LAS REDES BAYESIANAS COMO SISTEMAS DE SOPORTE A LA DECISIÓN EN LA EMPRESA

Mar Abad Grau
María Visitación Hurtado Torres
Rosana Montes Soldado
Miguel J. Hornos Barranco

Lenguajes y Sistemas Informáticos
Universidad de Granada

RESUMEN

En la empresa actual la cantidad de datos manipulados aumenta a velocidades vertiginosas. El recurso de la información constituye una valiosa fuente de diferenciación frente a las empresas del sector. Para gestionarla de forma adecuada es necesario poder procesar e interpretar grandes cantidades de datos, de forma que sea posible extraer el conocimiento necesario para una adecuada toma de decisiones. Frente a los clásicos sistemas expertos basados en reglas, actualmente se están utilizando otros basados en modelos gráficos probabilísticos (redes bayesianas, diagramas de influencia, etc.), los cuales permiten representar el grado de incertidumbre de las relaciones de dependencia entre las variables. Asimismo el experto puede decidir para su construcción, el nivel de certidumbre que el sistema debe asignar al conocimiento no extraído de los datos sino aportado directamente por él. Presentamos en este trabajo un ejemplo de red bayesiana para la concesión de créditos bancarios capaz de aprender a partir de las fuentes de datos sobre préstamos bancarios.

Palabras clave: Sistemas de Soporte a la Decisión, Sistemas expertos, Diagramas de influencia, redes bayesianas

1. INTRODUCCIÓN

La extracción de conocimiento a partir de una fuente de datos, constituye una valiosa tarea que hace de la información uno de los recursos más importantes de la empresa en el mundo globalizado en el que nos hallamos inmersos.

Son principalmente dos las razones que nos llevan a destacar el uso de herramientas automáticas de extracción de conocimiento. Por un lado, las cantidades cada vez más ingentes de información hacen larga y tediosa la tarea de extracción de conocimiento, *inducción* principalmente, por parte del experto. Por otro lado, la tarea necesaria para manejar tanto las fuentes de datos como los distintos modelos de conocimiento sobre los que aplicar los datos, conlleva una alta complicación para la memoria limitada propia del ser humano, si la comparamos con la de las máquinas.

En la sección 2 se realizará una breve introducción a los sistemas expertos y se definirán los llamados sistemas expertos probabilísticos. En la sección 3 se comentarán las redes bayesianas como caso concreto de sistema experto probabilístico. En la sección 4 se estudiará cómo una red bayesiana puede ser usada para la clasificación, de cara a la toma de decisiones del experto. Asimismo se explicarán diversos algoritmos para el aprendizaje de clasificadores basados en redes bayesianas. En la sección 5 se compararán distintos sistemas expertos entrenados para la clasificación según los algoritmos anteriormente expuestos. Para ello se utilizará como fuente de datos una muestra con 1000 casos sobre concesiones de préstamos en una entidad financiera. Se comparará la tasa de error con los casos nuevos una vez que el sistema experto ha sido entrenado con un subconjunto de casos (*conjunto de entrenamiento*) de los que se conoce el resultado (concesión/denegación del préstamo). Por último, en la sección 6 se expondrán las conclusiones y las líneas de trabajo futuro.

SISTEMAS EXPERTOS PROBABILÍSTICOS

Existe una herramienta capaz de aprender de forma automática, a partir de los datos y también de la opinión del experto, especialmente en su primera etapa de

funcionamiento, donde todavía el volumen de datos no es muy elevado y la opinión del experto es más valiosa. Esta herramienta es lo que se llama un *sistema experto* [5, 6].

En un sistema experto existen cuatro componentes básicos:

- El subsistema de *aprendizaje* o de *adquisición del conocimiento*, encargado de obtener tanto el conocimiento abstracto (reglas, espacios probabilísticos, etc.) como el concreto (casos). El proceso inductivo de adquisición de conocimiento abstracto a partir de la información concreta se denomina *aprendizaje automático*. En los inicios de un sistema experto, el aprendizaje suele ser realizado teniendo en cuenta la experiencia del experto.
- La *base de conocimiento*, encargada de almacenar el conocimiento abstracto.
- La *memoria de trabajo*, que almacena el conocimiento concreto o información.
- El *motor de inferencia*, que aplica el conocimiento abstracto al conocimiento concreto para sacar conclusiones. El procedimiento de inferencia por el cual a partir de conocimiento abstracto se sacan conclusiones concretas se denomina *deducción*.

Los sistemas expertos clásicos utilizan como modelo para la representación del conocimiento base de conocimiento un conjunto de reglas. Su principal inconveniente lo constituye la dificultad en la propagación de incertidumbre. Este problema puede ser grave a efectos prácticos, ya que las conclusiones a las que se llega pueden no ser correctas.

Los *sistemas expertos probabilísticos* utilizan como base de conocimiento, la estructura del espacio probabilístico y como motor de inferencia, probabilidades condicionales. Esto permite el manejo de incertidumbre. Dentro del espacio probabilístico, el modelo o base de conocimiento más utilizado es la llamada *red bayesiana*, la cual pretende que la propagación de probabilidades sea exacta, rápida y no cause problemas de excesivo número de parámetros [5]. Para ello se representan las relaciones de (in)dependencia condicionada mediante un grafo dirigido acíclico. Las redes bayesianas se están aplicando en los últimos años en la construcción de sistemas expertos. Así, algunos ejemplos en el campo de la

medicina y la biología pueden verse en [15, 18] y en el campo de las ciencias empresariales en [3, 12, 16]. Algunas variantes de los sistemas expertos probabilísticos son aquellos que permiten el tratamiento de la incertidumbre mediante otras teorías de incertidumbre diferentes a la de la probabilidad, por ejemplo, la teoría de la evidencia de Dempster-Shafer [23], los intervalos de probabilidad [14], etc.

3. REDES BAYESIANAS

Una red bayesiana está compuesta por:

1. Un grafo dirigido acíclico (GDA) donde cada nodo representa una variable aleatoria y los arcos representan dependencias probabilísticas entre variables. A esta parte de la red se la denomina *estructura* o *modelo*.
2. Una distribución de probabilidades condicionadas de la forma $P(x \mid \pi_x)$ para cada nodo x dado su conjunto de padres π_x. Estos son los llamados parámetros de la red bayesiana.

En una red bayesiana se considera que cada nodo es independiente de todos los nodos no descendientes dados sus padres. Así, a partir del producto de probabilidades condicionadas se puede obtener la distribución conjunta de probabilidades:

$$P(x_1, \ldots, x_i, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i \mid \pi_{xi})$$

3.1. APRENDIZAJE AUTOMÁTICO DE LA RED BAYESIANA

Si la base de conocimiento de un sistema experto la constituye una red bayesiana, ,sta debe cambiar conforme aumentan los datos concretos o casos. Así, en sus primeros estados, la red bayesiana puede ser construida según la opinión de los expertos o bien según la información de la que se parta o teniendo en cuenta ambas cosas. Conforme se añade información (conocimiento concreto) a la misma, se va modificando tanto su estructura como los parámetros mediante un proceso de aprendizaje.

En este trabajo consideraremos la construcción de la base de conocimiento

teniendo en cuenta exclusivamente los datos concretos, con el objeto de poder comparar los métodos de aprendizaje de la red bayesiana con otros métodos, ya sean estadísticos o pertenecientes también al propio campo de la inteligencia artificial.

4. CLASIFICADORES BAYESIANOS

Un clasificador es una función que asigna un valor de un atributo discreto, llamado *clase*, a instancias o ejemplos descritos mediante un conjunto de atributos, que pueden ser tanto continuos como discretos. Un sistema experto puede ser utilizado como clasificador. Así por ejemplo, un sistema experto en un hospital determinar que para un conjunto de síntomas presentados por un individuo, no es probable que exista cáncer. O bien, el sistema experto para ayuda a la decisión en una entidad financiera aconsejar no otorgar un préstamo a un cliente porque la probabilidad de impago sea muy elevada, a partir de una serie de atributos, fundamentalmente financieros, del mismo.

En el caso de que la base de conocimientos sea una red bayesiana, la función de clasificación estará definida a partir de probabilidades condicionadas. Otros modelos que suelen ser utilizados son los árboles de decisión [19], las redes neuronales [16, 20] o las más recientes máquinas de soporte vectorial [98].

Una de las redes bayesianas más eficientes en la clasificación es el llamado *clasificador simple* (del inglés *Naïve Bayes classifier*). La estructura de esta red bayesiana se basa en una fuerte restricción: todos los atributos que describen los casos son independientes entre sí dado el valor de la clase (Fig. 1).

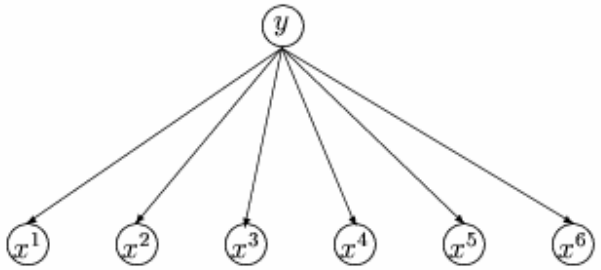

Figura 1: Gráfico correspondiente al clasificador simple bayesiano con 6 atributos de entrada.

Han sido propuestos otros modelos más sofisticados. Entre ellos, las *redes bayesianas simples aumentadas* (del inglés *Augmented Naïve Bayesian networks* (*AN*)), permiten arcos entre los atributos de entrada, de manera que se reduce la fuerte restricción propia de las redes simples.

Para decidir la estructura concreta de una red AN han sido propuestos diversos algoritmos en las publicaciones sobre Inteligencia Artificial de la última década [9, 21, 10]. Entre ellos cabe destacar el algoritmo de aprendizaje de *red bayesiana Simple Aumentada en árbol* (del inglés *Tree Augmented Naïve Bayesian network (TAN))* [10] y el de *red bayesiana Simple Aumentada Estructurada* [2] (del inglés *Structured Augmented Naïve Bayesian network (SAN)*).
El algoritmo TAN construye una red bayesiana con una estructura *TAN* (Fig. 2), es decir, una estructura en la que la variable clase no tiene padres y los atributos de entrada tienen como padres la clase y como máximo otro atributo más de entrada.

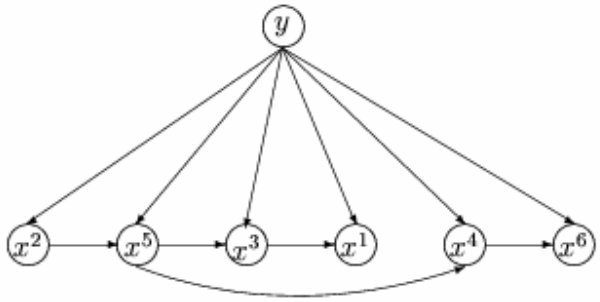

Figura 2: Un ejemplo de estructura TAN.

Para seleccionar el atributo padre *z* de un atributo *x* se utiliza como medida la llamada *información mutua condicionada* de *x* y *z* dada la clase *y*: Se trata de una medida del grado de independencia condicional de *x* y *z* dada la clase *y*: *I (x, z | y).*

Por otra parte, el algoritmo SAN es aun más flexible que TAN, en el sentido de que permite la construcción de estructuras AN menos restrictivas. Estas estructuras, llamadas también *SAN* (Fig. 3) se caracterizan porque la clase no tiene padres y los atributos de entrada pueden tener como padres además de la clase, cualquier número de atributos de entrada, siempre que no haya ciclos dirigidos, pues la estructura de una red bayesiana es siempre un GDA.

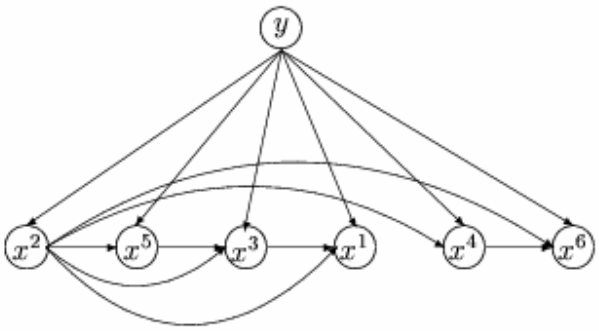

Figura 3: Un ejemplo de estructura SAN.

Sin embargo, cuanto más compleja es una estructura, es decir, cuantos más arcos existan en una estructura SAN, mayor es el riesgo de *sobreajuste*, es decir, mayor es el riesgo de que la estructura aprendida clasifique bien los casos usados para el aprendizaje pero tenga una baja eficiencia para casos nuevos. Así, su capacidad de generalización será baja y por tanto el aprendizaje no se puede considerar aceptable [20, 10, 1].

Para evitar el problema del sobreajuste en estructuras complejas el algoritmo de aprendizaje SAN utiliza un principio inductivo que favorece la creación de estructuras simples. Así, si el número de casos del que se dispone es pequeño, la estructura elegida ser más sencilla que si el número de casos es mayor. Dicho principio es el llamado Minimización del Riesgo Estructural (del inglés *Structural Risk Minimization*). Así, este principio define un equilibrio entre la calidad de un modelo dado un conjunto de datos y la complejidad del mismo.

Todos los algoritmos de aprendizaje que generan una estructura AN pertenecen al llamado *paradigma de la muestra* [7]. Los algoritmos pertenecientes a este

paradigma adolecen de los siguientes inconvenientes:

1. No son robustos frente a atributos superfluos.

2. Tienen una escasa capacidad de generalización cuando las muestras son pequeñas o el número de atributos de cada ejemplo muy elevado.

En [13] se define un algoritmo de aprendizaje inspirado en redes bayesianas. Dicho algoritmo, llamado *Transductivo Bayesiano (TB)* pertenece al *paradigma de diagnóstico* [7]. Este paradigma es más robusto frente a atributos superfluos y/o muestras pequeñas que el paradigma de la muestra. En la (Fig. 4) se muestra un ejemplo de estructura de red bayesiana creada por este algoritmo.

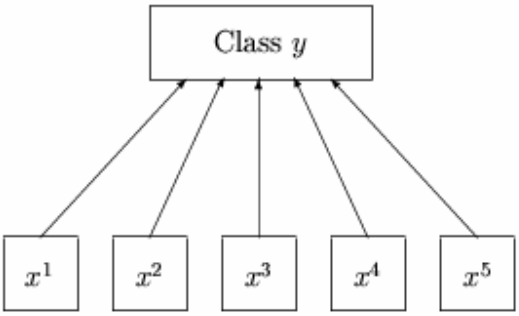

Figura 4: Estructura de la red bayesiana usada por el algoritmo TB.

Con este algoritmo sólo se calcula una distribución de probabilidad condicional. La distribución conjunta de probabilidad es proporcional a ella:

$$P(x_1,\dots,x_i,\dots,x_n) \propto P(K \mid x_1,\dots,x_i,\dots,x_n).$$

El problema que presenta la estimación de probabilidades condicionales con un elevado número de padres (tantos como atributos de entrada) es el alto riesgo de

sobreajuste dada la escasez de datos disponibles para la estimación. Por esta razón, el algoritmo BT no estima todas las probabilidades condicionadas. En su lugar, estima sólo los valores de interés. Este planteamiento del problema nos lleva al uso de un nuevo tipo de inferencia llamada *inferencia transductiva*} [25].

De forma parecida al algoritmo BT, los algoritmos *basados en instancias* [1] como el del *vecino más cercano* — del inglés *nearest neighbour (1nn)* — también estiman los valores de la clase para un conjunto de configuraciones de los atributos de entrada en las que se tiene interés.

Podemos considerar el algoritmo 1nn e incluso el BT más parecido a un sistema experto basado en reglas, que a un sistema experto probabilístico, dado que la decisión se toma según la "cercanía" del caso nuevo a los casos ya conocidos.

RESULTADOS EMPÍRICOS

Para comparar el comportamiento de estos algoritmos, hemos construido un sistema experto para cada uno de ellos. Así tendremos 6 sistemas expertos, que utilizarán como subsistema de aprendizaje los algoritmos Naïve Bayes, TAN, SAN, BT, 1nn y C4.5 [19]. Los cuatro primeros tienen como base de conocimiento una red bayesiana, el quinto una función del vecino más cercano y el último un árbol de decisión. Como antes hemos visto, dada la cantidad de parámetros a estimar con la red bayesiana de la estructura del algoritmo BT, dicho algoritmo utiliza una función de proximidad o cercanía y debe ser considerado como un tipo mixto entre los basados en redes bayesianas y los basados en casos o instancias. El algoritmo C4.5 se ha añadido dados sus buenos resultados en otros problemas. Se trata de un sistema basado en árboles o reglas de decisión, a diferencia de las redes bayesianas, que dan lugar a sistemas probabilísticos. En todos los casos consideraremos que en el aprendizaje no interviene la opinión del experto, es decir, se realizará un aprendizaje totalmente automático a partir de un conjunto de casos o ejemplos.

En todos se usará el mismo conjunto de casos, referentes a la concesión de préstamos en una entidad financiera alemana. Esta muestra ha sido obtenida del almacén de datos de la Universidad de California en Irvine [17]. Para cada caso se presentan 20 atributos de entrada. Algunos son: importe del préstamo, periodo, historial financiero del individuo, propósito del crédito, saldo de sus cuentas, contrato de trabajo, estado civil, edad, etc. La clase presenta sólo dos valores: concesión o denegación del préstamo.

El número de casos es de 1000. Como algunas de las variables son continuas, se ha utilizado como método de discretización el propuesto en [93]. Este método de discretización supervisada constituye el estado del arte en la actualidad [8, 19, 2].

Para medir la bondad de un algoritmo se ha utilizado el procedimiento de *dejar uno fuera* (del inglés *leave-one-out*) [20]. Así, cada sistema experto se entrena 1000 veces con 999 casos y se considera el resultado medio de cada caso que queda fuera en cada entrenamiento.

En la (Tabla 1) se muestran los resultado medios de fiabilidad para los 6 algoritmos de aprendizaje distintos: NB, TAN, SAN, BT, 1nn y C4.5. Como puede verse, los mejores resultados se obtienen con los algoritmos basados en redes bayesianas. Los algoritmos C4.5 y 1nn ofrecen peores resultados de fiabilidad.

Tabla 1: Fiabilidad media en la prueba alcanzada por diversos algoritmos de aprendizaje con la muestra sobre créditos bancarios.

Algoritmo de aprendizaje	Fiabilidad en la prueba
NB	76,6467
TAN	76,0479
SAN	75,4491
BT	73,9521
C4.5	71,2575
1nn	71,8563

CONCLUSIONES Y TRABAJOS FUTUROS

Como conclusión cabe destacar cómo para el ejemplo considerado de concesión de préstamos, los algoritmos basados en probabilidades aplicados para el aprendizaje de la base de conocimiento en un sistema experto se comportan mejor que aquellos basados en reglas. Así, en este caso parece que la fiabilidad de un conjunto de reglas para determinar si se concede un préstamo a un individuo es menor que si se consideran probabilidades. Así, el tratamiento de incertidumbre mediante probabilidades, como lo hacen los algoritmos basados en redes bayesianas NB, TAN y SAN obtiene mejores resultados.

También podemos observar que el algoritmo BT, a caballo entre las redes bayesianas y los algoritmos basados en casos, se comporta peor que cualquiera de los bayesianos puros.

Como trabajo futuro, nos proponemos estudiar las condiciones que se deben dar en una muestra para que sea más aconsejable el uso de sistemas expertos probabilísticos ya que, a pesar de que los resultados aquí obtenidos han sido mejores, existen otras muestras en las que los sistemas basados en reglas (en concreto los que usan C4.5) ofrecen buenos resultados [19, 13, 2]. En otros, son más aconsejables los basados en instancias [1]. En definitiva, se trata de conocer el *sesgo* [23, 26] de una muestra para aplicar el mejor algoritmo de aprendizaje. Para ello nos proponemos el desarrollo de una herramienta automática para la identificación de dicho sesgo.

BIBLIOGRAFÍA

[1] Albert, M; Aha D (1991). “Analyses of instances-based learning algorithms”, *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, AAAI Press.

[2] Abad, M. (2001). “Aplicación del principio inductivo de MEVR en la construcción de clasificadores”, *tesis doctoral*. Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, España.

[3] Ahan, J. H.; Ezawa, K. Z. “Decision support for real-time telemarketing operations through Bayesian network learning”, *Decision Support Systems*, 21, 17-27.

[4] Aha, D; Kibler, D.; Albert, M. K. (1991). “Instance-based learning algorithm”, *Machine Learning,* 6, 37-66.

[5] Castillo, E; Álvarez E. (1997). “Sistemas Expertos. Aprendizaje e incertidumbre”, Paraninfo, Madrid.

[6] Castillo, E; Gutiérrez, J. M.; Hadi, A. S. (1997) “Expert systems ans probabilistic networks models”, Springer Verlag.

[7] Dawid, A. P. (1986). “Probabilistic forecasting”. In Kotz, S; Johnson N. L.; Read, C. B. “*Encyclopedia of Statistical Sciences*”, 7, 210-218, New York.

[8] Dogherty, J; Kohavi, R; Sahami, M. (1995). “Supervised and unsupervised discretization of continuous features”, *Proceedings of the Twentyth International conference on Machine Learning*, San Francisco, CA.

[9] Ezawa K. J.; Schuermann T. (1995). “Fraud/uncollectable debt detection using a Bayesian network based learning system: A rare binary outcome with mixed data structures”, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence,* 157-166.

[10] Friedman, N.; Geiger, D; Goldszmidt, M. (1997). “Bayesian network classifier”, *Machine Learning*, 29, 131-163, 1997.

[11] Fayyad U. M.; Irani, K. B. (1993). “Multi-interval discretization of continuous-valued attributes for classification learning”, *Proceedings of the thirteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA.

[12] Garbolino, P; Taroni, F. (2002). “Evaluation of scientific evidence using bayesian networks”, *Forensic Science International* 125, 149-155.

[13] Hernández, L. ; Abad, M. (2001). “Clasificadores bayesianos robustos”. *Actas de la Novena Conferencia de la Asociación Española para la Inteligencia Artificial*, Gijón, España, 2001.

[14] Huete, J. F. (1995). “Aprendizaje de redes de creencia mediante la detección de independencias: Modelos no probabilísticos”. *Tesis Doctoral, Departamento de Ciencias de la Computación e Inteligencia Artificial,* Universidad de Granada, Granada, España.

[15] Imoto, S; Goto, T; Miyan, S. (2002). “Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression”, *Proceedings of the Pacific Symposium on Biocomputing,* 7, 175-186.

[16] Leigh, W; Purvis, R; Ragusa, J. (2002). “Forecasting the nyse composite index with technical analysis, pattern recognition, neural network, and genetic algorithm: a case study in romantic decision support”, *Decision Support Systems*, 32, 361-377.

[17] Murphy, P. M. And Aha, D. W. (1992) “UCI repository of machine learning databases”, University of California, Irvine, CA.

[18] Peér, D; Regev, A; Elidan, D; Friedman, N. (2001). “Inferring subnetworks to analyze expression data”, *Bioinformatics,* 17, 215-224.

[19] Quinlan, J. R. (1996). “Improved use of continuous attributes in C4.5”, *Journal of Artificial Intelligence Research*, 4, 77-90.

[20] Ripley, J. R. (1997). “Pattern Recognition and neural networks”, Cambridge University Press.

[21] Sahami, M. (1996). “Learning limited dependence Bayesian classifiers”, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining”, 335, 338, Menlo Park, CA.*

[22] Schaffer, C. (1994). “A conservation law for generalization performance”, *Proceedings of the eleventh international conference on Machine Learning*, Morgan Kaufmann.

[23] Shafer, G. (1976). “A mathematical Theory of Evidence”, Princeton University Press, Princeton, N. J.

[24] Vapnik, V. N. (1995). “The nature of Statistical Learning Theory”, Springer, New York.

[25] Vapnik, V. N. (1998) “Statistical Learning Theory”, J. Wiley, New York.

[26] Wilson D. R.; Martínez, T. R. (1997). “Bias and the probability of generalization”, *Proceedings of the International Conference on Intelligent Information Systems (IIS)*, 108-114.