Revista Electrónica de Metodología Aplicada
2012, Vol. 17 nº 1, pp. 1-17

Cómo Construir y Validar Redes Bayesianas con Netica

Jorge López Puga

Universidad de Almería

RESUMEN

Las redes bayesianas son herramientas de modelado estadístico destinadas a representar un conjunto de incertidumbres relacionadas. Su estructura gráfica y su fundamento probabilístico las hace apropiadas para modelar sistemas multivariados orientados a la clasificación, el diagnóstico y la toma de decisiones. En este trabajo se describen los pasos a seguir para estimar y validar redes bayesianas utilizando el software Netica. En concreto, se describe cómo construir la estructura gráfica del modelo, cómo estimar sus parámetros, cómo usar el modelo para evaluar el impacto de evidencias sobre las variables que contiene y cómo evaluar su bondad de ajuste. Todos los pasos se describen intentando hacerlos comprensibles para los investigadores en ciencias del comportamiento y la salud.

Palabras clave: Redes bayesianas, construcción, estimación, validación, Netica.

ABSTRACT

A Bayesian net is a kind of statistic modelling tool designed to represent a set of related uncertainties. It is based on a graphical structure and a set of conditional probability distributions which makes it is useful to model multivariate systems oriented to classification, diagnostic and making decisions tasks. In this work, the steps to build and validate Bayesian networks using Netica application are described. Specifically, graphical structure development, parameters estimation methods and some goodness of fit tests are described. Every step has been described trying to make them understandable to researchers in health and behaviour sciences.

Keywords: Bayesian nets, construction, estimation, validation, Netica.

Contacto:
Jorge López Puga
E-mail: jpuga@ual.es

1.- Introducción

Las redes bayesianas se han popularizado recientemente en el campo de la psicología debido a su utilidad para modelar procesos cognitivos como el aprendizaje y el razonamiento causal (p. e., Gopnik et al., 2004; Gopnik y Schulz, 2004; Glymour, 2001, 2003; Holyoak y Cheng, 2011). Sin embargo, no se ha prestado demasiada atención a la potencialidad que presenta este tipo de herramientas para ser utilizadas como técnicas de análisis de datos (López y García, 2011a; López, García, De la Fuente y De la Fuente, 2007). En nuestro contexto más cercano, las redes bayesianas se han utilizado para analizar datos en el campo de estudio de las actitudes emprendedoras y las actitudes hacia el medio ambiente (López y García, 2007; López y García, 2011b, López y García, en prensa; López, 2009; López, García, Cano, Gea y De la Fuente, 2009).

Una de las ventajas más importantes que tienen las redes bayesianas es que pueden representar de manera simultánea la dimensión cualitativa y la dimensión cuantitativa de un problema (p. e., Aguilera, Fernández, Fernández, Rumí, y Salmerón, en prensa; Edwards, 1998; Heckerman, 1995). Como señalan Heckerman, Mamdani, y Wellman (1995) esta ventaja ha sido propiciada por el aumento en la potencia de cómputo de los ordenadores personales y por el desarrollo de algoritmos de propagación de probabilidades que optimizan los recursos computacionales haciendo uso del teorema de Bayes. Por otro lado, dada su naturaleza bayesiana, estas herramientas pueden gestionar la presencia de casos perdidos en las muestras de manera eficiente (p. e., Nadkarni y Shenoy, 2004; Jansen et al., 2003). Otro aspecto a destacar de estas técnicas es que no suponen ninguna distribución o supuestos de partida subyacentes en los datos lo que las hace útiles en muchas situaciones aplicadas (Ruíz, García y Pérez, 2005). Lee, Abbott, y Johantgen (2005) también han destacado que las redes bayesianas tienen la ventaja de identificar efectos de interacción y modelar relaciones no lineales entre variables de manera que facilita la interpretación de los modelos. En cuanto al tema de la inferencia, ha sido destacado que las redes bayesianas permiten realizar inferencias bidireccionales; esto es, desde las causas a los efectos y desde los efectos a las causas. Adicionalmente, estas herramientas permiten realizar inferencias abductivas; o sea, encontrar la mejor explicación para un fenómeno a partir de un conjunto de evidencias (Gámez, 1998; Huete, 1998).

Existen varios programas informáticos que permiten crear y utilizar redes bayesianas. Algunos de ellos han sido desarrollados específicamente para generar este tipo de modelos (como Netica [Norsys Software Corp.], Elvira, Ergo [Noetic Systems Inc.] o Hugin [Hugin Expert A/S]) mientras que en otros casos han sido generados al amparo de programas estadísticos generales como son el caso de TETRAD, Neural Conection (SPSS Inc.), Bayes Net Toolbox-BNT (Matlab) o el paquete “deal” para R. Para más información sobre programas destinados a la estimación de redes bayesianas se pueden consultar los trabajos de Cowell, Dawid, Lauritzen, Spiegelhalter (1999) y Korb y Nicholson (2004). En este artículo se va a utilizar la versión 4.16 de Netica para Windows (2000/XP/Vista/7) que se puede descargar desde la página web de Norsys Software Corp. (http://www.norsys.com/download.html). El archivo Netica_Win.exe que podemos conseguir desde la citada página es un autoejecutable comprimido con WinZip que desencadena automáticamente el proceso de descompresión al hacer doble clic sobre él. El autoejecutable creará, por defecto, una carpeta llamada Netica(número de versión) en la unidad C: de nuestro equipo donde podremos encontrar todos los archivos necesarios para hacer funcionar el

programa. Para arrancar Netica tendremos que hacer doble clic en el archivo Netica.exe que hay en la carpeta creada anteriormente.

2.- Estimación.

El primer paso que hay que dar para construir una red bayesiana pasa por especificar su estructura gráfica (Cowel et al., 1999). En este sentido, podríamos decir que las redes bayesianas siguen un proceso de construcción parecido al que hay que seguir cuando se generan modelos de ecuaciones estructurales (Batista y Coenders, 2000). El hecho de que la estimación estructural sea un "pre-requisito", en comparación con la estimación paramétrica, para generar un modelo de red bayesiana ha hecho cuestionar, por ejemplo, parte de la investigación sobre juicios y/o aprendizaje causal publicados en la última mitad de siglo (Lagnado, Waldmann, Hagmayer y Sloman, 2007). Existen dos procedimientos genéricos para crear redes bayesianas (Mani, McDermmott, y Valtorta, 1997): uno basado en rutinas automáticas, donde se ponen en funcionamiento cierto número de algoritmos que son capaces de identificar la estructura gráfica subyacente en un conjunto de datos; y otro centrado en el juicio de expertos, donde se utiliza el conocimiento que un grupo de expertos tiene sobre un dominio particular para generar el modelo estadístico. En este trabajo no vamos a tratar ningún procedimiento automático de generación de redes bayesianas propiamente dicho. Más bien, se describirá la forma de crear estructuras y de estimar parámetros manualmente y de manera rápida a partir de una base de datos. Se recomienda a los lectores interesados en estos procedimientos automáticos a profundizar en el estudio de las referencias que se citan más abajo.

2.1. - Estimación estructural

Existen numerosos procedimientos automáticos (p. e., algoritmos PC y K2) destinados a generar grafos dirigidos acíclicos destinados a convertirse en redes bayesianas (p. e., Cooper y Herskovits, 1992; Cowell et al.,1999; Glymour, 2001; Gopnik et al., 2004; Herskovits y Dagher, 1997; Scheines, Spirtes, Glymour, Meek, y Richardson, 2005; Spirtes, Glymour, y Scheines, 2000). Sin embargo, Netica no incorpora ninguno de estos procedimientos automáticos de estimación estructural. Lo único que podemos hacer con Netica, como se describirá más abajo, es generar automáticamente un conjunto de nodos o variables con sus respectivos niveles o estados para, posteriormente, especificar manualmente la estructura gráfica. Así, partiendo del juicio de expertos en la materia a trabajar, o tomando como base estudios previos, podemos generar una estructura gráfica a partir de una base de datos. En este sentido, el trabajo de Nadkarni y Shenoy (2004) describe un procedimiento relativamente sencillo para construir estructuras causales bayesianas a partir del juicio de expertos. Consideremos el siguiente caso hipotético para construir nuestra red bayesiana.

> *Asumamos que trabajamos en una clínica especializada en el diagnóstico y tratamiento de procesos gripales y que solemos trabajar, de manera genérica, con dos tipos de enfermedades. La gripe A (o virus H1N1) es una enfermedad poco corriente (un 22,727% de la población la contrae cada año) mientras que la gripe común es más frecuente (supongamos que el resto de los casos que no son gripe A son gripe común). El dolor de cabeza y los problemas respiratorios están asociados con ambas enfermedades. El dolor de cabeza está presente en el 88,89% de los casos de la gripe común mientras que este síntoma está únicamente presente en el 66,67% de los casos que son diagnosticados como*

gripe A. Adicionalmente, los problemas respiratorios no están presentes en la mayoría de los casos de la gripe común (94,44%) mientras que están presentes en la mayoría de los casos en que se ha contraído el virus de la gripe A (en el 83,33% de los casos). ¿Cuál es la probabilidad de que una persona que tenga dolor de cabeza haya contraído la gripe común? ¿Cuál es la probabilidad de que una persona haya contraído la gripe común si tiene dolor de cabeza y sufre problemas respiratorios?

Para modelar este problema se podría generar una red bayesiana divergente (ver Figura 1), también llamado modelo de causa común, donde hubiese una variable que representase el tipo de enfermedad con dos posibles estados (gripe común y gripe A) y dos variables que representasen a los síntomas dolor de cabeza y problemas respiratorios (ambas variables con dos posibles estados: Sí y No).

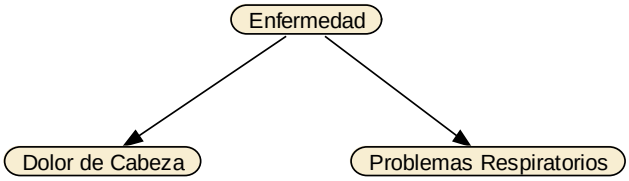

Figura 1. Estructura gráfica de la red *gripe*.

Para generar esta estructura de red bayesiana con Netica tendríamos que proceder del siguiente modo.

Generamos una nueva red utilizando el comando "File → New → Network" del menú principal o, alternativamente, presionando simultáneamente las teclas "Control" y "N" (algunos de los comandos más usuales se encuentra en forma de iconos bajo el menú principal pero, dado que el uso de estos iconos es más sencillo e intuitivo, aquí sólo se describirán los pasos haciendo referencia al menú principal y a los comandos abreviados de combinación de teclas). Hecho esto, seleccionamos la opción "Modify → Add → Nature Node" o, alternativamente, presionamos la tecla "F9". Seguidamente, hacemos clic con el botón izquierdo del ratón en el lugar donde deseemos crear nuestra variable. La variable creada, por defecto, tiene como nombre "A" y un estado o nivel llamado "state0". Para modificar las propiedades del nodo hacemos doble clic sobre él con el botón izquierdo del ratón. En el cuadro de diálogo que nos aparece (Figura 2) podemos poner el nombre del nodo ("Name") donde no se admiten ciertos tipos de caracteres como las tildes o los símbolos de interrogación y tampoco espacios. Se sugiere utilizar el guión bajo para separar diferentes palabras cuando nombremos una variable. En la opción "Title" sí que se pueden añadir más variedad de caracteres. La diferencia entre uno y otro campo es que el primero es utilizado por el programa para ejecutar rutinas mientras que el segundo se utilizará para representar elegantemente la estructura gráfica.

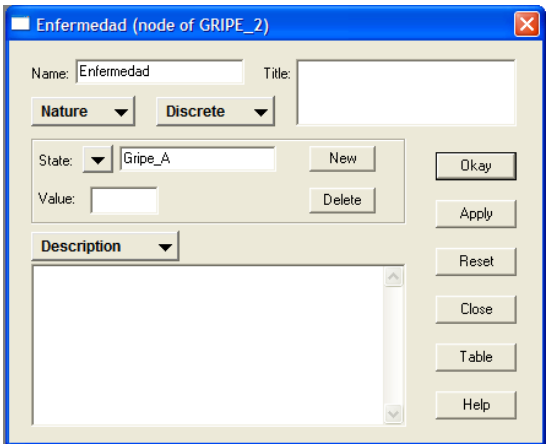

Figura 2. Cuadro de diálogo de las propiedades del nodo.

Supongamos que queremos crear el nodo "Dolor de Cabeza"; se sugiere escribir en la opción "Name" el texto "Dolor_de_Cabeza" mientras que en el cuadro de texto "Title" se podría escribir "Dolor de Cabeza". En el cuadro de texto desplegable llamado "State" escribimos "SI" y pulsamos el botón "New" para generar otro estado. Ahora escribimos "NO" y pulsamos en el botón "Okay". Repetimos el mismo procedimiento para generar los nodos "Enfermedad" y "Problemas Respiratorios". Hay que tener en cuenta que tampoco están permitidos los espacios al definir los estados o niveles de las variables por lo que se sugiere utilizar el guión bajo. Una vez creados los nodos que representarán las variables estableceremos los enlaces entre las variables. Para ello, hacemos clic en la opción "Modify → Add → Link" o presionamos la tecla "F12". A continuación, hacemos clic en la variable de origen, o "madre", (por ejemplo, en "Enfermedad") y luego sobre la variable de destino, o "hija", del enlace (por ejemplo, en "Dolor de Cabeza"). Se repite el procedimiento para el otro enlace hasta tener una estructura como la que aparece en la Figura 3.

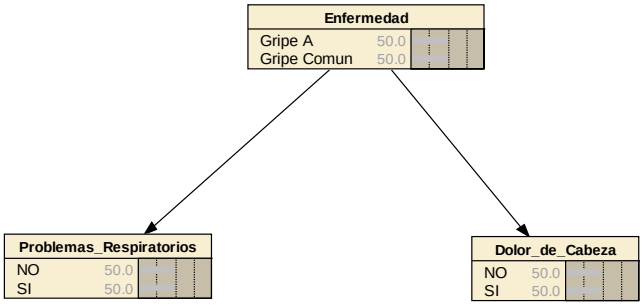

Figura 3. Estructura gráfica de la red *gripe* tras haber definido las propiedades de los nodos.

Cuando disponemos una base de datos que contiene un conjunto de casos para cada una de las variables que queremos modelar existe un procedimiento más rápido para generar los nodos. Aunque posteriormente tendremos que crear los enlaces manualmente, la definición de los nodos con sus respectivos estados, se puede crear de manera mecánica. Consideremos que disponemos de la base de datos que aparece en la Tabla 1, para generar nodos de manera automática hay que proceder del siguiente modo.

Caso	Dolor_de_Cabeza	Problemas_Respiratorios	Enfermedad
1	SI	NO	Gripe_Comun
2	SI	NO	Gripe_Comun
3	NO	SI	Gripe_A
4	SI	NO	Gripe_Comun
5	SI	NO	Gripe_Comun
6	SI	SI	Gripe_A
7	SI	NO	Gripe_Comun
8	SI	NO	Gripe_Comun
9	SI	NO	Gripe_Comun
10	SI	NO	Gripe_Comun
11	NO	NO	Gripe_Comun
12	SI	SI	Gripe_A
13	SI	NO	Gripe_Comun
14	SI	NO	Gripe_Comun
15	SI	SI	Gripe_A
16	SI	NO	Gripe_Comun
17	SI	NO	Gripe_Comun
18	SI	NO	Gripe_Comun
19	SI	NO	Gripe_Comun
20	SI	NO	Gripe_Comun

Tabla 1. Base de datos para la estimación.

En un archivo nuevo accedemos al comando "Cases → Add Case File Nodes…" y hacemos clic sobre él. Aparecerá un cuadro de diálogo titulado "Case file to obtain nodes from". En este momento tendremos que especificar un archivo que contenga nuestra base de datos. Netica es capaz de leer archivos de texto plano separado por tabulaciones (.txt), archivos de Excel (.xls o .xlsx) y un tipo de archivo propio de Netica (.cas) entre otros formatos. Al hacer doble clic sobre el archivo que contiene los datos que aparecen en la Tabla 1 aparecerán cuatro nodos en nuestra nueva red. A continuación habrá que crear los enlaces entre las variables como se ha descrito anteriormente para obtener una estructura divergente.

2.2.- Estimación paramétrica

De igual manera a como ocurre con la estimación estructural, la estimación de los parámetros de una red bayesiana puede hacerse automáticamente o a partir del juicio de expertos. Existen diversos algoritmos como el CB (Mani et al., 1997), el ELR (Greiner, Su, Shen, y Zhou, 2005; Greiner y Zhou, 2002; Shen, Su, Greiner, Musilek, y Cheng, 2003) o el EM (Cowell et al., 1999) que han sido diseñados para estimar los parámetros de una red bayesiana atendiendo a diferentes condicionantes estadísticos. Netica incorpora tres algoritmos diferentes para estimar los parámetros de una red bayesiana: un método basado en la frecuencia relativa conjunta, el algoritmo EM y un algoritmo que evalúa la reducción del gradiente de aprendizaje (similar a los implementados en las Redes Neurales Artificiales que

llevan a cabo computaciones del tipo backpropagation). En este trabajo se va a explicar cómo utilizar el método de estimación basado en la frecuencia relativa ya que es el más recomendable en casos en los que no se presuponen variables latentes y no hay una alta presencia de casos perdidos.

La versión más sencilla del algoritmo de máxima verosimilitud basado en las frecuencias relativas (conjuntas) queda expresado matemáticamente con la ecuación (1)

$$p(x_i \mid x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)})}{n(x_{\pi(i)})}, \tag{1}$$

donde $n(x_{\pi(i)})$ se refiere al número de casos que contiene la base de datos en los que las variables $X_{\pi(i)}$ toman el valor $x_{\pi(i)}$ y $n(x_i, x_{\pi(i)})$ es el número de casos en que $X_i = x_i$ y $X_{\pi(i)} = x_{\pi(i)}$. No obstante, el uso de este modelo de estimación puede dar lugar a dos tipos de problemas. Por un lado, podría generar estimaciones no definidas que se producen cuando alguna combinación particular de estados de variables no está presente y, por otro lado, cabría la posibilidad de incurrir en estimaciones sobreajustadas que generarían parámetros sesgados en el caso de que haya combinaciones de estados en las variables que estén sub-representadas o sobre-representadas. Por ello, Netica usa una función que introduce un factor de corrección en la ecuación (1) basado en la Ley de la Sucesión de Laplace (Morales, 2006; Ng y Jordan, 2002) y que dejaría la ecuación del siguiente modo:

$$p(x_i \mid x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)}) + 1}{n(x_{\pi(i)}) + \mid X_i \mid}, \tag{2}$$

donde $\mid X_i \mid$ se refiere al número de estados que tiene la variable X_i.

La estimación de probabilidades basadas en el juicio de expertos es, en la mayoría de los casos, un proceso subjetivo (Nadkarni y Shenoy, 2004) y consiste en rellenar tablas de probabilidad condicional. Por ejemplo, Das (2004) desarrolló un método para ayudar a los expertos humanos a estimar las probabilidades necesarias para parametrizar tablas de probabilidad condicional teniendo en cuenta los sesgos sistemáticos que se comenten al evaluar probabilidades (Kahneman, 2003; Kahneman, Slovic, y Tversky, 1982; Kahneman y Tversky, 1973; Tversky y Kahneman, 1974). Otra opción para parametrizar un modelo de red bayesiana podría ser utilizar resultados de investigaciones previas donde estuvieron involucradas las variables de nuestra red. Consideremos los datos introducidos en el ejemplo anterior y veamos cómo se ubican los parámetros del modelo en una red bayesiana.

> *En primer lugar parametrizaremos la variable "Enfermedad". Para ello, la seleccionaremos en el grafo creado en la sección anterior haciendo clic sobre ella con el botón izquierdo del ratón. A continuación haremos clic en la opción "Table → View/Edit" o presionaremos la tecla "Control" y "T" simultáneamente. Nos aparecerá un cuadro de diálogo que contiene una tabla con una fila y dos columnas. Las columnas son para indicar la probabilidad de ocurrencia de las enfermedades definidas por los estados de la variable. Dado que el ejemplo expuesto anteriormente indicaba que la probabilidad de contraer la gripe A era del 22,727%, insertaremos este dato en la celda correspondiente. Bajo el estado correspondiente a la gripe común escribiremos 77,273 al ser el*

número que al sumarlo al anterior daría como resultado el 100% (Figura 4). Pulsamos en la tecla "Okay".

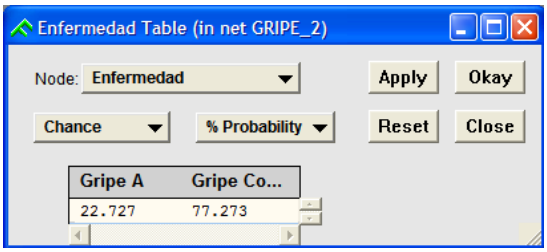

Figura 4. Tabla de probabilidad condicional para el nodo *Enfermedad*.

A continuación editaremos del mismo modo la tabla de probabilidad condicional del nodo "Problemas Respiratorios". En este caso la tabla es ligeramente más compleja ya que tiene tres columnas y tres filas. En la primera columna aparecerán los estados de la variable "Enfermedad" (gripe A y gripe común) mientras que en la columna dos y tres aparecerán los posibles estados que puede asumir la variable "Problemas respiratorios". Cada celda de la tabla de probabilidad condicional indica la probabilidad de que la variable "Problemas Respiratorios" tome un valor concreto (Sí o No) bajo la condición de que la variable "Enfermedad" tome un valor concreto. Ya que el problema especifica que la probabilidad de sufrir problemas respiratorios dado que se sufre la gripe común es del 83,33% [p(Problemas Respiratorios = Sí | Enfermedad = Gripe A) = 0,8333] escribiremos este valor en la casilla donde coinciden estos posibles estados de las variables. En la casilla vacía de la fila escribiremos el valor que al sumarlo a éste anterior de cómo resultado 100. De este modo podemos parametrizar la fila que nos queda y el nodo referido al dolor de cabeza (Figura 5).

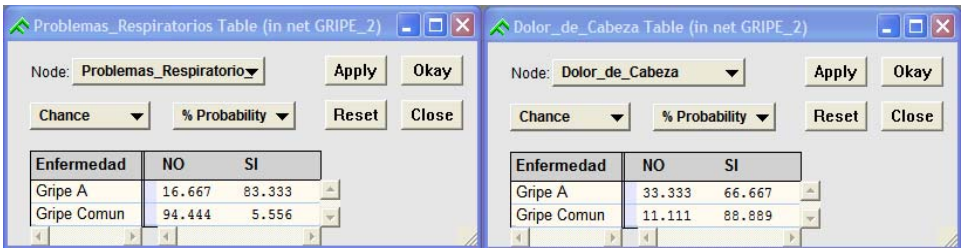

Figura 5. Tablas de probabilidad condicional para el nodo *Problemas respiratorios* (izquierda) y para el nodo *Dolor de Cabeza* (derecha).

Si queremos que Netica parametrize todas las tablas de probabilidad condicional que contiene nuestra red a partir de una base de datos podemos proceder del siguiente modo.

Hacemos clic en el comando "Cases → Incorp Case File". En el cuadro de diálogo que aparece se nos requiere que especifiquemos un archivo que contenga nuestra base de datos. La seleccionamos y aceptamos las opciones que nos dan por defecto. Una vez hecho esto, cada uno de los nodos de la red estarán parametrizados con base en el método de estimación máximoverosimil basado en la frecuencia y corregido con la Ley de la Sucesión de Laplace descrito

anteriormente. Para ver las tablas de probabilidad condicional podemos utilizar el comando anteriormente mencionado o presionar simultáneamente las teclas "Control" y "T".

3.- Uso

Antes de poder usar nuestra red para solucionar las preguntas que nos planteaba el ejemplo expuesto con anterioridad necesitamos compilar el modelo para que se ponga a funcionar en modo de inferencia. El proceso de compilación consiste en generar una distribución previa de las probabilidades asociadas a cada uno de los estados de cada variable utilizando el Teorema de la Probabilidad Total (Martínez, Martínez y Martínez, 2002; Serrano, 2003). Este teorema establece que si disponemos de un conjunto mutuamente excluyente de eventos $\varphi = \{A_1, A_2, A_1, \ldots A_n\}$ cuyas probabilidades suman la unidad, entonces la probabilidad arbitraria de un evento B vendría definida por la expresión

$$p(B) = \sum p(B \mid A_i) \times p(A_i). \qquad (3)$$

Para compilar nuestra red bayesiana tendremos que hacer clic en el comando "Network → Compile". Una vez hecho esto se observará que aparecen barras de color oscuro al lado de cada estado en cada variable precedidas por un valor (Figura 6). Las barras son una representación gráfica de la probabilidad previa estimada para cada estado. Por ejemplo, lo más probable que esté sucediendo a priori, sin conocer ninguna otra información sobre el paciente, cuando una persona llega a nuestra consulta es que esté sufriendo algún tipo de gripe común como se puede apreciar en su probabilidad asociada en el nodo "Enfermedad" (77,3% de posibilidades).

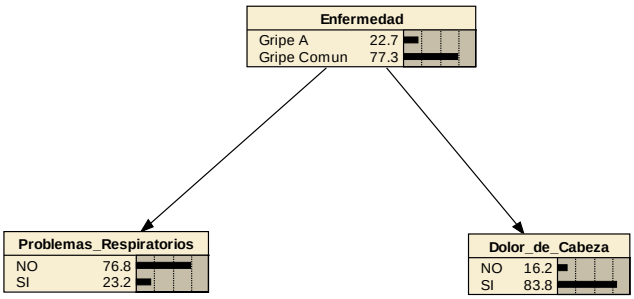

Figura 6. Red bayesiana compilada.

Sin embargo, si, como plantea el ejemplo anterior, el paciente nos dice que ha experimentado dolores de cabeza en los dos últimos días podríamos estimar la probabilidad de que esa persona sufra gripe común o gripe A. Para mostrarle esta evidencia al modelo, únicamente tenemos que hacer clic con el botón izquierdo del ratón sobre el estado "Sí" de la variable "Dolor de Cabeza" (diremos que la red bayesiana ha sido ***instanciada****). Al hacer esto, la probabilidad de este estado pasará a 100 y el resto de probabilidades del modelo se actualizarán (Figura 7). En este caso, ante esta nueva evidencia, la*

probabilidad de que el paciente sufra gripe común habrá aumentado hasta el 0,819.

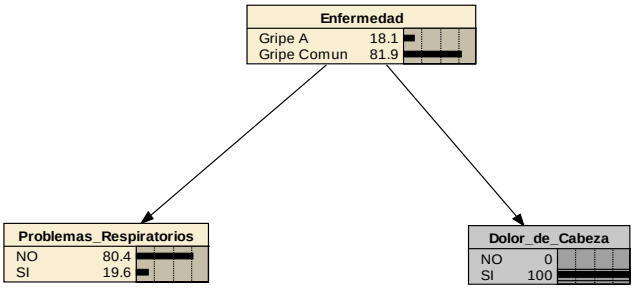

Figura 7. Red bayesiana instanciada con una evidencia sobre la variable *Dolor de Cabeza*.

No obstante, imaginemos que a lo largo de la consulta descubrimos que el paciente tiene serios problemas para respirar. En este caso, al introducir esta nueva evidencia en el modelo la enfermedad más probable para nuestro paciente ficticio ha pasado a ser la gripe A con un 76,8% de probabilidad (Figura 8).

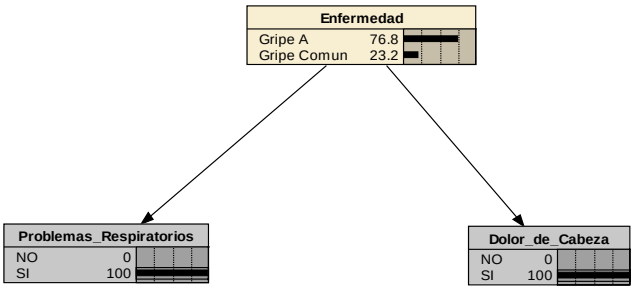

Figura 8. Red bayesiana instanciada con evidencias sobre las variables *Dolor de Cabeza* y *Problemas Respiratorios*.

4. – Validación

Tras haber creado nuestra red bayesiana tenemos la posibilidad de evaluar el grado en que su comportamiento se ajusta a un conjunto de datos. Por lo general, se suelen llevar a cabo estudios de validez cruzada. Esto es, se estima el modelo con una porción aleatoria de la muestra, por lo general del 70% o el 80%, y seguidamente se testa el modelo con el 30% o 20% restante respectivamente. En la medida en que el modelo se ajusta a este "nuevo" conjunto de datos podríamos decir que tenemos una evidencia sobre su validez. Por lo general, los estadísticos que genera Netica son aplicables a variables individuales y, aunque su interpretación se puede entender en términos globales, están referidos a la bondad de ajuste de una variable dentro del modelo.

Netica permite estimar tres estadísticos que evalúan el grado de ajuste del modelo en comparación con un conjunto de datos nuevos: la pérdida logarítmica, la pérdida cuadrática y la compensación esférica (López y García, 2011a; Pearl, 1978). La pérdida logarítmica oscila entre cero e infinito indicando cero la mejor bondad de ajuste. Por su parte, la pérdida cuadrática (o brier score) oscila entre cero y dos donde cero correspondería con una mejor

ejecución. Por último, la compensación esférica está acotada entre cero y uno, indicando uno un ajuste perfecto entre el modelo y los datos.

Netica también genera una matriz de confusión o tabla de clasificación donde se comparan las predicciones hechas por el modelo con lo realmente observado. Así, la matriz contendrá tantas filas y columnas como estados tenga el nodo que está siendo objeto del análisis. En las casillas de la matriz se representan el número de casos en que la red predijo un estado concreto en comparación con el estado que se observó en la base de datos de prueba. Un ajuste perfecto se concretaría con una diagonal que contenga frecuencias diferentes de cero y con ceros fuera de la diagonal. En relación con esto, Netica proporciona la tasa o el porcentaje global de errores (Error rate) en la clasificación de los nuevos datos que no han sido usados para estimar el modelo.

Cuando las variables son dicotómicas, el programa realiza un test de especificidad generando las coordenadas de una curva ROC (Receiver Operating Characteristic Curve) que evalúa la validez predictiva o clasificatoria del nodo. Sin embargo, los puntos de corte que utiliza son arbitrarios y no produce una estimación del área bajo la curva ROC. En caso de estar interesados en estos estadísticos, se recomienda usar la función de procesamiento de casos que se describirá a continuación y estimar el área bajo la curva ROC utilizando métodos clásicos (Hanley y McNeil, 1982, 1983) u otros programas informáticos (Franco y Vivo, 2007).

> *Supongamos que los datos que aparecen en la Tabla 2 son un fichero de texto plano delimitado por tabulaciones y que no han sido utilizados para parametrizar nuestro modelo de red bayesiana. Si hacemos clic el nodo "Enfermedad" con el botón izquierdo del ratón y accionamos el comando "Cases → Test With Cases" nos aparecerá un cuadro de diálogo que nos pide el archivo que contiene los datos de la Tabla 2. Al seleccionarlo y tras pulsar el botón "Abrir" los resultados del análisis aparecerán en una nueva ventana en formato de texto. Como se puede apreciar (Figura 9) ha habido cinco casos en los que la red ha predicho que la enfermedad era la "Gripe A" cuando realmente fue así, mientras que ha habido 13 casos en que la red clasificó correctamente a los pacientes cuando padecían "Gripe Común". Por su parte, únicamente dos casos fueron clasificados erróneamente, lo que supone un 10% de tasa de errores. Los estadísticos de pérdida logarítmica, pérdida cuadrática y compensación esférica también denotan un ajuste bastante aceptable. Por un lado, la pérdida logarítmica y la pérdida cuadrática están muy cercanas a cero; mientras que la compensación esférica tiene un valor muy cercano a uno.*

.

Caso	Problemas_Respiratorios	Enfermedad	Dolor_de_Cabeza
1	NO	Gripe_Comun	SI
2	SI	Gripe_A	SI
3	NO	Gripe_Comun	SI
4	SI	Gripe_A	NO
5	SI	Gripe_A	NO
6	NO	Gripe_Comun	SI
7	NO	Gripe_Comun	SI
8	SI	Gripe_A	SI
9	NO	Gripe_Comun	SI
10	NO	Gripe_Comun	SI
11	NO	Gripe_Comun	SI
12	NO	Gripe_Comun	SI
13	NO	Gripe_A	SI
14	NO	Gripe_Comun	SI
15	SI	Gripe_A	NO
16	NO	Gripe_Comun	SI
17	NO	Gripe_Comun	SI
18	NO	Gripe_Comun	SI
19	SI	Gripe_Comun	SI
20	NO	Gripe_Comun	SI

Tabla 2. Base de datos para la validación.

```
Confusion:
  ...Predicted..
  Gripe_  Gripe_    Actual
  ------  ------    ------
       5       1    Gripe_A
       1      13    Gripe_Comun

Error rate = 10%

Scoring Rule Results:

  Logarithmic loss = 0.2994
  Quadratic loss   = 0.1657
  Spherical payoff = 0.9112
```

Figura 9. Resultados de testar el modelo de red bayesiana con casos diferentes a los usados para la estimación.

Para generar un archivo que contenga las probabilidades estimadas para cada caso y respecto a un estado de la variable tenemos que, en primer lugar, generar un fichero de control y un archivo que contenga los casos de las variables que queremos utilizar como observaciones a evaluar. Por ejemplo, consideremos que

queremos estimar la probabilidad a posteriori del estado "Gripe_A" para el conjunto de datos que aparece en la Tabla 3. En este caso tendríamos que generar un archivo de control (un archivo de texto plano con extensión .txt) que contuviese la siguiente expresión:

IDnum()
bel (Enfermedad, Gripe_A)

A continuación tendríamos que ejecutar el comando "Cases → Process Cases". La ventana que aparece nos pide el archivo de control que contiene la sintaxis indicada anteriormente. Cuando especificamos cual es el archivo del control el programa nos pide el archivo que contiene los casos a procesar (Tabla 3) y, a continuación, nos demanda un nombre y una ubicación para el archivo que contendrá las probabilidades estimadas para el estado "Gripe_A" del nodo "Enfermedad" para cada caso del archivo procesado. Estas probabilidades pueden utilizarse para calcular estadísticos de verosimilitud como la lejanía o "deviance" (López y García, 2011a) o estadísticos relativos al porcentaje de varianza explicada por el modelo (DeMaris, 2002; Long, 1997).

.

Caso	Dolor_de_Cabeza	Problemas_Respiratorios
1	SI	NO
2	SI	SI
3	NO	SI
4	SI	NO
5	SI	SI

Tabla 3. Base de datos para el procesado de casos.

5.- Referencias

Aguilera, P. A., Fernández, A. Fernández, R., Rumí, R., y Salmerón, A. (En prensa). Bayesian networks in environmental modelling. *Environmental Modelling & Software*. doi: 10.1016/j.envsoft.2011.06.004

Batista, J. M., y Coenders, G. (2000). *Modelos de ecuaciones estructurales.* Madrid: Muralla/Hespérides.

Cooper, G. F., y Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309-347.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., y Spiegelhalter, D. J. (1999). *Probabilistic netwoks and expert systems*. Harrisonburg, VA: Springer.

Das, B. (2004). Generating conditional probabilities for bayesian networks: easing the knowledge acquisition problem. Descargado el 10 de Septiembre, 2005, desde http://arxiv.org/abs/cs.AI/0411034.

DeMaris, A. (2002). Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociological Methods & Research, 31*, 27–74.

Edwards, W. (1998). Hailfinder. Tools for and experiences with bayesian normative modeling. *American Psychologist, 53*, 416–428.

Franco, M., y Vivo, J. M. (2007). *Análisis de curvas ROC. Principios y aplicaciones.* Madrid: La Muralla.

Gámez, J. A. (1998). Abducción en modelos gráficos. En J. A. Gámez y J. M. Puerta (Eds.), *Sistemas expertos probabilísticos* (pp. 79–111). Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha.

Glymour, C. (2001). *The mind's arrows. Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitives Sciences, 7*, 43–48.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., y Danks, D. (2004). A theory of causal learning in children: causal and bayes nets. *Psychological Review, 111*, 3–32.

Gopnik, A., y Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitives Sciences, 8*, 371–377.

Greiner, R., Su, X., Shen, B., y Zhou, W. (2005). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Machine Learning, 59*, 297–322.

Greiner, R., y Zhou, W. (2002). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Proceedings of the Eighteenth Annual National Conference on Artificial Intelligence, Aug, 2002*, 167–173.

Hanley, J. A., y McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Hanley, J. A., y McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Heckerman, D. (1995). *A tutorial on learning with bayesian networks* (Rep. Téc. MS-TR-95-06). Redmon, WA: Microsoft Research.

Heckerman, D., Mamdani, A., y Wellman, M. P. (1995). Real-world applications of bayesian networks. *Communications of the Association for Computing Machinery, 38* (3), 24–26.

Herskovits, E. H., y Dagher, A. P. (1997). *Applications of bayesian networks to health care* (Rep. Téc. NSI-TR-1997-02). Baltimore, MD: Noetic Systems.

Holyoak, K.H., y Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology, 62*, 135-163.

Huete, J. F. (1998). Sistemas expertos probabilísticos: modelos gráficos. En J. A. Gámez y J. M. Puerta (Eds.), *Sistemas expertos probabilísticos* (pp. 1–40). Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha.

Jansen, R., Yu, H., Greenbaum, D., Kluger, N., Krogan, N. J., Chung, S., et al. (2003, Octubre 17). A bayesian network approach for predicting proteinprotein interactions from genomic data. *Science, 302*, 449–453.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist, 58*, 723–730.

Kahneman, D., Slovic, P., y Tversky, A. (1982). *Judgement under uncertainty: heuristic and biases*. New York: Cambridge University Press.

Kahneman, D., y Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Korb, K. B., y Nicholson, A. E. (2004). *Bayesian artificial intelligence*. Boca Raton, FL: Chapman & Hall/CRC.

Lagnado, D. A., Waldman, M. R., Hagmayer, Y., y Sloman, S. A. (2007). Beyond covariation: cues to causal structure. En A. Gopnik y L. Schulz (Eds.), *Causal learning: psychology, philosophy, and computation* (pp. 154-172). Londres: Oxford University Press.

Lee, S. M., Abbott, P., y Johantgen, M. (2005). Logistic regression and Bayesian networks to study outcomes using large data sets. *Nursing Research, 2*, 133–138.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE Publications.

López, J. (2009). *Modelos predictivos en actitudes emprendedoras: análisis comparativo de las condiciones de ejecución de las redes bayesianas y la regresión logística.* Tesis doctoral no publicada, Facultad de Psicología, Universidad de Almería.

López, J. y García, J (2011a). *Utilidad de las redes bayesianas en psicología.* Almería: Servicio de Publicaciones de la Universidad de Almería.

López, J. y García, J (2011b). Eventos por variable en regresión logística y redes bayesianas para predecir actitudes emprendedoras. *Revista Electrónica de Metodología Aplicada, 16*, 43-58.

López, J. y García, J. (en prensa) Comparative Study on Entrepreneurial Attitudes Modelled with Logistic Regression and Bayes Nets. *The Spanish Journal of Psychology*.

López, J., García, J., Cano, C. J., Gea, A. B., y De la Fuente, L. (2009, Septiembre). *A definition of potential entrepreneur from a probabilistic point of view*. Comunicación presentada en el XI Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

López, J., García, J., De la Fuente, L., y De la Fuente, E. I. (2007). Las redes bayesianas como herramienta de modelado en psicología. *Anales de Psicología, 23*, 307–316.

López, J., y García, J. (2007). Valores, actitudes y comportamiento ecológico modelados con una red bayesiana. *Medio Ambiente y Comportamiento Humano, 8*, 159–175.

Mani, S., McDermmott, S., y Valtorta, M. (1997). MENTOR: a bayesian model for prediction of mental retardation in newborns. *Research in Developmental Disabilities, 18*, 303–318.

Martínez, A., Martínez, S., y Martínez, H. (2002). *Estadística empresarial.* Almería: Servicio de Publicaciones de la Universidad de Almería.

Morales, M. E. (2006). *Modelización y predicción en estadística universitaria.* Tesis doctoral no publicada, Facultad de Ciencias Experimentales, Universidad de Almería.

Nadkarni, S., y Shenoy, P. P. (2004). A causal mapping approach to constructing bayesian networks. *Decision Support Systems, 38*, 259–281.

Ng, A. Y., y Jordan, M. I. (2002). On discriminative vs. generative classifiers: a comparation of logistic regresion and naive Bayes. *Advances in Neural Information Processing Systems, 14*, 841–848.

Pearl, J. (1978). An economic basis for certain methods of evaluating probabilistic forecastst. *International Journal of Man-Machine Studies, 10*, 175-183.

Ruiz, F., García, M. E., y Pérez, A. (2005). *Estilos de vida en ciudad de la Habana. Hábitos físico-deportivos y de salud*. Madrid: Gymnos.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., y Richardson, T. (2005). TETRAD 3: tools for causal modeling. User's manual. Descargado el 14 de Febrero de 2005, desde http://www.phil.cmu.edu/projects/tetrad/.

Serrano, J. (2003). *Iniciación a la estadística bayesiana*. Madrid: Muralla / Hespérides.

Shen, B., Su, X., Greiner, R., Musilek, P., y Cheng, C. (2003, Noviembre). *Discriminative parameter learning of general bayesian network classifiers*. Comunicación presentada en la 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03). Sacramento, California.

Spirtes, P., Glymour, C., y Scheines, R. (2000). *Causation, prediction and search* (2ª ed.). Cambridge, MA: MIT Press.

Tversky, A., y Kahneman, D. (1974, Septiembre 27). Judgement under uncertainty: heuristic and biases. *Science, 185*, 380–400.