# Un algoritmo de diagnóstico para modelado del alumno basado en test adaptativos y redes bayesianas

Eva Millán y José Luis Pérez-de-la-Cruz

Departamento de Lenguajes y Ciencias de la Computación
E.T.S.I. Informática, Universidad de Málaga. Apdo. 4114, Málaga 29080. Spain
e-mail: eva, perez@lcc.uma.es

***Resumen***: En este artículo presentamos un algoritmo de diagnóstico para modelado del alumno basado en el uso conjunto de redes bayesianas y test adaptativos informatizados. La validez de este enfoque integrado ha sido puesta a prueba mediante el uso de alumnos simulados. Los resultados obtenidos son prometedores, ya que muestran que la sola aplicación del algoritmo de diagnóstico bayesiano tiene un comportamiento excelente, cuya eficiencia aún es posible mejorar (tanto en precisión como en rapidez) introduciendo criterios de selección de preguntas adaptativos.

## 1. Introducción

Las nuevas tecnologías han aportado al campo de la educación aspectos innovadores que suponen una mejora cualitativa en las formas de enseñar y aprender. Una de las principales innovaciones introducidas desde los primeros programas de enseñanza asistida por ordenador han sido los llamados *sistemas tutores inteligentes* (STI), que, a diferencia de los programas tradicionales, muestran la capacidad de adaptarse a cada uno de los alumnos que los usan para aprender, proporcionando así una *enseñanza individualizada*, que según se demuestra en (Bloom, 1984) es el mejor método de enseñanza.

Por tanto, si la característica clave de un sistema tutor inteligente es la capacidad de adaptarse al alumno, la componente clave de dicho sistema es el denominado *modelo del alumno*, donde se almacena la información relativa al alumno. Dicha información debe ser inferida por el propio sistema a partir de la información que tenga disponible: datos previos, respuestas a preguntas, etc. Este proceso de inferencia se denomina *diagnóstico*, y es sin duda el proceso más complicado dentro de un sistema tutor inteligente, dado que además de la dificultad que supone conlleva tratamiento de información que en muchos casos es incierta o imprecisa. Sin embargo, en muchos de los STI existentes se prefiere el desarrollo de heurísticos ad-hoc (que por su falta de fundamento teórico muestran a veces comportamientos no deseados) antes que utilizar alguna de las técnicas disponibles para razonamiento aproximado dentro del campo de la Inteligencia Artificial (IA), o teorías con un fundamento teórico más sólido como las que subyacen en los Test Adaptativos Informatizados (TAIs).

El objetivo de nuestro trabajo ha sido profundizar en el estudio de diversos procedimientos de diagnóstico con objeto de obtener modelos del alumno más precisos. Para ello se ha buscado la aplicación de técnicas procedentes de la IA que tengan un fundamento teórico consistente, pero poniendo especial énfasis en simplificar su uso de modo que no suponga una carga excesiva de trabajo adicional a la tarea ya de por sí considerable de desarrollar un sistema tutor inteligente. Además,

hemos buscado mecanismos que permitan aumentar la eficiencia de dichas técnicas en el proceso de diagnóstico (donde por una mejora en eficiencia entendemos una mejora tanto en la precisión del diagnóstico como en el tiempo que cuesta obtenerlo). La solución propuesta se basa en la integración en un mismo marco formal de dos teorías diferentes: las redes bayesianas (RB) y los TAIs.

El presente trabajo se estructura de la siguiente forma: en la sección 2 presentamos de forma sencilla y concisa las RBs. En la sección 3 describimos en primer lugar la RB que representará el conocimiento del alumno y sus interacciones con el sistema y posteriormente nos centramos en los test adaptativos basados en RBs. En la sección 4 evaluamos el modelo integrado propuesto (RB + Algoritmo TAI bayesiano) utilizando alumnos simulados. Tras una breve revisión de trabajos relacionados, finalizamos presentando las conclusiones junto con las líneas de trabajo futuro.

## 2. Fundamentos: redes bayesianas

De un modo informal, podemos decir que una RB es un conjunto de nodos y arcos. Cada nodo corresponde a una variable, que a su vez representa una entidad del mundo real, y los arcos que unen los nodos indican relaciones de *influencia causal* entre las variables. Veamos un ejemplo sencillo.

***Ejemplo 1.***

La RB no trivial más simple que podemos imaginar consta de dos variables, y un arco dirigido entre ellas. Supongamos que *C* representa el conocimiento del alumno sobre cierto concepto C y $P_1$ su capacidad de resolver correctamente cierta pregunta $P_1$ relativa a dicho concepto. Entonces, que el alumno sepa el concepto C tiene influencia causal en que sea capaz de responder bien a la pregunta $P_1$, lo cual se expresa mediante el arco dirigido que aparece en la Figura 1.

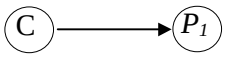



*Figura 1 RB con dos nodos.*

La notación que usaremos será la siguiente: si *X* es una variable binaria, denotaremos por *x* a cualquier valor de la variable X, por $+x$ la presencia de aquello a lo que representa y por $\neg x$ a su ausencia. Así, por ejemplo en este caso $+c$ significará "el alumno conoce el concepto C" y $\neg p_1$ "el alumno no es capaz de resolver correctamente la pregunta $P_1$".

La información cuantitativa (parámetros) de una RB viene dada por:

- La probabilidad a priori de los nodos que no tienen padres.
- La probabilidad condicionada de los nodos con padres.

Por tanto, en nuestro ejemplo, los datos que debemos conocer son $P(c)$ y $P(p_1 / c)$. Así, la RB completa sería la que se muestra en la Figura 2.

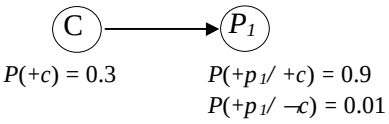



*Figura 2 RB con parámetros.*

Veamos qué significado tienen en este caso estos valores:

- $P(+c) = 0.3$ indica que el 30% de los alumnos del grupo en estudio conocen el concepto.
- $P(+p_1 / +c) = 0.9$ indica que el 90% de los alumnos que conocen el concepto C responden correctamente a la pregunta $P_1$. Esto quiere decir que incluso los alumnos que conocen el concepto pueden tener un despiste y contestar mal a la pregunta (en una proporción del 10%).
- $P(+p_1 / \neg c) = 0.01$ significa que sólo el 1% de los alumnos que no conocen el concepto C son capaces de contestar correctamente a la pregunta $P_1$. Este parámetro indica por tanto qué alumnos que no conocen el concepto pueden adivinar la respuesta correcta a la pregunta $P_1$.

En el campo de la medicina, estos parámetros tienen una interpretación muy sencilla: supongamos que tenemos una red que representa la relación entre padecer o no cierta enfermedad E y el resultado de un test T que se utiliza para el diagnóstico de la enfermedad E. La RB se representa en la Figura 3.

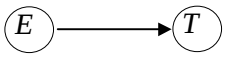



*Figura 3 RB para diagnóstico médico.*

donde

- $P(+e)$ representa el tanto por ciento de la población en estudio que padece la enfermedad E, es decir, la *prevalencia* de E.
- $P(+t / +e)$ indica el tanto por ciento de pacientes que dan positivo en el test T entre los que padecen la enfermedad E. Esto se conoce como *sensibilidad* del test.
- $P(+t / \neg e)$ indica el tanto por ciento de pacientes que dan positivo en el test T entre los que no padecen la enfermedad E. A la probabilidad complementaria $P(\neg e / \neg t)$, es decir, a la proporción de pacientes que dan negativo en el test entre los que no padecen la enfermedad se le llama *especificidad* del test.

En medicina siempre se buscan los tests con mayor grado de sensibilidad y especificidad. Esta semántica puede extenderse al caso del modelado del alumno, así que a partir de ahora hablaremos también de la *sensibilidad* y *especificidad* de una pregunta para un concepto.

La potencia de las RBs radica en que, supuestas las llamadas *condiciones de independencia condicional* (que son unas condiciones de independencia entre las variables del grafo que resultan *razonables*), una vez especificada la estructura de la red es posible realizar cualquier tipo de inferencia dada la información disponible. Es decir, podemos hacer inferencias predictivas (si el alumno conoce el concepto C, ¿cuál es la probabilidad de que responda correctamente a la pregunta $P_1$?) o abductivas (si el alumno ha contestado incorrectamente a la pregunta $P_1$, ¿cúal es la probabilidad de que conozca el concepto C?). De esta forma, un mismo nodo puede ser tanto fuente de información como objeto de predicción. Dichas inferencias se realizan aplicando

algoritmos de propagación de probabilidades que se han desarrollado específicamente para tal fin, y que dependiendo del tipo de inferencia que queramos hacer y la estructura de la red pueden resultar más o menos complicados.

De este modo, para utilizar una RB lo único que debemos hacer es identificar las variables y las relaciones de influencia causal entre ellas, y cuantificar estas relaciones de influencia causal asignando las probabilidades condicionadas y a priori. Una vez definida la red podemos utilizar alguno de los paquetes software comerciales o de libre disposición como HUGIN y NETICA para realizar las inferencias que nos interesen. Esta sección pretende ser una introducción sencilla a las RBs. El lector interesado puede encontrar información más rigurosa y detallada en (Charniak, 1991; Neapolitan, 1990; Pearl, 1988; Castillo, Gutiérrez, & Hadi, 1997).

## 3. Test adaptativos bayesianos como algoritmo de diagnóstico

A pesar de la gran potencia y versatilidad de las RBs en los problemas de diagnóstico, su utilización en el problema del modelado del alumno no es muy frecuente en comparación con la gran cantidad de sistemas desarrollados. Así por ejemplo en las últimas conferencias de mayor relevancia en el campo (Intelligent Tutoring Systems, ediciones 1998 y 2000 y Artificial Intelligence in Education, ediciones 1999 y 2001) vemos que en cada edición hay solamente entre dos y cuatro trabajos que utilizan RBs. En nuestra opinión, la principal causa de esta escasez de modelos del alumno bayesianos estriba en que aplicar un modelo de RBs exige mucho más esfuerzo que aplicar otros modelos de razonamiento aproximado (factores de certeza, lógica difusa) o desarrollar un heurístico ad hoc para la definición y actualización del modelo del alumno. Este esfuerzo adicional viene provocado principalmente por dos causas:

a) *Especificación de la red* (estructura y parámetros). Especificar una RB exige estudiar cuidadosamente cuáles son las variables que intervienen en el sistema y las relaciones de influencia causal entre ellas. Además, una vez definida la red se deben estimar las probabilidades condicionadas asociadas, que normalmente son un número bastante grande de parámetros difíciles de estimar.

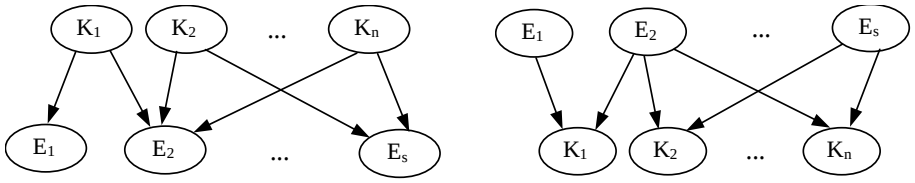
b) *Dificultad de implementar los algoritmos de propagación de probabilidades*, que además de ser más o menos complejos, son computacionalmente muy costosos.

Con nuestro trabajo (Millán & Pérez-de-la-Cruz 2002, Millán 2000 y Millán, Pérez-de-la-Cruz & Triguero , 1998) pretendemos paliar en lo posible la dificultad de especificar la red, proporcionando una estructura fija para la RB que representa el conocimiento del alumno y su relación con las variables que actuarán como fuentes de información (preguntas y ejercicios) y una forma sencilla de especificar los parámetros. Veremos además que al utilizar criterios adaptativos de selección de preguntas es posible mejorar la precisión y rapidez del proceso de diagnóstico. En la sección 3.1 vamos a describir la red bayesiana que servirá de soporte al algoritmo de diagnóstico basado en test adaptativos bayesianos que se presenta en la sección 3.2 y se evalúa en la sección 4.

### 3.1 Modelo estructural

Pasamos pues a describir las redes bayesianas que utilizaremos para modelar las relaciones entre los nodos que representarán el conocimiento del alumno (*nodos de*

*conocimiento*) y los nodos que utilizaremos para recolectar la evidencia disponible (*nodos evidenci*a). Sean pues $K_1$, ..., $K_n$ nodos de conocimiento y $E_1$, …, $E_s$ nodos evidencia. Las dos estructuras posibles para la red se muestran en la Figura 4.




*Alternativa* 1 *Alternativa 2*

*Figura 4 Alternativas para modelar las relaciones conocimiento-evidencia.*

La primera alternativa se basa directamente en relaciones de causalidad, tal como las percibimos: el conocimiento de ciertas partes del currículum tiene *influencia causal* en que las situaciones en que debemos hacer uso de ese conocimiento se resuelvan o no correctamente. La segunda alternativa se corresponde con una estructuración del conocimiento en forma de reglas: si alguna de las situaciones para las que hace falta tener determinado conocimiento se resuelve correctamente es porque el conocimiento se posee. Además para cada alternativa tenemos que:

a) En la alternativa 1, los parámetros a especificar serían las probabilidades a priori de conocer los $K_i$, $\{P(K_i), i = 1, ..., n\}$, y las probabilidades condicionadas de los $E_j$ dados sus padres, es decir, $\{P(E_j / \{K_i / K_i \in Pa(E_j)\})$ para $j=1, ..., s\}$. En cuanto a independencias, esta red implicaría:

   - la independencia a priori de los $K_i$, para $i = 1, ..., n$;
   - $K_i$ es independiente a priori de $E_j$ para todo $E_j$ que no sea hijo de $K_i$, $i=1 ..., n$;
   - $E_j$ es independiente de todo $E_i$ (con $i \neq j$) dado $pa(E_j)$, $j = 1, ..., s$;
   - $E_j$ es independiente de $K_i$ para todo $i$ tal que $K_i \notin pa(E_j)$, $j =1 , ..., s$.

b) En la alternativa 2, los parámetros a especificar serían: la probabilidad a priori de los $E_j$, $\{P(E_j), j = 1, ..., s\}$, y las probabilidades condicionadas de los $K_i$ dados sus padres, es decir, $\{P(K_i / Pa(K_i), i = 1, ..., n\}$. Esta estructura implica las siguientes independencias:

   - la independencia a priori de los $E_j$ , para $j = 1, ..., s$;
   - $E_j$ es independiente a priori de $K_i$ para todo $K_i$ que no sea hijo de $E_j$ y para cada $j$ $=1, ..., s$;
   - $K_i$ es independiente de todo $K_j$ (con $i \neq j$) dado $pa(K_i)$, $i = 1, ..., n$;
   - $K_i$ es independiente de $E_j$ para todo $j$ tal que $E_j \notin pa(K_i)$, $i = 1, ..., n$.

   Vemos por tanto que esta red implicaría la independencia de los $K_i$ conocidas las evidencias, lo cual no es cierto. Veamos un contraejemplo simple: supongamos que para contestar a una pregunta $P$ se requiere tener conocimiento sobre $K_1$ y $K_2$, y que la pregunta $P$ se ha respondido incorrectamente. Entonces, saber que el alumno conoce $K_1$ debería implicar que no conoce $K_2$. Pero como las variables $K_1$ y $K_2$ están d-separados por $P$, son condicionalmente independientes en esta red, y, por tanto, la evidencia sobre $K_1$ no afectará a la probabilidad de conocer $K_2$ de la forma que debiera.

Por tanto optamos por la primera alternativa, que es la que más adecuadamente describe el comportamiento que queremos que tenga la red en este caso. De este modo, consideraremos que las relaciones entre nodos de conocimiento y nodos evidencia tienen la dirección representada en la Figura 5.

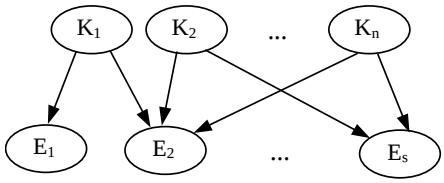



*Figura 5 Relaciones entre nodos de conocimiento y nodos evidencia.*

Los parámetros que tendremos que definir en esta parte de la red serán las probabilidades a priori de los conceptos y las condicionadas de las preguntas dados los conceptos. Estas probabilidades condicionadas pueden ser muy complicadas de estimar para los profesores que definen el test. Para simplificar en la medida de lo posible la especificación de estos parámetros hemos partido del enfoque presentado en (VanLehn, Niu, Siler, & Gertner, 1998), que consiste en considerar que:

- La probabilidad de que una pregunta se responda correctamente dado que se dominan todos los conceptos relativos a esa pregunta es 1-*s*, donde *s* es el factor de descuido.
- La probabilidad de que una pregunta se responda correctamente dado que alguno de los conceptos relativos a esa pregunta no se domina es *k/n*, donde *n* es el número de posibles respuestas y *k* es un factor que representa la probabilidad de que el alumno intente adivinar la respuesta correcta.

El inconveniente que presenta este enfoque es que considera igualmente probable que el alumno responda correctamente cuando sólo le falta conocer uno de los conceptos necesarios que cuando no conoce ninguno de ellos. Nosotros consideramos que esta probabilidad debe ser mayor conforme más conocimiento tenga el alumno, especialmente en preguntas tipo test en las que la elección de la respuesta correcta puede basarse en descartar aquellas que son incorrectas. El enfoque que nosotrosproponemos es el siguiente: sea $F(x)$ la función que determina la CCI en el modelo de tres parámetros en la TRI, es decir:

$$F(x) = c + \frac{1-c}{1+exp(-1.7a(x-b))} \qquad x \in I\!R$$

donde $c = 1/n$, $a$ es el índice de discriminación asociado a la pregunta y $b$ es el nivel de dificultad de la pregunta. A partir de $F$ definimos una función $G$ que no es más que una transformación lineal de F mediante la expresión:

$$G(x) = 1 - \frac{(1-c)(1+exp(-1.7ab))}{1+exp(1.7a(x-b))} \quad x \geq 0$$

El efecto de esta modificación se ilustra en la Figura 6.

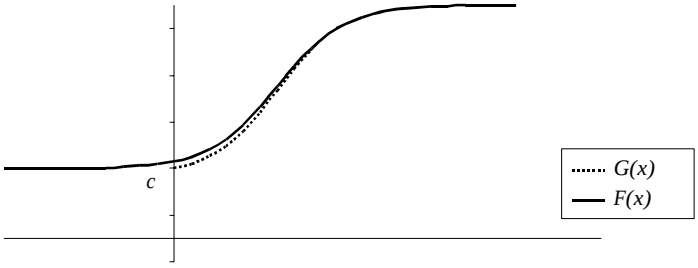



*Figura 6 Modificación de la CCI.*

Esta modificación la hacemos para que $G(0) = c$, puesto que la función $G$ será la que usaremos para obtener los valores que asignaremos a la probabilidad de responder correctamente a la pregunta según el número de conceptos conocidos por el alumno. Cuando el alumno no conoce ningún concepto, su probabilidad de contestar correctamente a la pregunta será $G(0) = c$. Cuando los conoce todos, será $1-s$. El resto de los valores se interpolan entre $c$ y $1-s$, según se ilustra en la Figura 7.

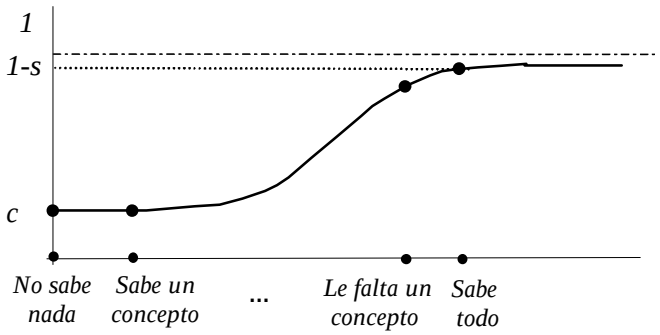



*Figura 7 Uso de la función $G(x)$ para calcular las probabilidades condicionadas.*

Para más detalles sobre el uso de la función G, ver (Millán, 2000) o (Millán & Pérez-de-la-Cruz, 2002).

Cabe resaltar que la estructura que hemos discutido en esta sección se engloba dentro de un marco general más amplio, en el que se modela el conocimiento del alumno a varios niveles de granularidad (conceptos / temas / asignatura). En la Figura 8 podemos ver este marco general, donde los nodos etiquetados con $T_i$ ($i$ =1, ..., $s$) representan los temas y el etiquetado con A la asignatura. Utilizando esta red, se lleva a cabo un proceso evaluador (representado también Figura 8), que consta de dos etapas:

- *Etapa de diagnóstico*, en la que se utilizará la parte de la red en la que tenemos los conceptos, las preguntas y las relaciones entre ellos. El objetivo de esta etapa será determinar, a partir de las respuestas dadas por el alumno, el conjunto de conceptos que éste conoce y que no conoce.
- *Etapa de evaluación*, donde, a partir de los resultados de la etapa anterior, se usará la propagación de probabilidades para determinar la calificación del alumno a dos niveles de granularidad diferentes, es decir, determinaremos la calificación para la asignatura y para cada uno de los temas de los que consta.

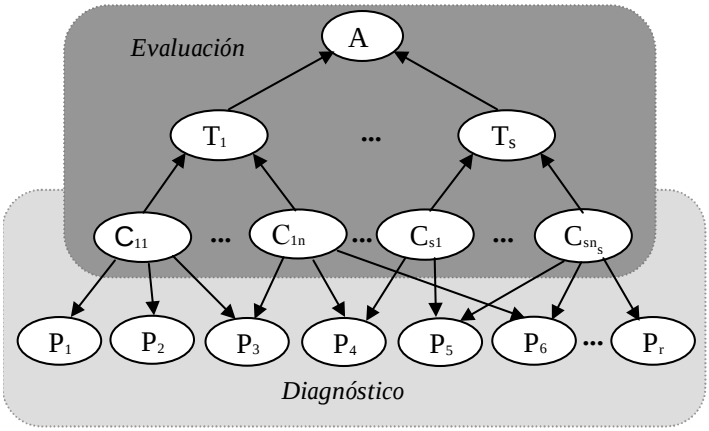



*Figura 8 Uso de la red bayesiana en el proceso evaluador*

Sin embargo, y dado que el objetivo principal de este artículo es describir el proceso de diagnóstico, no vamos a entrar en detalles sobre la jerarquía de granularidad ni sobre la etapa de evaluación. El lector interesado puede encontrar más información en en (Millán, 2000).

## 3.2 Test adaptativos basados en redes bayesianas

En esta sección discutiremos el algoritmo que proponemos para realizar tests adaptativos basados en RBs, que permite diagnosticar más de una habilidad simultáneamente. El algoritmo TAI bayesiano propuesto se lleva a cabo sobre la estructura de la red representada en la Figura 5. Para definirlo vamos a utilizar la descripción de los elementos básicos de un TAI que aparece en (Weiss & Kingsbury, 1984): modelo de respuesta asociado a cada pregunta, método de puntuación, banco de ítems o preguntas, nivel inicial, método de selección de preguntas y criterio de parada. Algunos de estos elementos vienen ya dados por el uso de redes bayesianas, mientras que otros necesitan de una mayor elaboración.

### 3.2.1 Modelo de respuesta asociado a cada pregunta

Una vez definida la red, el modelo de respuesta de cada pregunta viene dada por la distribución de probabilidad de la pregunta condicionada a sus padres. En la sección 3.1. se ha propuesto una simplificación para especificar estas probabilidades, como por ejemplo el uso de una función tipo logístico para medir la relación entre conocer una serie de conceptos y contestar correctamente a una pregunta relacionada con ellos.

### 3.2.2 Método de puntuación

Este elemento viene dado por el uso de un modelo bayesiano, ya que los algoritmos de propagación de probabilidades proporcionan un método consistente para valorar las respuestas, es decir, para estimar el nivel de conocimiento de los conceptos que intervienen en las preguntas de acuerdo con las respuestas dadas por el alumno.

### 3.2.3 Banco de preguntas

Respecto al banco de preguntas, el uso de la función logística proporciona una forma sencilla de especificar los parámetros (y por tanto de calibrar las preguntas), que tiene

en cuenta no sólo los errores no intencionados y la posibilidad de que el alumno adivine la respuesta correcta, sino también el hecho de que la probabilidad de dar la respuesta correcta aumenta conforme el conjunto de conocimientos del alumno es más completo. Además, permite introducir en las preguntas algunos de los parámetros usuales en la TRI: factor de adivinanza, dificultad, descuido y discriminación.

### 3.2.4 Nivel inicial

Como nivel inicial se puede usar la información disponible sobre el alumno o grupo de alumnos que van a tomar el test. En ausencia de información parece razonable considerar que es igualmente probable que el alumno conozca o no los conceptos elementales, es decir, asignar una distribución uniforme a los conceptos elementales.

### 3.2.5 Criterios de selección de preguntas

Como métodos de selección de preguntas proponemos varios criterios diferentes, que posteriormente han sido evaluados mediante el uso de alumnos simulados. El uso de estos criterios hará que las preguntas seleccionadas se adapten al nivel de conocimiento que ha demostrado el alumno hasta el momento. De este modo se pretende satisfacer el objetivo principal de un test adaptativo: mejorar la precisión del diagnóstico reduciendo el número de preguntas. A continuación vamos a describir tales criterios.

#### 3.2.5.1 Criterio aleatorio

El más sencillo es el criterio aleatorio, en el que cada pregunta de la base de datos tiene la misma probabilidad de ser elegida. Al criterio aleatorio lo denotaremos por $C_A$. Con este criterio se pretende simular un test tradicional, en el que de antemano se fija un número de preguntas que se le van a hacer al alumno. Sin embargo, el test basado en el criterio aleatorio es diferente de los test tradicionales, puesto que el método de diagnóstico y evaluación es bayesiano. Por supuesto este criterio no puede considerarse adaptativo, pero se introduce con el objeto de comparar posteriormente los resultados.

#### 3.2.5.2 Criterios adaptativos

En los criterios adaptativos la selección de la siguiente pregunta se hace en base al rendimiento que haya mostrado el alumno en las preguntas anteriores, y más concretamente en la estimación del nivel del conocimiento que se tenga del alumno en base a las respuestas a las preguntas anteriores. Hemos definido dos tipos de criterios adaptativos diferentes: los criterios basados en la cantidad de información que aporta cada pregunta y los criterios condicionados, que se basan en potenciar el comportamiento que ha demostrado el alumno hasta el momento.

##### 3.2.5.2.1 Criterios basados en la información

Definamos primero qué se entiende por utilidad de una pregunta *P* para un nodo de conocimiento *C*.

*Definición 1*

Dada una pregunta *P* y un nodo de conocimiento *C*, se define la *utilidad*$_1$ del nodo *P* para el nodo *C* como

$$U_1(P, C) = \left|P(C=1/P=1) - P(C=1)\right| P(P=1) + \left|P(C=0/P=0) - P(C=0)\right| P(P=0).$$

La interpretación de esta medida de utilidad es sencilla: la utilidad de un nodo evidencia se define como la *ganancia esperada de información*, puesto que lo que hacemos es calcular cuánto cambiará la probabilidad de *C* según el resultado del nodo evidencia *P*, y ponderar este cambio con la probabilidad de cada resultado. Por tanto, el nodo evidencia más informativo para cierto ítem será aquel que tenga utilidad máxima.

Por la forma de las relaciones en nuestra red, en la expresión de la utilidad$_1$ podemos prescindir de los valores absolutos, ya que dado que cuando se responda correctamente la probabilidad de conocer el concepto aumentará, y cuando se responda incorrectamente disminuirá. Por tanto podemos trabajar con la siguiente medida de utilidad:

$$U_1(P, C) = ( P(C=1/P=1) - P(C=1) ) P(P=1) + (P(C=0/P=0) - P(C=0) ) P(P=0).$$

En el contexto de los tests adaptativos, la pregunta más informativa será aquella de mayor utilidad. Vemos también que la utilidad de una pregunta se ve asimismo afectada por la estimación actual del nivel de conocimiento del alumno, ya que ponderamos por las probabilidades de responder la pregunta bien o mal que evidentemente dependerán del nivel de conocimiento actual del estudiante.

En el estudio realizado por Collins (Collins et al., 1996) sobre test adaptativos bayesianos, se define el concepto de utilidad como

$$U_C(P, C) = \left|P(C=1/P=1) - P(C=0/P=0)\right|.$$

Aunque los autores declaran haber tenido resultados satisfactorios en las simulaciones realizadas, en nuestra opinión esta medida de utilidad no es buena, ya que en todo caso si se quiere trabajar con estas probabilidades, ambas deberían ser maximizadas con lo cual no tiene ningún sentido maximizar el valor absoluto de la diferencia.

La medida de utilidad que proponemos tiene un inconveniente. En un test adaptativo, calcular la utilidad de las preguntas del banco de preguntas supone instanciar la red dos veces para cada pregunta (suponiendo respuesta correcta e incorrecta). Como el número de preguntas en un buen banco debe ser grande, este proceso puede ser demasiado costoso desde el punto de vista computacional, ya que el tiempo de espera del alumno debe minimizarse.

Sin embargo este problema es sencillo de solucionar. Para ello no hay más que aplicar el Teorema de Bayes en la definición del concepto de utilidad, y obtenemos que:

$$U_1(P,C) = ( P(P=1/C=1) - P(P=1) ) P(C=1) + (P(P=0/C=0) - P(P=0) ) P(C=0).$$

Lo cual supone instanciar los conceptos en lugar de las preguntas, resultando en un gran ahorro computacional, ya que en nuestras redes el número de conceptos es mucho

menor que el de preguntas. Asimismo, a la hora de realizar las instanciaciones necesarias para calcular la utilidad de una pregunta los cálculos se realizan en el subgrafo de nodos relevantes para el cálculo generado por el algoritmo orientado a un objetivo. De esta forma se ha conseguido que el tiempo que el estudiante tiene que esperar para que se le presente la siguiente pregunta sea muy pequeño (menor que un segundo en las pruebas realizadas con una red de catorce conceptos y cien preguntas).

Vamos a dar una definición alternativa para el concepto de utilidad:

*Definición 2*

Dado un nodo evidencia *P* y un nodo de conocimiento *C*, se define la *utilidad*$_2$ del nodo *E* para el nodo *C* como

$$U_2(P, C) = P(P=1/C=1)\, P(C=1) + P(P=0/C=0)\, P(C=0).$$

También esta medida de utilidad tiene una interpretación sencilla: lo que estamos haciendo es priorizar aquellas preguntas con mayor grado de sensibilidad y especificidad[1], o lo que es lo mismo, con menor tasa de *falsos positivos* (alumnos que responden bien sin saber el concepto) y *falsos negativos* (alumnos que responden mal aún sabiendo el concepto).

Otra interpretación para esta medida de utilidad viene de simplificar un poco la fórmula de cálculo:

$$U_2(P, C) = P(P=1 \wedge C=1) + P(P=0 \wedge C=0) = P(P = C).$$

Es decir, sería la probabilidad de que las variables *P* y *C* tomen el mismo valor.

Tenemos así dos definiciones diferentes para el concepto de utilidad: la basada en el aumento esperado de información y la basada en los conceptos de sensibilidad y especificidad.

Una vez calculada la utilidad de una pregunta para cada uno de los conceptos que en ella intervienen, queda por definir la utilidad global de la pregunta en función de las utilidades de los conceptos con ella relacionados. Según la definición de la utilidad global, se proponen dos criterios diferentes:

- *Criterio de la suma*, en el que la utilidad global de una pregunta se define como la suma de las utilidades de la pregunta para cada uno de los conceptos con ella relacionados, es decir:

$$U(P) = \sum_{C \in pa(P)} U(P,C)$$

- *Criterio del máximo*, en el que la utilidad global de una pregunta se define como el máximo de las utilidades de la pregunta para cada uno de los conceptos con ella relacionados, es decir:

[1] Véase la interpretación de los parámetros de una RB en el contexto de la medicina que se hizo en el ejemplo 3. 1.

$$U(P) = \max_{C \in pa(P)} U(P,C)$$

Combinando estas definiciones, tenemos cuatro criterios adaptativos basados en el concepto de utilidad:

- Criterio de la *suma* de las utilidades, definiendo la utilidad como la *ganancia de información*, que denotaremos por $C_{SG}$.
- Criterio del *máximo* de las utilidades, definiendo la utilidad como la *ganancia de información*, que denotaremos $C_{MG}$.
- Criterio de la *suma* de las utilidades, definiendo la utilidad en base a los conceptos de *especificidad* y *sensibilidad*, que denotaremos $C_{SE}$
- Criterio del *máximo* de las utilidades, definiendo la utilidad en base a los conceptos de *especificidad* y *sensibilidad*, que denotaremos $C_{ME}$.

#### 3.2.5.2.2 Criterios condicionados

Estos criterios se basan en potenciar que el diagnóstico vaya en la dirección que define el comportamiento del alumno en las preguntas previas. La utilidad de la pregunta se va a definir como la sensibilidad o la especificidad de la misma, según si el alumno está demostrando mayor o menor conocimiento. Hemos propuesto dos criterios diferentes:

- *Criterio condicionado a la probabilidad del concepto.* La utilidad de una pregunta se calcula mediante la siguiente expresión:

$$U(P) = \max_{C \in Pa(P)} U'(P,C),$$

donde $U'(P,C)$ se define como:

$$U'(P,C) = \begin{cases} P(P=1/C=1) & \text{si } P(C=1) > P(C=0) \\ P(P=0/C=0) & \text{en otro caso.} \end{cases}$$

La idea de este criterio consiste en elegir la pregunta más específica o más sensible según si el alumno está demostrando poseer conocimiento acerca de los conceptos o no poseerlo. Lo denotaremos por $C_{CC}$.

- *Criterio condicionado a la probabilidad de la pregunta.* La utilidad de una pregunta se calcula mediante la siguiente expresión:

$$U(P) = \begin{cases} \max_{C \in pa(P)} P(P=1/C=1) & \text{si } P(P=1) > P(P=0) \\ \max_{C \in pa(P)} P(P=0/C=0) & \text{en otro caso.} \end{cases}$$

Este criterio es similar al anterior, pero en lugar de elegir la sensibilidad o la especificidad en función a la probabilidad del concepto se elige en función de la probabilidad de la pregunta. Lo denotaremos por $C_{CP}$.

Para resumir, los siete criterios que analizaremos y compararemos son[2]:

- Criterio aleatorio, $C_A$.
- Criterio de la suma de las utilidades, donde la utilidad se define como la ganancia esperada de información, $C_{SG}$.
- Criterio del máximo de las utilidades, donde la utilidad se define como la ganancia esperada de información, $C_{MG}$.
- Criterio de la suma de las utilidades, donde la utilidad se define en términos de los conceptos de sensibilidad y especificidad, $C_{SE}$.
- Criterio del máximo de las utilidades, donde la utilidad se define en términos de los conceptos de sensibilidad y especificidad, $C_{ME}$.
- Criterio condicionado por la probabilidad del concepto, $C_{CC}$.
- Criterio condicionado por la probabilidad de la pregunta, $C_{CP}$.

Discutiremos los resultados de este estudio en la sección 4.

### 3.2.6 Criterios de parada

Como criterio de parada hemos utilizado una combinación de dos criterios: el test termina cuando se alcanza un número máximo de preguntas o bien cuando todos los conceptos han sido evaluados[3]. Para determinar si un concepto ha sido evaluado, fijamos cierto nivel *s*. Si la probabilidad de dominar el concepto es mayor o igual que *1-s* se considera que el concepto se ha diagnosticado como sabido, y, si es menor que *s*, que se ha diagnosticado como no sabido. Todos aquellos conceptos cuya probabilidad esté comprendida entre *s* y 1-*s* se considerarán no diagnosticados. Por tanto, un test puede finalizar aún cuando algunos conceptos no hayan sido diagnosticados si se alcanza el número máximo de preguntas establecido. Este mecanismo evita tests demasiado largos, puesto que dependiendo de la regularidad de las respuestas del alumno puede haber conceptos que no llegaran a diagnosticarse.

## 4. Evaluación

Para la evaluación de nuestra propuesta hemos utilizado alumnos simulados (VanLehn et al., 1998), (Collins et al., 1996). Ello ha permitido realizar el estudio sin necesidad de definir un test para una asignatura concreta y de disponer de un grupo de alumnos a los que aplicar dicho test. Además, el uso de alumnos simulados tiene las siguientes ventajas:

- No parece adecuado probar con personas un método de evaluación sin antes haber comprobado su validez. Desde luego es impensable utilizar un método no comprobado para calificar a unos alumnos en una asignatura. También se podría

[2] Algunos otros criterios han sido también considerados y evaluados, como por ejemplo dividir la suma por el número de conceptos involucrados y criterios basados en la definición tradicional de la *ganancia de información* en *Teoría de la Información*. Sin embargo, los resultados obtenidos con tales criterios eran muy pobres en comparación con los resultados obtenidos con los siete criterios propuestos.

[3] Excepto en el caso del criterio de selección de preguntas aleatorio, en el que la longitud del test es de tamaño fija.

haber pedido la participación de los alumnos en las pruebas sin utilizar sus resultados como base de la calificación, pero en este caso la motivación de los alumnos para contestar adecuadamente a las preguntas no es ni mucho menos comparable a la que tienen cuando contestan a un test de verdad.

- Aún contando con un grupo de alumnos suficientemente motivados, las estimaciones del nivel de conocimiento que se obtuviesen con el sistema iban a ser comparadas con las estimaciones que hiciese el profesor, bien por conocimiento directo o bien utilizando otros métodos tradicionales de evaluación como exámenes, tests, etc. La imposibilidad de comprobar las estimaciones con el *verdadero* estado de conocimiento del alumno hace que la evaluación del método se dificulte, puesto que nunca podremos estar seguros de que sean peores o mejores que las que hace un tutor humano, que nunca pueden considerarse como totalmente objetivas.

Por otro lado, esta técnica también presenta sus inconvenientes (VanLehn et al., 1995). Al menos, dos cuestiones merecen ser mencionadas:

- *Limitaciones en la tecnología de IA*. De hecho, no podemos simular adecuadamente la forma en que los alumnos reales interactúan por medio del lenguaje natural, comunicación no verbal, etc.
- *Limitaciones del modelo.* Muchas de las características de los alumnos reales no se representan en nuestro modelo (motivación, seguridad en sí mismo, etc.) De esta forma, se debería realizar una evaluación empírica del modelo propuesto para asegurar que los resultados experimentales obtenidos se pueden generalizar para los alumnos reales.

El funcionamiento de un alumno simulado es el siguiente: sean $\{C_1, \ldots, C_n\}$ los conceptos de la red de diagnóstico asociada a la asignatura que se pretende evaluar. Dado un valor $k \in [0,1]$ , se define el *alumno simulado tipo k* como un alumno que conoce el $100 \cdot k\%$ de los conceptos $\{C_1, \ldots, C_n\}$, donde el conjunto de los conceptos conocidos se genera aleatoriamente. De esta forma se obtienen alumnos simulados del mismo nivel pero cuyo conjunto de conceptos conocidos es diferente. Una vez generado el alumno simulado, se utiliza la red para calcular las probabilidades de responder correctamente a cada una de las preguntas. Dicha probabilidad se utilizará para simular el comportamiento del alumno en el test de la siguiente forma: supongamos que la probabilidad de responder correctamente a una pregunta P es $p$. Si el test plantea la pregunta $P$, se genera un número aleatorio $a$ en el intervalo [0,1]. Si $p \geq a$, se considera que el alumno ha respondido correctamente a la pregunta, y si $p < a$, que la ha respondido incorrectamente. Tras obtener la respuesta, el algoritmo de diagnóstico la utiliza para actualizar las probabilidades de los conceptos y elige la siguiente pregunta para proponerle al alumno. Como se ve este sencillo mecanismo permitirá comparar el diagnóstico obtenido tras la aplicación del test con el estado real de conocimiento del alumno.

Para las simulaciones hemos utilizado una *red de pruebas* compuesta por catorce conceptos $C_1, \ldots, C_{14}$ y cien preguntas $P_1, \ldots, P_{100}$. A modo de ejemplo, en la Figura 9 aparecen las relaciones entre los conceptos y las veinte primeras preguntas preguntas. Como puede observarse, cada pregunta está relacionada con uno, dos o tres conceptos. De esta forma se modela el hecho de que para responder a una pregunta hay que hacer uso de todos los conceptos que en ella intervienen.

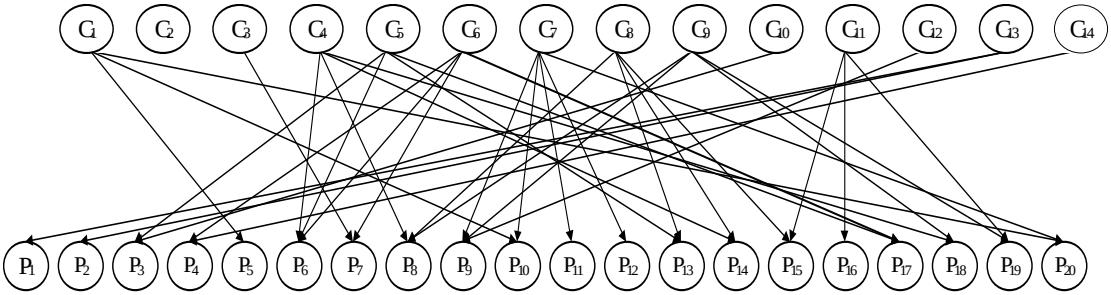



*Figura 9 Relaciones entre conceptos y preguntas de la 1 a la 20 en la red de pruebas.*

Cada pregunta tiene seis posibles respuestas, y por tanto un factor de adivinanza de 1/6. Hay diez niveles de dificultad (de 1 a 10) y diez preguntas en cada nivel. Hay cuatro grupos diferentes de veinticinco preguntas cada uno, cuyos factores de adivinanza s e índices de discriminación a aparecen representados en la Tabla 1.

| | Factor de adivinanza *s* | Índice de discriminación *a* |
|---|---|---|
| Grupo 1 | 0.001 | 2 |
| Grupo 2 | 0.01 | 1.2 |
| Grupo 3 | 0.01 | 0.3 |
| Grupo 4 | 0.2 | 1.2 |

*Tabla 1 Grupos de preguntas según su factor de adivinanza e índice de discriminación.*

Como vemos, hemos numerado los grupos según su calidad psicométrica (a menor índice mayor calidad psicométrica. Por ejemplo, los mejores ítems (menor factor de adivinanza y mayor índice de discriminación) son los del grupo 1.

Se generaron 30 alumnos de seis tipos distintos: alumnos 0.0 (no conocen ningún concepto), alumnos 0.2 (conocen un 20% de los conceptos), alumnos 0.4 (conocen el 40% de los conceptos), alumnos 0.6 (conocen el 60% de los conceptos), alumnos 0.8 (conocen el 80% de los conceptos) y alumnos 1.0 (conocen todos los conceptos), lo cual hace un total de 180 alumnos simulados[4].

## 4.1 Resultados

Empezaremos por analizar los resultados obtenidos al final del test, y después evaluaremos más detalladamente los resultados correspondientes a aquellos criterios de selección que han demostrado un mejor comportamiento.

### 4.1.1 Resultados al final del test

Para estudiar el comportamiento de los diferentes criterios vamos a calcular el número de conceptos que se han dejado sin evaluar al final del test, así como el número de conceptos que han sido correcta o incorrectamente diagnosticados. Un concepto ha sido correctamente diagnosticado si el alumno simulado conocía el concepto y se ha diagnosticado que lo conocía, o bien si no lo conocía y se ha diagnosticado que no lo conocía. Un concepto ha quedado sin evaluar si su probabilidad está comprendida

[4] A la hora de determinar el número de conceptos conocidos el redondeo se hizo siempre tomando la parte entera del número de conceptos conocidos, es decir, si por ejemplo un alumno 0.6 debía conocer 8.4 conceptos de los 14 conceptos de los que consta la red de pruebas, se consideró que conocía 8.

entre los umbrales mínimo y máximo fijados previamente por el profesor (en estas pruebas, 0.3 y 0.7). Los resultados obtenidos se muestran en la Tabla 2.

| | | Basados en información | | | | Condicionados | |
|---|---|---|---|---|---|---|---|
| **Diagnóstico** | **$C_A$** | **$C_{SG}$** | **$C_{MG}$** | **$C_{SE}$** | **$C_{ME}$** | **$C_{CC}$** | **$C_{CP}$** |
| Correcto | 2275 | 2304 | 2262 | 2225 | 2096 | 1965 | 2382 |
| Incorrecto | 77 | 209 | 256 | 124 | 65 | 141 | 58 |
| Sin Evaluar | 168 | 7 | 2 | 171 | 359 | 414 | 80 |
| **Número medio preguntas** | 60 | 16.88 | 15.06 | 55.44 | 51.99 | 58.9 | 55.14 |

*Tabla 2 Resultados al final del test con cada uno de los criterios propuestos.*

Para comparar los resultados quizás sea mejor trabajar con el porcentaje de conceptos sin diagnosticar, evaluados correctamente y evaluados incorrectamente, que aparece en la Tabla 3.

| | | Basados en información | | | | Condicionados | |
|---|---|---|---|---|---|---|---|
| **Diagnóstico** | **$C_A$** | **$C_{SG}$** | **$C_{MG}$** | **$C_{SE}$** | **$C_{ME}$** | **$C_{CC}$** | **$C_{CP}$** |
| Correcto | 90,27% | 91,43% | 89,76% | 88,29% | 83% | 77,98% | 94,53% |
| Incorrecto | 3,06% | 8,29% | 10,15% | 4,92% | 3% | 5,60% | 2,30% |
| Sin Evaluar | 6,67% | 0,28% | 0,01% | 6,79% | 14% | 16,42% | 3,17% |
| **Número medio preguntas** | 60 | 16.88 | 15.06 | 55.44 | 51.99 | 58.9 | 55.14 |

*Tabla 3 Resultados al final del test en porcentajes.*

Lo primero que llama la atención en esta tabla es el buen comportamiento que muestra el criterio aleatorio, que diagnostica bien el 90.28% de los conceptos, mal el 3.06% y deja solamente el 6.67% de los conceptos sin evaluar. Teniendo en cuenta que el test consta de solamente sesenta preguntas, y que los conceptos diagnosticados son catorce, podemos calificar los resultados obtenidos como muy buenos. Sin duda ello se debe a la consistencia teórica del modelo utilizado, ya que como hemos comentado en anteriormente las RBs constituyen un modelo teórico perfectamente fundamentado que funciona muy bien en problemas de clasificación y diagnóstico.

En segundo lugar, nos resultó sorprendente comprobar que sólo uno de los criterios adaptativos propuestos demuestra un rendimiento claramente superior al criterio aleatorio. Pensamos que ello puede deberse a que el modelo permite situaciones *anómalas*[5], es decir, que alumnos sin conocimiento alguno acierten la pregunta, y alumnos que conocen todos los conceptos la fallen. Analicemos el rendimiento de cada grupo de criterios:

- Si nos fijamos en los criterios basados en la utilidad definida como la ganancia de información, cuando estas situaciones anómalas se producen la ganancia de información es en el sentido contrario al que deseamos. De esta forma, al estar seleccionando aquellas preguntas que producen una ganancia máxima, dicha ganancia también es máxima en estos casos anómalos, distorsionando el proceso de

[5] Aunque utilizamos el término *anómalo* para referirnos a este tipo de situaciones, en la práctica son muy habituales, porque especialmente en los exámenes tipo test los alumnos pueden acertar las respuesta correcta o fallar una pregunta que saben, siendo más probable lo primero que lo segundo.

diagnóstico y resultando en un número mayor de conceptos mal evaluados. Cabe destacar sin embargo la gran reducción en el número de preguntas necesarias.

- Referente a los criterios basados en el concepto de utilidad definida en base a los conceptos de sensibilidad y especificidad, cabe destacar que para los alumnos de comportamiento más *predecible*, esto es, para los alumnos 0.0 y 1.0, ambos criterios producen mejores resultados que el aleatorio. Sin embargo, los resultados van empeorando conforme el comportamiento del alumno es más *impredecible* (para alumnos 0.2, 0.4, 0.6 y 0.8) lo que empeora el resultado global.
- En cuanto al criterio condicionado a la probabilidad del concepto es el que peores resultados ha generado, debido a que puede darse el caso de que para la misma pregunta se definan las utilidades $U'(P,C)$ como la sensibilidad para aquellos conceptos cuya $P(C)$ fuese mayor que 0.5 y como la especificidad para aquellos en los que $P(C)$ fuese menor que 0.5, con lo cual no parece tener mucho sentido coger como utilidad de la pregunta $U(P)$ el máximo de estas utilidades $U'$.
- Por último el criterio de condicionar la definición de utilidad según la probabilidad de la pregunta demostró el mejor comportamiento, consiguiendo los diagnósticos más precisos y reduciendo el número de preguntas. La distribución del número de preguntas se muestra en la Figura 10, donde en el eje horizontal se han agrupado el número de preguntas necesarias en intervalos de tamaño 5[6] y en el vertical se representa el número de alumnos:

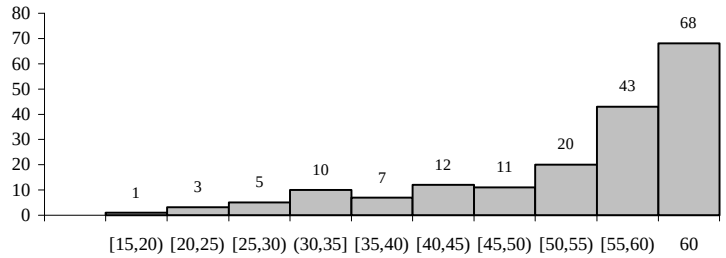



*Figura 10 Distribución del número de preguntas con el criterio condicionado a la pregunta.*

La medida de preguntas necesarias para evaluar todos los conceptos con el test adaptativo es de 51.98, con una desviación estándar de 10.53. Es cierto que la reducción en el número de preguntas no es demasiado significativa, lo cual puede deberse en parte al buen funcionamiento del modelo bayesiano como algoritmo de diagnóstico, pero si la unimos a la mayor precisión obtenida y a la simplicidad del criterio merece la pena su aplicación. En la siguiente sección haremos un análisis comparativo más detallado de los dos criterios que han demostrado un mejor comportamiento, es decir, el aleatorio y el condicionado a la pregunta, que en adelante llamaremos *adaptativo*[7].

### 4.1.2 Comparativa entre los criterios aleatorio y adaptativo

Para realizar este análisis comparativo detallado, vamos a estudiar la evolución del test analizando los resultados obtenidos tras 15, 30, 40, 50 preguntas y al final del test y los resultados por tipo de alumno.

---

[6] Excepto el caso de necesitar las 60 preguntas.

[7] En el análisis no hemos incluido los criterios basados en la ganancia de información puesto que el tanto por ciento de conceptos mal diagnosticados es alto (alrededor del 10%). Sin embargo la gran reducción alcanzada en la longitud del test puede hacer que en algunos casos merezca la pena su aplicación.

#### 4.1.2.1 Evolución del test

Para estudiar la evolución de los test aleatorio y adaptativo, se muestra en la Tabla 4 el número de conceptos que se dejan sin evaluar, son diagnosticados correctamente e incorrectamente tras un número fijado de preguntas (15, 30, 40, 50) y al final del test. Dichos datos aparecen representados en las Figuras 11 a 13.

| | **Correctas** | | **Incorrectas** | | **Sin Evaluar** | |
|---|---|---|---|---|---|---|
| | **Aleatorio** | **Adaptativo** | **Aleatorio** | **Adaptativo** | **Aleatorio** | **Adaptativo** |
| **15** | 857 | 922 | 153 | 89 | 1510 | 1509 |
| **30** | 1514 | 1648 | 134 | 73 | 872 | 799 |
| **40** | 1878 | 1971 | 117 | 69 | 525 | 480 |
| **50** | 2100 | 2247 | 97 | 60 | 323 | 213 |
| **Final** | 2275 | 2382 | 77 | 58 | 168 | 80 |

*Tabla 4 Evolución de los resultados del test.*

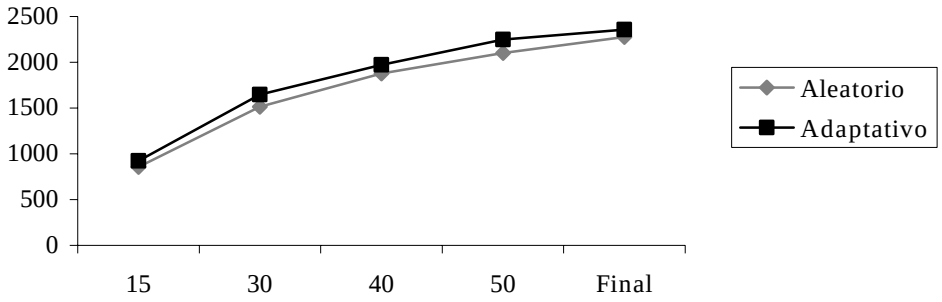



*Figura 11 Conceptos diagnosticados correctamente según el número de preguntas realizadas.*

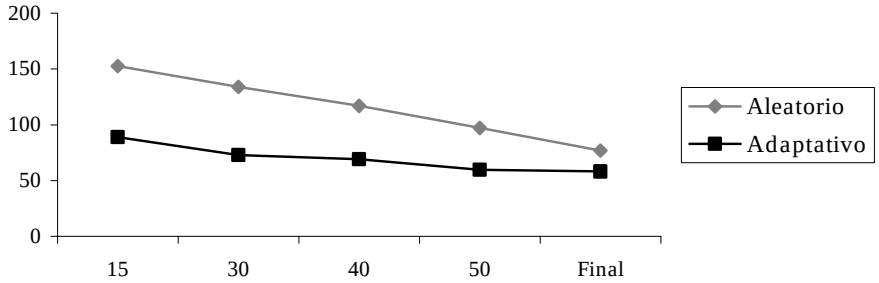



*Figura 12 Conceptos diagnosticados incorrectamente según el número de preguntas realizadas.*

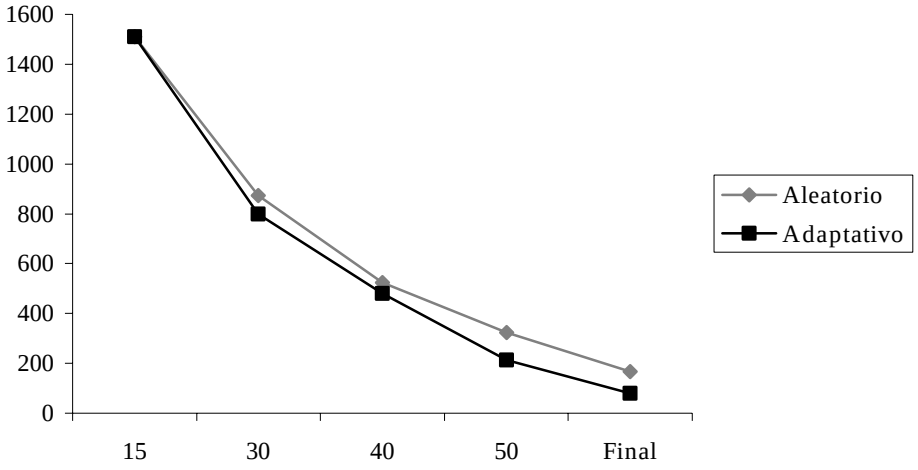



*Figura 13 Conceptos sin evaluar según el número de preguntas realizadas.*

Quizás se podría destacar que en las figuras 11 a 13 las escalas son distintas, y en especial que el rango en la figura 12 el rango es mucho menor. Las gráficas muestran que el comportamiento del test adaptativo es siempre mejor que el del test aleatorio, y por tanto generará siempre test más cortos de resultados más precisos.

A continuación vamos a ver la tendencia que muestra el algoritmo de diagnóstico, es decir, vamos a analizar si tiende a evaluar por exceso a los alumnos (diagnosticar como sabidos conceptos que no se conocen) o a evaluarlos por defecto (diagnosticar como no sabidos conceptos que el alumno se sabe). Para ello vamos a volver a los resultados finales obtenidos con los dos métodos, que se pueden ver en la Tabla 2 y se representan en porcentajes en la Figura 14.

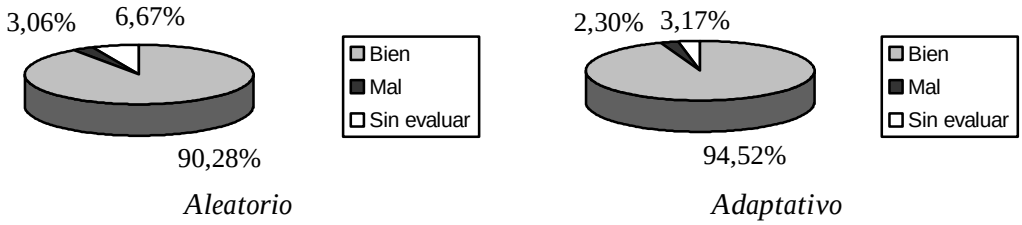
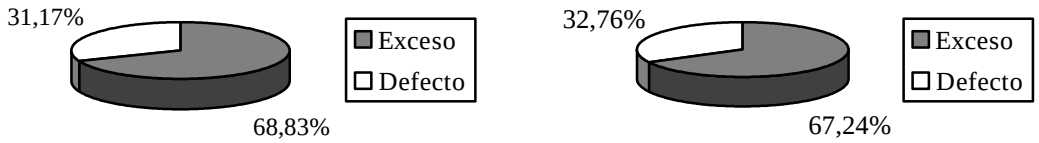



*Figura 14 Resultados al finalizar el test.*

Vamos ahora a desglosar los conceptos mal diagnosticados en conceptos mal evaluados por exceso y conceptos mal evaluados por defecto. Los resultados aparecen en la Tabla 5 y se representan en porcentajes en la Figura 15.

| Diagnóstico | Aleatorio | Adaptativo |
|---|---|---|
| Incorrecto (exceso) | 53 | 39 |
| Incorrecto (defecto) | 24 | 19 |
| Incorrecto (Total) | 77 | 58 |

*Tabla 5 Conceptos mal diagnosticados por exceso/por defecto al final del test.*

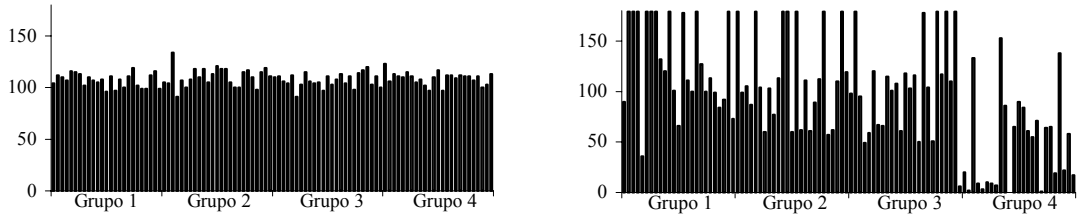



*Aleatorio* *Adaptativo*

*Figura 15 Tendencia a estimar por exceso/por defecto de cada test para los conceptos mal evaluados.*

Observamos que ambos métodos tienden a estimar por exceso, pero pensamos que esto no es una característica del método bayesiano de diagnóstico sino del banco de ítems utilizado. Un alumno que no conoce los conceptos necesarios para acertar una pregunta tiene una probabilidad de acertarla de 0.16667, mientras que un alumno que tiene los conocimientos necesarios para una pregunta tiene una probabilidad media de fallar de 0.05715[8]. De esta forma, la tendencia del test vendrá determinada por el banco de preguntas (en este caso se tiende a sobreestimar a los alumnos, ya que adivinar la respuesta correcta es mucho más fácil que fallar cuando se conocen todos los conceptos relacionados).

Es interesante también comprobar cuántas veces se ha utilizado cada pregunta en cada tipo de test. Estos datos se reflejan en la Figura 16.




***Criterio aleatorio*** ***Criterio adaptativo***

*Figura 16 Distribución del número de veces que se utilizó cada pregunta.*

Observamos que el test aleatorio tiende a escoger las preguntas uniformemente, de forma que cada pregunta ha sido elegida un mínimo de 91 y un máximo de 134 veces. Sin embargo, el test adaptativo usa las preguntas de forma diferente, puesto que mientras que hay preguntas que apenas se usan (por ejemplo, $P_{54}$ no se usa ninguna vez, $P_{41}$ tan sólo una vez) hay otras que llegan a usarse hasta 179 veces (como por ejemplo $P_1$, $P_2$ y $P_{70}$). Se observa cierta tendencia a usar más aquellas preguntas con alto índice de discriminación y bajo factor de descuido, y a usar menos aquellas preguntas con bajo índice de discriminación y alto factor de descuido, lo cual parece lógico dado que desde el punto de vista psicométrico la calidad de las primeras es superior. El número medio de veces que cada tipo de pregunta aparece en el test viene dado en la Tabla 6, donde podemos ver que el uso de las preguntas es mayor cuanto mayor es su calidad psicométrica.

[8] Como ya comentamos en la nota al pie 5, esta es la situación usual en un examen tipo test.

| | Aleatorio | Adaptativo |
|---|---|---|
| Grupo 1 (*s*=0.001, *a*=2) | 107.12 | 129.32 |
| Grupo 2 (*s*=0.01, *a*=2) | 110.4 | 113.76 |
| Grupo 3 (*s*=0.01, *a*=0.3) | 107.08 | 104.32 |
| Grupo 4 (*s*=0.2, *a*=1.2) | 109.04 | 49.68 |

*Tabla 6 Uso medio de los items de cada grupo.*

#### 4.1.2.2 Resultados por tipo de alumno

A continuación vamos a analizar los resultados por tipo de alumno. Como ya hemos comentado con anterioridad, hemos considerado seis tipos de alumnos diferentes. En primer lugar mostramos en la Tabla 7 el número medio de preguntas que se necesitaron para cada tipo de alumno en el test adaptativo:

| Tipo alumno | Nº medio preguntas | Tipo alumno | Nº medio preguntas |
|---|---|---|---|
| Alumno 0.0 | 54,23 | Alumno 0.6 | 51,8 |
| Alumno 0.2 | 52,67 | Alumno 0.8 | 54,76 |
| Alumno 0.4 | 41,73 | Alumno 1.0 | 56,73 |

*Tabla 7 Número medio de preguntas por tipo de alumno.*

Los resultados por tipo de alumno aparecen en la Tabla 8.

| Tipo Alumno | Diagnóstico | Aleatorio | Adaptativo |
|---|---|---|---|
| Alumno 0.0 | Correcto | 371 | 395 |
| | Incorrecto | 17 | 4 |
| | Sin Evaluar | 32 | 21 |
| Alumno 0.2 | Correcto | 366 | 385 |
| | Incorrecto | 17 | 14 |
| | Sin Evaluar | 37 | 21 |
| Alumno 0.4 | Correcto | 357 | 387 |
| | Incorrecto | 24 | 20 |
| | Sin Evaluar | 59 | 13 |
| Alumno 0.6 | Correcto | 376 | 400 |
| | Incorrecto | 9 | 10 |
| | Sin Evaluar | 35 | 10 |
| Alumno 0.8 | Correcto | 390 | 402 |
| | Incorrecto | 10 | 8 |
| | Sin Evaluar | 20 | 10 |
| Alumno 1.0 | Correcto | 415 | 413 |
| | Incorrecto | 0 | 2 |
| | Sin Evaluar | 5 | 5 |

*Tabla 8 Resultados por tipo de alumno.*

Vemos que para todos los tipos de alumno (excepto para el tipo 1.0) el resultado del test adaptativo es significativamente mejor que el del test aleatorio, puesto que se dejan menos conceptos sin evaluar y se diagnostican menos conceptos mal, lo cual resulta en un mayor número de conceptos correctamente diagnosticados. La mejora más significativa se produce para los alumnos de tipo 0.4, con un 11% más de conceptos

evaluados correctamente y con los test más cortos de todos (una media de 41,73 preguntas). En el único caso en que los resultados del test aleatorio parecen ser algo mejores es en el caso de los alumnos tipo 1.0, si bien esta mejora no es significativa dada la componente aleatoria inherente en el proceso y que prácticamente en ambos casos todos los conceptos son diagnosticados correctamente.

## 5. Trabajos relacionados

Las RBs se han utilizado con éxito para definir los modelos del alumno de varios sistemas. Por el contrario, es difícil encontrar aplicaciones de los TAI al problema del modelado del alumno, a pesar de la gran mejora en precisión y rapidez que pueden aportar al proceso de diagnóstico. A continuación revisaremos brevemente aquellos trabajos que tienen una mayor relación con el nuestro, muchos de los cuales han sido ya discutidos en secciones anteriores.

- HYDRIVE (Mislevy & Gitomer, 1996) modela la capacidad del alumno para diagnosticar averías en sistemas hidráulicos para aviación. El conocimiento del alumno se caracteriza en términos de constructos generales (variables dimensionales) y se usa una RB para actualizar estas variables, utilizando las acciones del alumno como evidencias.

- ANDES (Conati et al., 1997) es un STI para la enseñanza de la Física mediante resolución de problemas tutorizada. El sistema se construyó sobre sistemas anteriores: OLAE (Martin & VanLehn, 1995b) y POLA (Conati & VanLehn, 1996a), y utiliza RBs para llevar a cabo diferentes procesos: reconocimiento del plan, evaluación a largo plazo, predicción de las acciones del alumno durante la resolución de los problemas. En (VanLehn et al., 1998), se usaron algoritmos de diagnóstico basados en tests para encontrar las probabilidades a priori que se necesitaban para el sistema ANDES. Este trabajo ya ha sido comparado con el nuestro en la sección 3.1.

- Quizás el trabajo que guarda más relación con el nuestro sea el presentado en (Collins et al., 1996), donde se aplican las redes bayesianas junto con jerarquías de granularidad, y se utilizan preguntas tipo test BNs como evidencias para determinar si el alumno domina o no los objetivos de aprendizaje definidos. Se comparan tres estructuras diferentes para la RB, en términos de el número de parámetros requeridos y la longitud y cobertura del test. Sin embargo, hemos ya mencionado que el método de selección de preguntas propuesto parecía poco efectivo. También es interesante notar que el rendimiento del algoritmo de diagnóstico fue evaluado en términos de la longitud y cobertura del test, pero no de la precisión del diagnóstico. Además, la evaluación se realizó utilizando solamente alumnos buenos (0.8 y 1) y malos (0 y 0.2), lo cual puede distorsionar los resultados ya que obviamente los alumnos intermedios son los de comportamiento más impredecible y por tanto más difíciles de diagnosticar.

## 6. Conclusiones y trabajos futuros

En este trabajo hemos presentado y evaluado un algoritmo de diagnóstico para modelado del alumno basado en test adaptativos bayesianos. El algoritmo se lleva a

cabo sobre una red bayesiana en la que los nodos tienen una semántica bien definida y los enlaces describen de una forma precisa las relaciones entre ellos. El modelo propuesto permite una reducción sustancial en el número de parámetros que se deben especificar, ya que las probabilidades condicionadas se calculan de forma automática a partir de los factores de descuido, adivinanza, dificultad y discriminación.

La validez del enfoque propuesto se ha probado intensivamente utilizando alumnos simulados. Los resultados obtenidos son muy prometedores, ya que muestran que el modelo bayesiano integrado produce estimaciones muy precisas del estado cognitivo del alumno. Sin embargo, los alumnos simulados son sólo instancias del modelo presentado, y por tanto hay muchos otros factores que influyen en el comportamiento del alumno que no se representan. Por tanto, para asegurar la validez de los resultados aquí presentados en el mundo real sería necesario realizar estudios con alumnos reales. En particular, una de las limitaciones mayores del modelo es que supone que el conocimiento del alumno no cambia durante el test, lo cual puede ser una hipótesis válida para este trabajo pero puede ser considerada poco realista.

Aún cuando los resultados obtenidos son muy satisfactorios (90.27% de conceptos diagnosticados correctamente en un entorno en que se permiten fallos y adivinanzas), ha sido posible mejorarlos un poco más mediante la introducción de criterios adaptativos de selección de preguntas (alcanzando un 94.52% de conceptos correctamente diagnosticados).

En cuanto a trabajos futuros, identificamos varias líneas que agrupamos en dos categorías principales: a) *mejoras en el modelo estructural*, para permitir la inclusiónd e otros tipos de relaciones (pre-requisitos, etc.) como de otras fuentes de evidencia, como qué episodios instructores ha superado el alumno, opiniones del profesor, etc. y b) *aplicaciones del modelo,* como puede ser el desarrollo de un *Sistema Inteligente Bayesiano de Evaluación mediante Tests*, accesible a través de la web, que permita que profesores sin conocimientos de programación ni de redes bayesianas definan sus propios test adapativos, que después estarán disponibles para que sus alumnos los realicen también a través de la web.

## Referencias


1. HUGIN [Web Page]. URL http://www.hugin.com.
2. NETICA [Web Page]. URL http://www.norsys.com/netica.html.
3. Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher, 13, 4-15.
4. Castillo, E., Gutiérrez, J. M., & Hadi, A. (1997). Expert Systems and Probabilistic Network Models. New York: Springer Verlag.
5. Charniak, E. (1991). Bayesian Networks Without Tears. AI Magazine, 12(4), 50-63.
6. Collins, J. A., Greer, J. E., & Huang, S. H. (1996). Adaptive Assessment Using Granularity Hierarchies and Bayesian Nets. In Lecture Notes in Computer Science: Vol. 1086. Proceedings of 3rd International Conference ITS'96 (pp. 569-577). Berlin Heidelberg: Springer Verlag.
7. Conati, C., Gertner, A., VanLehn, K., & Druzdzel, M. (1997). On-line Student Modelling for Coached Problem Solving using Bayesian Networks. Proceedings of the 6th International Conference on User Modelling UM'97 (pp. 231-242). Vienna,

New York: Springer Verlag.

8. Conati, C., & VanLehn, K. (1996a). POLA: A Student Modeling Framework for Probabilistic On-line Assessment of Problem Solving Performance. Proceedings of the 5th International Conference on User Modeling UM'96 (pp. 75-82). User Modeling Inc.
9. Martin, J., & VanLehn, K. (1995b). Student Assessment using Bayesian Nets. International Journal of Human-Computer Studies, 42, 575-591.
10. Millán, E. (2000). Bayesian System for Student Modeling. Unpublished doctoral dissertation, Universidad de Málaga, Dpto. de Lenguajes y Ciencias de la Computación.
11. Millán, E., & Pérez-de-la-Cruz, J. L. (2002). A Bayesian Diagnostic Algorithm for Student Modeling. User Modeling and User-Adapted Interaction, (Special Issue on Empirical Evaluation of User Models).
12. Millán, E., Pérez-de-la-Cruz, J. L., & Triguero F. (1998). Using Bayesian networks to build and handle the student model in exercise based domains. In Lecture Notes in Computer Science: Vol. 1452. Intelligent Tutoring Systems. Proceedings of 4th International Conference ITS'98 (p. 612). Berlin Heidelberg: Springer Verlag.
13. Mislevy, R., & Gitomer, D. H. (1996). The Role of Probability-Based Inference in an Intelligent Tutoring System. User Modeling and User-Adapted Interaction, 5, 253-282.
14. Neapolitan, R. (1990). Probabilistic Reasoning in Expert Systems: Theory and Algorithms. New York: John Wiley & Sons.
15. Pearl, J. (1988). Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann Publishers, Inc.
16. VanLehn, K., Niu, Z., Siler, S., & Gertner, A. S. (1998). Student Modeling from Conventional Test Data: A Bayesian Approach Without Priors. In Lecture Notes in Computer Science: Vol. 1452. Intelligent Tutoring Systems. Proceedings of 4th International Conference ITS'98 (pp. 434-443). Berlin Heidelberg: Springer Verlag.
17. VanLehn, K., Ohlsson, S., & Nason, R. (1995). Applications of Simulated Students: An Exploration. Journal of Artificial Intelligence and Education, 5(2), 135-175.
18. Weiss, D., & Kingsbury, G. (1984). Application of Computerized Adaptive Testing to Educational Problems. Journal of Educational Measurement, 12(361-375).