga que $x_1, x_2, \dots, x_n$ son variables aleatorias que permiten predecir el valor de una variable c llamada Clase de acuerdo a este teorema, la probabilidad de que una observación pertenezca a la clase C es:

$$p(C = C / X_1, \dots, X_n) = \frac{p(C = C) p(X_1, \dots, X_n / C = C)}{p(X_1, \dots, X_n)}$$

El proceso de clasificación se realiza asignando la observación en aquella clase que tenga la mayor probabilidad usando la ecuación anterior. Las Redes Bayesianas pueden ayudarnos a simplificar la representación de la función de

probabilidad conjunta $p(x_1, x_2, \dots, x_n/C)$ considerando las relaciones de dependencia que existen entre las variables.⁸

El modelo se puede construir con el conocimiento del experto o a partir de datos. En el contexto de las Redes Bayesianas la probabilidad se interpreta como una métrica numérica del grado de creencia consistente en una proposición, la consistencia se establece de acuerdo a los datos disponibles. Entre las principales ventajas de las Redes Bayesianas se considera que es una representación compacta, intuitiva y robusta.

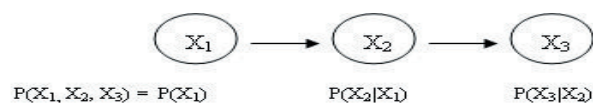


Figura 1: Ejemplo de una Red Bayesiana.

OBJETIVO

Evaluar el desempeño de clasificación de datos médicos a partir de algoritmos basados en Redes Bayesianas, y así poder determinar si son una herramienta poderosa y eficaz para los médicos en la predicción y el diagnóstico médico.

MATERIALES Y MÉTODOS

Las bases de datos que se utilizaron para este artículo, fueron tomadas del repositorio de datos de la Universidad de California,⁶ cuyo contenido se encuentra relacionado con diferentes padecimientos médicos, los cuales se describen en las tablas siguientes. (Ver tablas de la I a IV)

Característica Observada	Definición
Edad	Edad del paciente en base a los intervalos: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
Menopausia	Valores de si, no o premeno dependiendo si la paciente está o no en la menopausia ó premenopausia.
Tamaño	Tamaño en milímetros del tumor, considerando los intervalos: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
Maligno	Valores de si o no dependiendo si el tumor detectado a la paciente es o no maligno.
Seno	Valores de izquierdo o derecho dependiendo en cual seno se encuentre el tumor.
Cuadrante	Valores de izquierdo-arriba, izquierdo-abajo, derecho-arriba, derecho-abajo o central, dependiendo de la ubicación exacta del tumor en el cuadrante de su seno.
Radiación	Valores de si o no dependiendo si la paciente se someterá a un tratamiento de radiación.
Variable Clase	Valores de si o no dependiendo si la enfermedad es recurrente o no.

Tabla I. Variables de Entrada de la base de datos "Breast Cancer". La información contenida pertenece a 286 pacientes.

Característica Observada	Definición
Pre-Glucosa	Cantidad de Azúcar en la sangre del paciente en ayunas.
Presión	Valor correspondiente a la presión arterial del paciente,
Piel	Valores que describen la apariencia de la piel del paciente.
Insulina	Cantidad de insulina requerida para el paciente.
Masa Corporal	Cantidad de masa corporal en el cuerpo del paciente
Inactividad	Valores de si o no dependiendo si el paciente tiene actividad o no.
Edad	Edad del paciente en años.
Variable Clase	Valores de si o no dependiendo si el paciente tiene diabetes.

Tabla II. Variables de Entrada de la base de datos “Diabetes”. La información contenida pertenece a 768 pacientes.

Característica Observada	Definición
Edad	Edad del paciente en años.
Sexo	Valores de masculino y femenino según sea el caso
Esteroides	Valores de si o no dependiendo si el paciente ha consumido esteroides.
Antigripales	Valores de si o no dependiendo si el paciente ha tomado o no algún medicamento antigripal.
Fatiga	Valores de si o no dependiendo si el paciente sufre de fatiga.
Anorexia	Valores de si o no dependiendo si el paciente padece de anorexia
Hígado Grande	Valores de si o no dependiendo si el paciente tiene el hígado grande.
Hígado Firme	Valores de si o no dependiendo si el paciente tiene el hígado firme.
Varices	Valores de si o no dependiendo si el paciente tiene várices.
Bilirrubina	Cantidad de bilirrubina que el paciente tiene en la sangre.
Albumina	Cantidad de albumina que el paciente tiene en la sangre.
Fosfato	Cantidad de fosfato que el paciente tiene en la sangre.
Antecedentes	Valores de si o no si el paciente tiene el antecedente de hepatitis.
Variable Clase	Valores de Vivo o Muerto

Tabla III. Variables de Entrada de la base de datos “Hepatitis”. La información contenida pertenece a 155 pacientes.

Característica Observada	Definición
Edad	Edad del paciente considerando los intervalos: <30, 30-59, >=60.
Sexo	Valores de masculino y femenino según sea el caso
Tipo	Tipo histológico del tumor, con valores de epidérmico, adenoide, anaplastico
Grado	Grado de definición del tumor, con valores probables de bueno, moderado y pobre.
Hueso Medular	Valores de si o no dependiendo de si el tumor se encuentra en la médula espinal del paciente.
Pulmón	Valores de si o no dependiendo de si el tumor se encuentra en el pulmón.
Peritoneo	Valores de si o no dependiendo del problema del paciente.
Hígado	Valores de si o no dependiendo si el tumor se encuentra en el hígado.
Cerebro	Valores de si o no dependiendo si el tumor se encuentra en el cerebro
Piel	Valores de si o no dependiendo de la lesión en la piel del paciente
Cuello	Valores de si o no dependiendo de si el tumor se encuentra alojado en el cuello.
Axila	Valores de si o no dependiendo si el tumor se encuentra en la axila.
Abdominal	Valores de si o no dependiendo si el tumor se encuentra en un área abdominal.
Otro Órgano	Valores de estómago, ovarios, próstata, riñón, páncreas, tiroides, matriz, vagina, seno. Dependiendo donde se aloje el tumor.

Tabla IV. Variables de Entrada de la base de datos “Primary Tumor”. La información contenida pertenece a 339 pacientes.

El proceso de evaluación de las redes bayesianas a partir de los datos médicos descritos anteriormente, se llevó a cabo a través de la ejecución de algoritmos de clasificación, cuya precisión y rapidez de procesamiento son los más utilizados en este tipo de problemas. Los algoritmos son los siguientes:

- 1. Naive Bayes:** Es uno de los algoritmos de clasificación más efectivo, sus principales cualidades son su simplicidad y precisión, y aunque su estructura siempre es fija (la variable clase apuntando a cada nodo) se ha mostrado que tiene una alta precisión de clasificación y un error mínimo. En términos simples, Naive Bayes aprende a partir de una muestra de datos de entrenamiento, la probabilidad condicional de cada atributo dada la clase. Entonces, una vez que un nuevo caso se da, usa el teorema de Bayes para calcular la

probabilidad condicional de cada nodo seleccionando el valor de la clase con la probabilidad mayor.⁴

2. **Tan:** Algoritmo conocido también como Red Bayesiana Aumentada a Árbol. El cual consiste en construir un árbol de dependencias entre las variables que se van a predecir y que a su vez son hijas de la variable clase. Por lo que la probabilidad de dichas variables se calculará aplicando el teorema de Bayes con base en la probabilidad de la variable clase.⁸
3. **Hill-Climber:** Algoritmo de clasificación que inicia con una red generada de manera aleatoria. Para cada nodo o variable, el algoritmo agrega o borra relaciones de manera aleatoria, calculando a partir de la probabilidad conjunta de la variable clase la probabilidad de cada nodo que forman la red. El algoritmo elige la red óptima, es decir aquella que tiene la mejor calidad, eliminando aquellas que no alcanzan su nivel.⁷
4. **K2:** Algoritmo de clasificación que inicia con la red más simple posible, es decir, una red sin arcos, y supone que los nodos están ordenados. Para cada variable del problema, el algoritmo añade a su conjunto de padres el nodo con menor probabilidad que conduce a un máximo incremento de la calidad correspondiente a la medida de calidad elegida en el proceso de clasificación. Este proceso se repite hasta que, o bien no se incrementa la calidad o se ha llegado a una red completa.⁹

METODOLOGÍA Y RESULTADOS

Considerando los años de estudio que necesita un médico para convertirse en un especialista cuyo tiempo aproximado es de 4 a 5 años, lo cual nos da una idea del proceso de entrenamiento tan complejo por el que deben pasar los médicos, aunado a esto se tiene la evidencia de que en el proceso de interpretación existe un importante ingrediente de subjetividad. Razón por la cual se desea evaluar si las Redes Bayesianas pueden ser una herramienta de soporte que ayude a disminuir este problema o al menos que ayude a identificar las posibles causas de esta subjetividad.

La figura II muestra la Red Bayesiana generada para la base de datos de Breast Cancer, así como las probabilidades condicionales de algunos nodos o variables de la red. Lo anterior para darnos una idea mas clara de la estructura y contenido de una red de este tipo.

La evaluación del proceso de clasificación de las Redes Bayesianas se llevó a cabo a través de la ejecución de los algoritmos descritos anteriormente, para cada una de las bases de datos. Una vez que se obtuvieron los resultados, se obtuvo el promedio de dichas ejecuciones, para poder contar con un balance más exacto del proceso de clasificación que se llevó a cabo. En la tabla V se muestran los resultados obtenidos.

Cabe mencionar que los resultados obtenidos dependen tanto del proceso de clasificación de las Redes Bayesianas, como de la experiencia del médico, quien es el que lleva a cabo el proceso de captura e interpretación de los padecimientos presentados.

En términos generales podemos observar que el proceso de clasificación de las Redes Bayesianas es satisfactorio ya que los porcentajes de clasificación obtenidos para cada una de las variables están por encima de 70% de exactitud, a excepción de la base de datos relacionada con los tumores cancerígenos.

En la figura III podemos observar como los porcentajes de clasificación para cada uno de los algoritmos basados en Redes Bayesianas se comportan de manera similar y que la diferencia más significativa se presenta en los porcentajes de clasificación de una base de datos a otra.

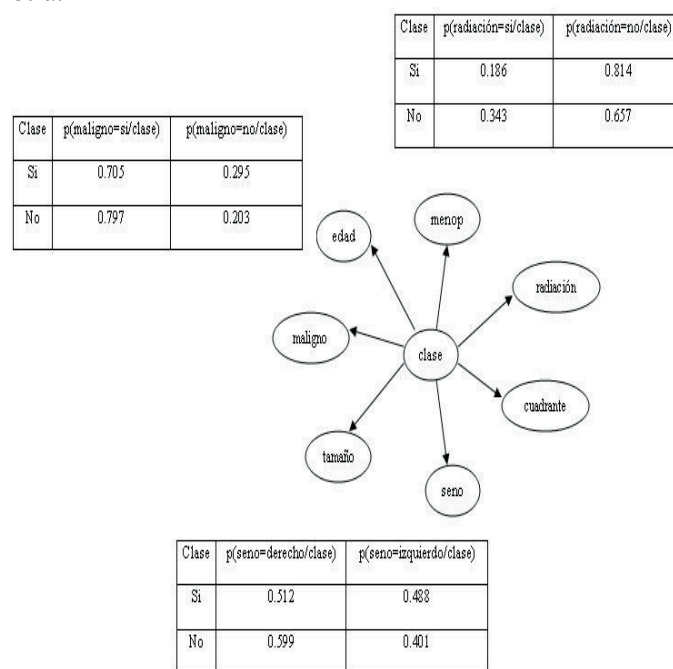


Figura 2. Ejemplo de una Red Bayesiana para el problema de cáncer de seno.

Base de Datos	Naive Bayes	K2	Tan	Hill Climber	Porcentaje Promedio
Breast Cancer	73.5936	73.3867	74.0345	73.4557	73.617625
Diabetes	75.6777	75.0775	75.7567	75.0776	75.397375
Hepatitis	81.5256	88.8462	88.8462	88.8462	87.01605
Tumor	43.9429	37.9394	39.3565	37.9394	39.79455

Tabla V. Resultados obtenidos al evaluar bases de datos con información médica, con algoritmos de clasificación basados en Redes Bayesianas.

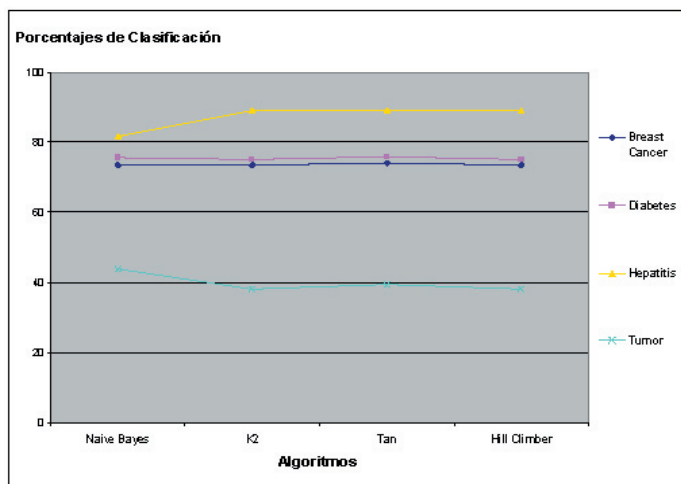


Figura 3. Figura que muestra el comportamiento de los porcentajes de clasificación que se obtuvieron al analizar las bases de datos con los algoritmos de clasificación utilizando Redes Bayesianas.

CONCLUSIONES

Como podemos observar, los resultados anteriores son constantes, puesto que la diferencia entre los porcentajes es mínima. En base a lo anterior, podemos concluir que el comportamiento de los algoritmos de clasificación basados en redes bayesianas presentan resultados en términos generales satisfactorios ya que el error obtenido en la clasificación es mínimo, a excepción de la base de datos *Tumor Primary*, cuyo porcentaje es muy bajo. Lo anterior pudo ser resultado de dos situaciones, que la cantidad de casos o variables utilizadas no sean suficientes y su contenido es carente de precisión, o que la estructura de la Red Bayesiana utilizada en los algoritmos que se ejecutaron no permiten la generación de dichos datos. Para poder mejorar este porcentaje, el experto en la materia deberá evaluar si es posible agregar información y/o especificar la información proporcionada a la base de datos.

Es importante mencionar que aunque desde el punto de vista matemático los resultados obtenidos en esta evaluación son satisfactorios por lo dicho anteriormente,

desde el punto de vista médico pueden provocar incertidumbre los resultados obtenidos, por lo delicado de su trabajo, ya que el mínimo error en este tipo de situaciones podría ser fatal. Por lo que es importante trabajar en lo que a la interpretación de las variables se refiere, y así tener la confianza de que se puede lograr con éxito el diagnóstico médico y la toma de decisiones.

BIBLIOGRAFÍA

1. Heckerman D., Geiger D. and Chickering, D. *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*. Machine Learning 20, 1995; 197-243.
2. Jiménez-Andrade JL, *BayesN: Un Algoritmo para Aprender Redes Bayesianas Clasificadoras a partir de datos*. Tesis de Maestría en Inteligencia Artificial. Xalapa, Ver.: Facultad de Física e Inteligencia Artificial, U.V., 2003; 1-10.
3. Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*. 2da. Edición Elsevier 2005.
4. Cruz-Ramírez N, Acosta-Mesa HG, Barrientos-Martínez RE and Nava-Fernández LA. *How Good are Bayesian Information Criterion and the Minimum Description Length Principle for Model Selection? A Bayesian Networks Analysis*. Advances in Artificial Intelligence. Vol. 4293, 2006; 494-504.
5. Pagina Web del repositorio de datos de la Universidad de California, <http://kdd.ics.uci.edu/>.
6. Gutierrez-Fragoso K, *Análisis del comportamiento de MDL en el contexto del aprendizaje de la estructura de Redes Bayesianas a partir de datos*. Tesis de Maestría en Inteligencia Artificial. Xalapa, Ver.: Facultad de Física e Inteligencia Artificial, U.V., 2007.
7. Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Nava-Fernández LA and Barrientos-Martínez RE. *Diagnosis of Breast Cancer using Bayesian Networks: A case study*. Computers in Biology and Medicine. Vol. 37, 2007; 1553-1564.
8. Pérez A, Larrañaga P and Inza I. *Modelos gráficos probabilísticos para la clasificación supervisada empleando la estimación basada en kernels Gaussianos esféricos*. III Taller Nacional de Minería de Datos y Aprendizaje. 2005; 125-134.