

# OPTIMIZACIÓN DE REDES BAYESIANAS BASADO EN TÉCNICAS DE APRENDIZAJE POR INDUCCIÓN

Felgaer, P.<sup>1</sup>, Britos, P.<sup>2,3</sup>, Sicre, J.<sup>2</sup>, Servetto, A.<sup>4,3</sup>, García-Martínez, R.<sup>2,1</sup> y Perichinsky, G.<sup>4,3</sup>

1.- Laboratorio de Sistemas Inteligentes  
Facultad de Ingeniería.  
Universidad de Buenos Aires.  
Paseo Colón 850 4to Piso. Ala Sur. (1063) Capital Federal

2.- Centro de Ingeniería del Software e Ingeniería del  
Conocimiento (CAPIS)  
Instituto Tecnológico de Buenos Aires  
Av. Madero 399. (1106) Capital Federal.

3.- Programa de Doctorado en Ciencias Informáticas  
Facultad de Informática  
Universidad Nacional de La Plata.  
Buenos Aires

4.- Laboratorio de Sistemas Operativos y Bases de Datos.  
Facultad de Ingeniería.  
Universidad de Buenos Aires  
Paseo Colón 850 4to Piso. Ala Sur. (1063) Capital Federal

**Resumen:** Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística. Son utilizadas para proveer una forma compacta de representar el conocimiento y métodos flexibles de razonamiento. El obtener una red bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico. En este trabajo, se define un método de aprendizaje que optimiza las redes bayesianas aplicadas a clasificación, mediante la utilización de un método de aprendizaje híbrido que combina las ventajas de las técnicas de inducción de los árboles de decisión (TDIDT-C4.5) con las de las redes bayesianas

**Palabras Clave:** Redes Bayesianas. Aprendizaje por Inducción. Clasificación.

**Pertinencia:** Congreso en General

## 1. Estado de la cuestión

La habilidad de aprender es considerada como una característica central de los “sistemas inteligentes”, y es por esto que se ha invertido esfuerzo y dedicación en la investigación y el desarrollo de esta área. El aprendizaje puede ser definido como “cualquier proceso a través del cual un sistema mejora su eficiencia” [Simon, 1983]. El desarrollo de los sistemas basados en conocimientos motivó la investigación en el área del aprendizaje con el fin de automatizar el proceso de adquisición de conocimientos el cual se considera uno de los problemas principales en la construcción de sistemas basados en conocimientos.

Se denomina Minería de Datos al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos [Piatetski-Shapiro *et al.*, 1991; Chen *et al.*, 1996; Mannila, 1997].

El término Minería de Datos Inteligente [Evangelos & Han, 1996; Michalski *et al.*, 1998] refiere específicamente a la aplicación de métodos de aprendizaje automático [Michalski *et al.*, 1983; Holsheimer & Siebes, 1991], para descubrir y enumerar patrones presentes en los datos.

Se desarrollaron un gran número de métodos de análisis de datos basados en la estadística [Michalski *et al.*, 1982]. En la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos. Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que el análisis de datos tradicional supone que las hipótesis ya están construidas y validadas contra los datos, mientras que la minería de datos supone que los patrones e hipótesis son automáticamente extraídos de los datos [Hernández Orallo, 2000].

Las tareas de la minería de datos se pueden clasificar en dos categorías: minería de datos descriptiva y minería de datos predictiva [Piatetski-Shapiro *et al.*, 1996; Han, 1999]. Algunas de las técnicas más comunes de minería de datos son los árboles de decisión (TDIDT), las reglas de producción y las redes neuronales.

Por otro lado, un aspecto importante del aprendizaje por inducción, es el de obtener un modelo que represente el dominio de conocimiento y que sea accesible para el usuario. En particular, resulta importante obtener la información de dependencia entre las variables involucradas en el fenómeno, en los sistemas donde se desea predecir el comportamiento de algunas variables desconocidas basados en ciertas variables conocidas. Una representación del conocimiento que es capaz de capturar esta información sobre las dependencias entre las variables son las redes bayesianas.

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres. La variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables y sus dependencias condicionales dada otra(s) variable(s). Dichas dependencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades).

El obtener una red bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico [Pearl, 1988]. La primera de ellas, consiste en obtener la estructura de la red bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas. La segunda etapa, tiene como finalidad obtener las probabilidades a priori y condicionales requeridas a partir de una estructura dada.

Las redes bayesianas [Pearl, 1988] son utilizadas en diversas áreas de aplicación como por ejemplo el diagnóstico médico [Beinlinch *et al.*, 1989]. Las mismas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento - basados en las teorías probabilísticas - capaces de predecir el valor de variables no observadas y explicar las observadas. Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos [Heckerman *et al.*, 1995; Díaz & Corchado, 1999], evitan el sobre-ajuste de los datos y pueden manejar bases de datos incompletas [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996].

## 2. Descripción del Problema

Las redes bayesianas están diseñadas para hallar las relaciones de dependencia e independencia entre todas las variables que conforman el dominio de estudio. Esto permite realizar predicciones sobre el comportamiento de cualquiera de las variables desconocidas a partir de los valores de las otras variables conocidas. Esto presupone que cualquier variable de la base de datos puede comportarse como incógnita o como evidencia según el caso.

Las redes bayesianas pueden realizar la tarea de clasificación como un caso particular de la tarea de predicción mencionada anteriormente. Se caracteriza por tener una sola de las variables de la base de datos (clasificador) que se desea predecir, mientras que todas las otras son los datos propios del caso que se desea clasificar. Pueden existir una gran cantidad de variables en la base de datos, algunas de las cuales estarán directamente relacionados con la variable clasificadora que se quiere predecir pero también pueden existir otras variables que no son influyentes sobre dicha clase.

En este trabajo, se define un método de aprendizaje que ayuda en la pre-selección de variables del dominio, optimizando la configuración de la red bayesiana en problemas de clasificación.

## 3. Solución Propuesta

Para solucionar el problema de las redes bayesianas aplicadas a la tarea de clasificación, lo que propone en este trabajo es utilizar un método de aprendizaje híbrido que combine las ventajas de las técnicas de inducción de los árboles de decisión (TDIDT – C4.5) con las de las redes bayesianas.

Para ello, lo que se propone es integrar al proceso de aprendizaje estructural y paramétrico de las redes bayesianas, un proceso previo de preselección de variables. En el mismo, se elige a partir de todas las variables del dominio, un subconjunto de ellas con la finalidad de generar la red bayesiana para la tarea particular de clasificación y de esta forma, optimizar la performance y mejorar la capacidad predictiva de la red.

El método para aprendizaje estructural de redes bayesianas se basa en el algoritmo desarrollado por Chow y Liu (69) para aproximar una distribución de probabilidad por un producto de probabilidades de segundo orden, lo que corresponde a un árbol. La probabilidad conjunta de  $n$  variables se puede representar como:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{j(i)})$$

donde  $X_{j(i)}$  es la causa o padre de  $X_i$ .

Se plantea el problema como de optimización y lo que se desea es obtener la estructura en forma de árbol que más se aproxime a la distribución “real”. Para ello se utiliza una medida de la diferencia de información entre la distribución real ( $P$ ) y la aproximada ( $P^*$ ):

$$I(P, P^*) = \sum_x P(X) \log(P(X) / P^*(X))$$

Entonces, el objetivo es minimizar  $I$ . Para ello se define una diferencia en función de la información mutua entre pares de variables, que se define como:

$$I(X_i, X_j) = \sum_x P(X_i, X_j) \log(P(X_i, X_j) / P(X_i)P(X_j))$$

Chow [1968] demuestra que la diferencia de información es una función del negativo de la suma de las informaciones mutuas (pesos) de todos los pares de variables que constituyen el árbol. Por lo que encontrar el árbol más próximo equivale a encontrar el árbol con mayor peso. Basado en lo anterior, el algoritmo para determinar la red bayesiana óptima a partir de datos es el siguiente:

1. Calcular la información mutua entre todos los pares de variables ( $n(n-1)/2$ ).
2. Ordenar las informaciones mutuas de mayor a menor.
3. Seleccionar la rama de mayor valor como árbol inicial.
4. Agregar la siguiente rama mientras no forme ciclo, si es así, desechar.
5. Repetir (4) hasta que se cubran todas las variables ( $n-1$  ramas).

Rebane y Pearl [1989] extendieron el algoritmo de Chow y Liu para poliárboles. En el caso de un poliárbol, la probabilidad conjunta es:

$$P(X) = \prod_{i=1}^n P(X_i | X_{j1(i)}, X_{j2(i)}, \dots, X_{jm(i)})$$

donde  $\{X_{j1(i)}, X_{j2(i)}, \dots, X_{jm(i)}\}$  es el conjunto de padres de la variable  $X_i$ .

#### 4. Verificación Experimental

A continuación se procede a realizar una comparación experimental entre las redes bayesianas puras y las redes bayesianas preprocesadas con algoritmos de inducción C4.5. Para realizar esta comparación, se utilizaron bases de datos obtenidas del *Irving Repository of Machine Learning* de la Universidad de California.

A continuación se resumen las características principales de las bases de datos utilizadas

|            | Variables | Variables C4.5 | Clases | Casos Totales | Casos Control | Casos Contraste |
|------------|-----------|----------------|--------|---------------|---------------|-----------------|
| Votaciones | 16        | 6              | 2      | 435           | 300           | 135             |
| Hongos     | 22        | 7              | 2      | 8124          | 5416          | 2708            |

La metodología utilizada para llevar a cabo los experimentos con cada una de las bases de datos evaluadas, se detalla a continuación.

- Dividir la base de datos en dos. Una de control o entrenamiento (aproximadamente 2/3 de la base total) y otra de contraste o validación (con los datos restantes)
- Procesar la base de datos de control mediante el algoritmo de inducción C4.5 para obtener el subconjunto de variables que conformarán la red bayesiana C4.5
- Repetir para el 10%, 20%, ..., 100% de los datos de la base de control
  - Repetir 30 veces

- Tomar al azar el X% de la base de datos de control según el porcentaje que corresponda a la iteración
- Mediante ese subconjunto de casos de la base de control, realizar el aprendizaje estructural y paramétrico de las redes bayesianas Completa y C4.5
- Evaluar el poder predictivo de ambas redes utilizando la base de datos de contraste (este se mide como el porcentaje de casos clasificados correctamente sobre el total de casos evaluados). Cuando el porcentaje de certeza sobre la clasificación de un caso no sea del 100%, se considerará la clase con mayor probabilidad
  - Calcular el poder predictivo promedio (a partir de las 30 iteraciones)
- Graficar el poder predictivo de ambas redes (Completa y C4.5) en función de los casos de entrenamiento

Gráfico del poder predictivo de las redes bayesianas completas y C4.5 en función de la cantidad de casos de aprendizaje para el dominio “Votaciones”.

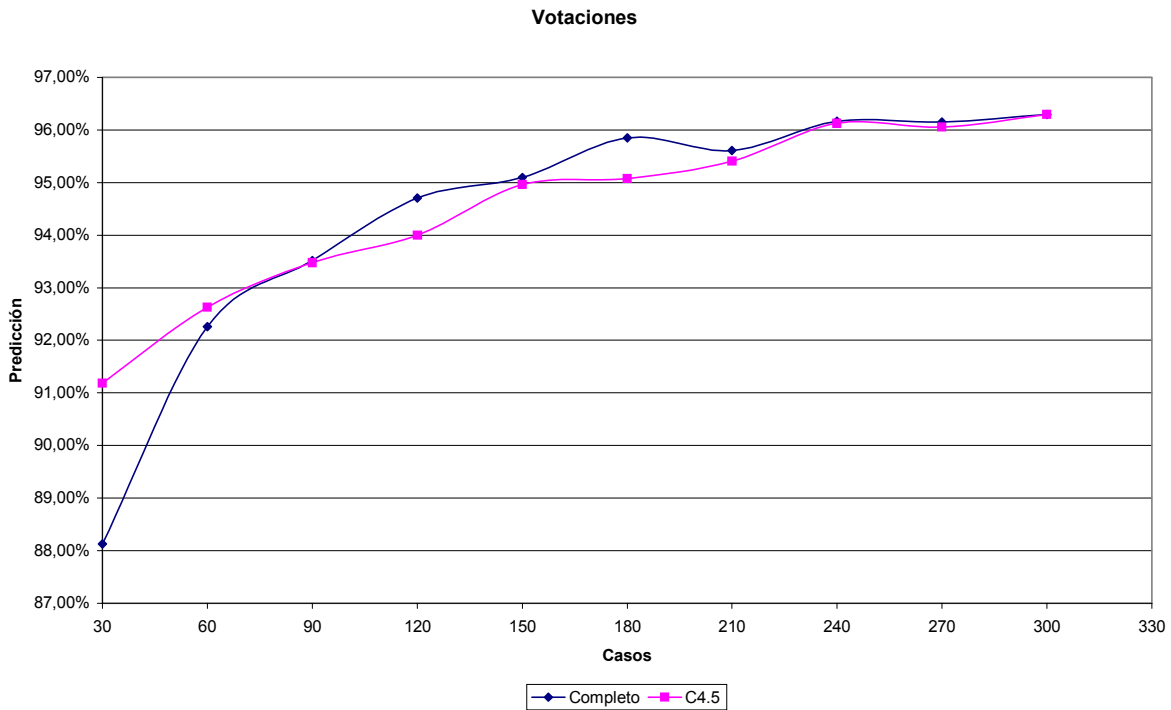
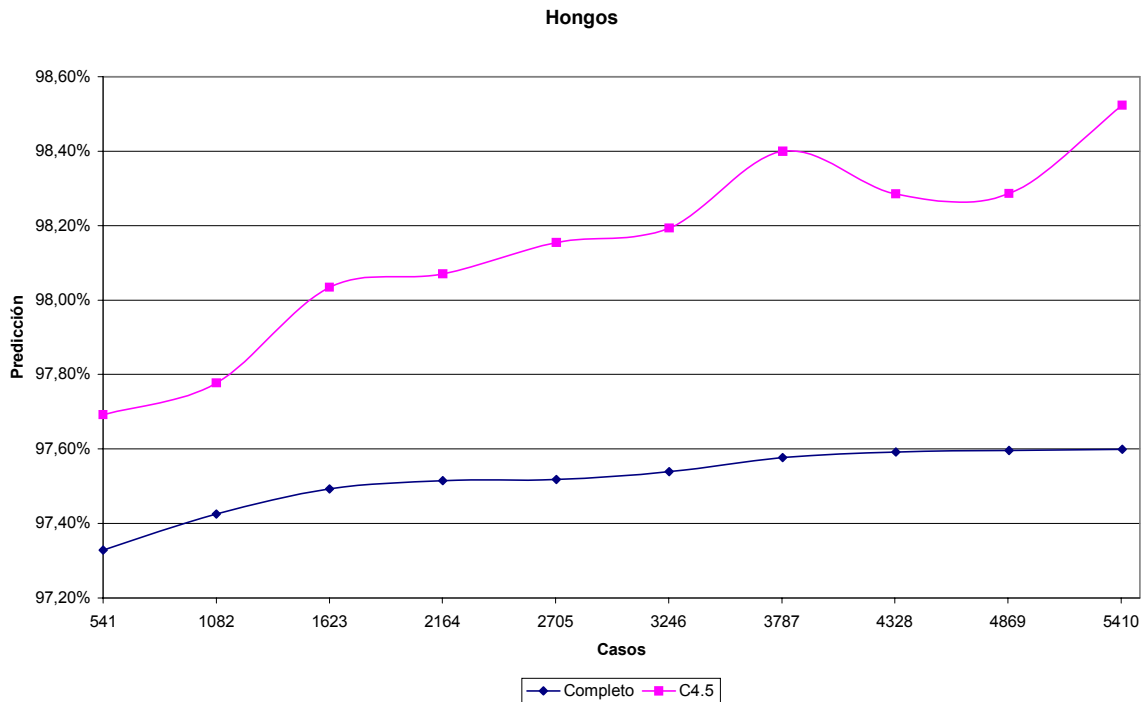


Gráfico del poder predictivo de las redes bayesianas completas y C4.5 en función de la cantidad de casos de aprendizaje para el dominio “Hongos”



## 5. Conclusiones

Como se puede observar, las curvas que representan el poder predictivo en función de la cantidad de casos de entrenamiento son crecientes. Este fenómeno se da independientemente del dominio de datos utilizado y del método evaluado (red bayesiana Completa o C4.5). Esto se traduce en la conclusión de que cuanto mayor sea la base de datos que se disponga para la tarea de entrenamiento, mejor será el poder predictivo de la red bayesiana obtenida ya que la misma, será una representación más fidedigna del dominio real que intenta representar el modelo.

Por otro lado, del análisis de los resultados obtenidos en la experimentación, podemos concluir que el método híbrido de aprendizaje utilizado en bases de datos que abarquen un orden pequeño de casos, no genera ventajas sobre el método de aprendizaje tradicional (como se puede ver en el caso de “Votaciones”). Sin embargo, en bases de datos de volúmenes más grandes, como por ejemplo la de “Hongos”, la diferencia en el poder predictivo es clara y en favor del método propuesto en este trabajo.

## 6. Referencias

- Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F. (1989). *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks*. In proceedings of the 2<sup>nd</sup> European Conference on Artificial Intelligence in Medicine.
- Chen, M., Han, J., Yu, P. (1996). *Data mining: An overview from database perspective*. IEEE Transactions on Knowledge and Data Eng.

- Diaz, F., Corchado, J.M. (1999). *Rough sets bases learning for bayesian networks*. International workshop on objective bayesian methodology, Valencia, Spain.
- Evangelos, S., Han, J. (1996). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, EE.UU.
- Han, J. (1999). *Data Mining*. Urban and Dasgupta (eds.), Encyclopedia of Distributed Computing, Kluwer Academic Publishers.
- Heckerman, D. (1995). *A tutorial on learning bayesian networks*. Technical report MSR-TR-95-06, Microsoft research, Redmond, WA.
- Heckerman, D., Chickering, M., Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data*. Machine learning 20: 197-243
- Heckerman, D., Chickering, M. (1996). *Efficient approximation for the marginal likelihood of incomplete data given a bayesian network*. Technical report MSR-TR-96-08, Microsoft Research, Microsoft Corporation.
- Hernández Orallo, J. (2000). *Extracción automática de conocimiento de bases de datos e ingeniería de software*. Programación declarativa e ingeniería de la programación.
- Holsheimer, M., Siebes, A. (1991). *Data Mining: The Search for Knowledge in Databases*. Report CS-R9406, ISSN 0169-118X, Amersterdam, The Netherlands.
- Mannila, H. (1997). *Methods and problems in data mining*. In Proc. of International Conference on Database Theory, Delphi, Greece.
- Michalski, R.S., Baskin, A.B., Spackman, K.A. (1982). *A Logic-Based Approach to Conceptual Database Analysis*. Sixth Annual Symposium on Computer Applications on Medical Care, George Washington University, Medical Center, Washington, DC, EE.UU.
- Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (1983). *Machine learning I: An AI Approach*. Morgan Kaufmann, Los Altos, CA.
- Michalski, R.S., Bratko, I., Kubat, M. (1998). *Machine Learning and Data Mining, Methods and Applications*. John Wiley & Sons Ltd, West Sussex, England.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA.
- Piatetski-Shapiro, G., Frawley, W.J., Matheus, C.J. (1991). *Knowledge discovery in databases: an overview*. AAAI-MIT Press, Menlo Park, California.
- Piatetsky-Shapiro, G., Fayyad, U.M., Smyth, P. (1996). *From data mining to knowledge discovery*. AAAI Press/MIT Press, CA.
- Ramoni, M., Sebastiani, P. (1996). *Learning bayesian networks from incomplete databases*. Technical report KMI-TR-43, Knowledge Media Institute, The Open University.