Actas de las XXIII JENUI. Cáceres, 5 al 7 de julio de 2017
ISBN: 978-84-697-4077-4
Páginas: 217-220

Diseño de una aplicación de apoyo a la dirección de los centros universitarios basada en redes bayesianas

Carmen Lacave, Félix O. García, José A. Cruz-Lemus, Ramón Hervás, Ana I. Molina, Eduardo Fernández-Medina
Dpto. de Tecnologías y Sistemas de la Información, UCLM
{carmen.lacave, felix.garcia, joseantonio.cruz, ramon.hervas, anaisabel.molina, eduardo.fdezmedina}@uclm.es

Resumen

En este trabajo se plantea el diseño de una aplicación de apoyo a la toma de decisiones de los equipos directivos en los centros universitarios, orientadas a mejorar la calidad de las diferentes actuaciones que han de realizarse. Para ello, la aplicación se basará en el análisis de la información almacenada en las bases de datos de la Universidad mediante el uso de redes bayesianas, que permiten obtener modelos predictivos en función de una serie de datos.

Abstract

This paper proposes the design of an application to support the decision-making of management teams in university centers, aimed at improving the quality of the different actions to be carried out. To do this, the application will be based on the analysis of the information stored in the University databases by using Bayesian networks, which allow to obtain predictive models based on a series of data.

Palabras clave

Perfiles de estudiante, redes bayesianas, predicción del rendimiento, ayuda a la toma de decisiones.

1. Introducción

En la actualidad, uno de los factores estratégicos de la dirección de los centros universitarios es el de garantizar la calidad de las decisiones que aseguren la mejora continua de las titulaciones que en ellos se imparten. Así pues, es importante ser cuidadoso con la forma en la que dichas decisiones se toman. Algunos equipos directivos se apoyan en herramientas que les ayudan a visualizar y a analizar la información, como el sistema SID[1] usado en la UCLM, que está basado en tecnologías de *Data Warehouse*.

Sin embargo, la mayoría de estas herramientas o bien se limitan a proporcionar una gran cantidad de datos estadísticos meramente descriptivos que son difíciles de interpretar, o bien si recomiendan alguna decisión no son capaces de justificar el proceso de razonamiento seguido para ello.

En la última edición de JENUI se presentó un trabajo que describía el uso de redes bayesianas[2] para la identificación del perfil del estudiante que abandona los estudios de Informática en la Universidad de Castilla-La Mancha (UCLM) a partir de una base de datos proporcionada por la propia universidad y que contenía los datos de matriculación de los alumnos [6]. En dicho trabajo se trató de replicar el desarrollado previamente con éxito con datos del alumnado de la Universidad de Almería para la obtención de perfiles educativos [9] y para el análisis del rendimiento universitario [4]. Para ello, en [6] se utiliza una base de datos como entrada a un algoritmo de aprendizaje automático [3,5,8] que proporciona la red bayesiana que representa gráficamente las relaciones de dependencia e independencia probabilística entre el conjunto de variables que definen los campos de dicha base de datos. Sobre la red obtenida se aplica posteriormente un algoritmo de abducción (total o parcial) [10], que proporciona la configuración de variables de la red de máxima probabilidad y que es lo que permite definir el perfil del alumno en función de los indicadores representados en la base de datos.

Si bien los resultados de dicho trabajo no fueron concluyentes debido a las deficiencias de la base de datos utilizada [6], éste sirvió para poner de manifiesto el interés de la metodología empleada, la cual permite

[1] http://www.socinfo.es/contenido/seminarios/1417clamancha3/InteligenciaUCLM.pdf

[2] Una red bayesiana se define mediante un grafo dirigido acíclico, en el que cada nodo representa una variable aleatoria y los enlaces representan las dependencias probabilísticas entre dichas variables. Además, cada nodo lleva asociada la distribución de probabilidad del mismo condicionada en sus padres, la cual definirá, junto con la estructura del grafo, la relación de dependencia probabilística existente entre ellos.

extraer conocimiento relevante e interesante de los datos más allá del que se ve a simple vista.

Teniendo en cuenta el potencial de la propuesta anterior, la dirección de la Escuela Superior de Informática de Ciudad Real (ESI), junto con un equipo de profesores del centro, se ha planteado utilizarla como modelo de soporte a la toma de decisiones orientadas a mejorar las diferentes actuaciones (docentes, de calidad, de captación de alumnos, etc.) que desde hace tiempo se vienen desarrollando. De este modo se dispondría de un sistema inédito hasta el momento que permitiría hacer estudios sistematizados para mejorar el proceso enseñanza-aprendizaje a partir de la obtención de perfiles de alumnado y de modelos predictivos.

Además, aunque las redes bayesianas representan de forma intuitiva las relaciones entre un conjunto de variables y existen programas con los que se hace relativamente sencillo inferir nuevo conocimiento y obtener los resultados deseados, se ha planteado el desarrollo de una aplicación que facilite aún más la interacción del usuario con este tipo de modelos probabilísticos de manera que le resulten totalmente transparentes. Esta aplicación tendrá como objetivo dar respuestas específicas a las necesidades de información relacionadas con la obtención del perfil del alumnado de la ESI en base a distintos indicadores académicos, sociales y/o familiares; permitiendo predecir su rendimiento académico. El público objetivo de esta herramienta será el equipo directivo de la ESI. Además, el diseño de la aplicación incluirá la posibilidad de mejorar los resultados proporcionados, permitiendo incorporar los datos de los alumnos que se vayan obteniendo cada año. Su uso permitirá replicar el trabajo de análisis y apoyo a la toma de decisiones en otros centros de nuestra Universidad o en otras universidades.

En la próxima sección se resumen los objetivos principales planteados, para describir posteriormente el método de trabajo, que incluye tanto la obtención del modelo como la arquitectura de la aplicación. Finalmente se presentan las conclusiones de este trabajo.

2. Objetivo del trabajo

En líneas generales, el principal objetivo de este trabajo se puede enunciar así:

OG: Proporcionar soporte al equipo de dirección de la ESI en la toma de decisiones relacionadas con las actuaciones orientadas a mejorar el rendimiento de los alumnos en el Grado en Ingeniería Informática.

Este objetivo se puede concretar en dos más específicos:

OE1: Analizar los datos de matriculación y de rendimiento académico de los alumnos matriculados en el Grado en Ingeniería Informática mediante redes bayesianas.

OE2: Diseñar una aplicación que sirva de interfaz entre el equipo de dirección y los modelos de análisis de datos.

3. Método de trabajo

El procedimiento a seguir para alcanzar el objetivo general planteado consta de dos fases diferenciadas, encaminadas cada una de ellas a la consecución de los dos objetivos específicos enunciados, y que se están desarrollando de forma paralela:

- Con relación a la consecución del objetivo **OE1**, una fase del trabajo está relacionada con la generación de las redes bayesianas a partir de la base de datos proporcionada y la obtención de los resultados, que denominaremos el *Sistema de Información o de Conocimiento*. Puesto que el objetivo es dar soporte a distintas decisiones que involucran en cada caso a distintos conjuntos de variables, este Sistema de Conocimiento estará integrado por varias redes bayesianas independientes, proporcionando cada una de ellas información relativa al conjunto de variables implicadas en cada caso. Por ejemplo, habrá una red bayesiana para proporcionar información sobre el perfil socioeconómico del alumno que abandona los estudios de Informática, que será distinta e independiente de la red bayesiana que predice la evolución del rendimiento de un alumno en determinadas asignaturas en función de las notas obtenidas en otras asignaturas hasta el momento.
- Por otra parte, la fase relacionada con la consecución del objetivo **OE2** comprende todos los aspectos implicados en el *Desarrollo de una Aplicación Software*, incluyendo el análisis y diseño de los elementos de interacción y comunicación del Sistema de Conocimiento con el usuario final, que será el equipo directivo de la ESI.

3.1. Sistema de Conocimiento

En esta sección se describen las etapas a seguir para el desarrollo del Sistema de Conocimiento:

1. *Obtención de los datos de una base de datos.* Para paliar las deficiencias de la base de datos utilizada en el trabajo previo [6] la Oficina de Planificación y Calidad (OPyC) nos ha proporcionado la información personal y académica disponible en UXXI-Académico[3] de los 1133 alumnos matriculados en el Grado en Ingeniería Informática en la ESI desde el curso 2010/11, en el que la titulación comenzó a impartirse. Para garantizar la

[3] http://www.ocu.es/productos/universitas-xxi-academico/

confidencialidad y privacidad de los datos, el acceso a los mismos no permite la posibilidad de identificar a la persona en cuestión. La información se ha entregado en una base de datos Access junto con un documento explicativo, en el que se explica de forma resumida el contenido de las tablas y las relaciones entre ellas, las cuales dan lugar a una estructura en estrella. En este tipo de relaciones existe una tabla central, denominada **tabla de hechos**, unida a varias tablas periféricas, denominadas **dimensiones**. Las dimensiones representan el detalle de los datos y la tabla de hechos las relaciones entre dichas dimensiones, junto con información adicional que permite obtener indicadores o métricas (número de alumnos, notas medias, etc.). La base de datos proporcionada contiene 15 tablas para representar dimensiones y 7 tablas de hechos.

2. *Preparación de las bases de datos.* A partir de las tablas proporcionadas y haciendo uso del sistema administrador de base de datos Microsoft SQL Server, versión Developer, se obtienen las distintas bases de datos en formato *.csv* necesarias en la fase de aprendizaje de las distintas redes. Además, para evitar los problemas de heterogeneidad de las bases de datos, producidos fundamentalmente por una excesiva granularidad de las variables, habrá que simplificar las variables que puedan tomar demasiados valores, como las relacionadas con la edad, el municipio del domicilio familiar, los estudios y las profesiones paternas, así como las relacionadas con el total de asignaturas (matriculadas, convalidadas, etc.).
3. Aprendizaje y validación de las redes bayesianas obtenidas, haciendo uso de los programas OpenMarkov[4], Elvira[5] y Hugin[6], pues cada uno de ellos proporciona funcionalidades distintas, útiles para la obtención e interpretación de los resultados. Así, por ejemplo, OpenMarkov es gratuito y de código abierto, incluye la opción de ejecutar diversos algoritmos de aprendizaje paso a paso, permite la edición de la red bayesiana en cualquier momento, ofrece opciones de pre-procesamiento de datos y consta de una interfaz que facilita todo el proceso [2]. Elvira proporciona herramientas de explicación que facilitan identificar la relevancia que tiene cada variable sobre otras, así como el tipo de influencia que ejerce [7]. Y Hugin dispone de distintas opciones de análisis muy útiles para la interpretación tanto del modelo como de los resultados [1].
4. *Obtención e interpretación de los resultados* mediante la aplicación de distintos tipos de razonamiento en función de la evidencia disponible. Para ello, dependiendo de la pregunta a responder, se aplicarán distintos procedimientos. Así, si lo que se quiere es obtener el perfil del estudiante que abandona el grado en función de todas las variables que forman parte de la red, se tendrá que realizar un proceso de abducción total para obtener la configuración de máxima probabilidad que incluye a todas las variables. Si se desea predecir el rendimiento académico de un alumno en una asignatura en función de determinados indicadores académicos, lo que habrá que aplicar es un proceso de propagación de la evidencia disponible [10]. Este tipo de algoritmos se basan en el teorema de Bayes y proporcionan la probabilidad "a posteriori" de las variables de interés. Se denomina *evidencia* al conjunto de hallazgos que determinan, con certeza, el valor de las variables observadas. Por ejemplo, si conocemos de un alumno que es varón, de 20 años y ha accedido a la titulación con una nota de 6,7.

3.2. Desarrollo de la Aplicación

En la Figura 1 se muestra la arquitectura de la herramienta propuesta, la cual consta de tres módulos principales:

- **Módulo de Conocimiento**, que contiene las distintas redes que se obtienen a partir de los datos proporcionados por la UCLM, así como los algoritmos de aprendizaje y de propagación necesarios para obtener la información deseada, y los que permiten actualizar las redes creadas a medida que se van incorporando nuevos datos a la base de datos.
- **Cuadro de Mando**, que es el módulo encargado de facilitar la interacción con el usuario mediante la representación de los indicadores de relevancia

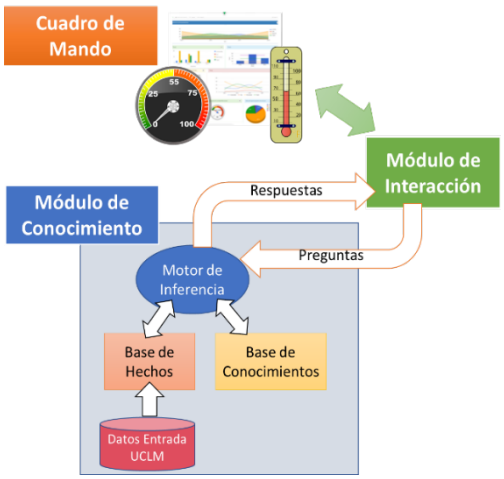

Figura 1. Arquitectura propuesta para la aplicación

[4] http://openmarkov.org
[5] http://leo.ugr.es/elvira
[6] http://www.hugin.com

que faciliten la toma de decisiones. Por ejemplo, un indicador de relevancia podría ser el riesgo de abandono de un alumno (expresado como un porcentaje), y que puede calcularse a partir de sus datos de rendimiento académico, demográficos y sociales. De esta forma, en base a la información que proporcionen dichos indicadores, la dirección del centro puede tomar decisiones estratégicas de actuación para la prevención de ciertos abandonos. Para garantizar la usabilidad de la aplicación desarrollada se hará uso de distintos recursos visuales, como puede ser el uso de códigos de colores o metáforas de visualización de los indicadores (por ejemplo, empleando la metáfora del termómetro), los cuales permitirán que, de una forma rápida, el tomador de decisiones pueda ver qué alumnos están en riesgo de abandono y poder aplicar las políticas del centro pertinentes en dichos casos.

- **Módulo de Interacción,** que es el encargado de facilitar la comunicación entre los módulos anteriores. Para ello, se encargará de recoger información que los usuarios de la aplicación introduzcan en el Cuadro de Mando y lanzar los algoritmos necesarios en el Módulo de Conocimiento para obtener la información deseada. Asimismo, tendrá que traducir dicha información para proporcionarla al usuario de la forma más entendible posible. Además, debe sincronizar en todo momento el estado de la red bayesiana y del Cuadro de Mando. En particular, se encargará de realizar las preguntas al motor de inferencias de la red, de modo que se pueda generar el contenido a mostrar en el Cuadro de Mando.

4. Conclusiones

Este documento recoge los principales objetivos del trabajo que se pretende desarrollar por parte de la dirección de la Escuela Superior de Informática de Ciudad Real (ESI), junto con un grupo de docentes de la misma, orientado a facilitar la toma de decisiones orientadas a mejorar el desempeño de los alumnos del Grado en Ingeniería Informática de la ESI. Para ello, se propone el uso de técnicas de minería de datos basadas en redes bayesianas que, a partir de los datos de matriculación (académicos y socio-familiares) del alumnado, junto con los de su rendimiento académico, sea capaz de proporcionar información sobre distintos perfiles así como de predecir su evolución. Además, para hacer más sencilla la interacción con este tipo de sistemas, se plantea el desarrollo de una aplicación que haga de interfaz entre el conocimiento que proporcionan los modelos probabilísticos obtenidos y el equipo directivo que debe gestionar dicha información.

Como se ha puesto de manifiesto, el trabajo se encuentra en una fase inicial, en el que se han definido los objetivos y la metodología a seguir, así como la arquitectura de la aplicación a desarrollar. En pocos meses esperamos tener construidos los modelos probabilísticos a partir de los que obtener los primeros resultados, así como el primer prototipo de la aplicación.

Referencias

[1] S. Andersen, K. Olesen y F. Jensen, «HUGIN-A shell for building Bayesian belief universes for expert systems,» de *Readings in uncertain reasoning*, G. S. a. J. P. (Eds.), Ed., San Francisco, CA: Morgan Kauffmann Publishers Inc., 1990, pp. 332-337.

[2] I. Bermejo, J. Oliva, F. J. Díez y M. Arias, «Interactive learning of Bayesian networks using OpenMarkov».

[3] G. F. Cooper y E. Herskovits, «A Bayesian method for the induction of probabilistic networks from data,» *Machine Learning,* vol. 9, nº 4, pp. 309-347, 1992.

[4] A. Fernández, M. Morales, C. Rodríguez y A. Salmerón, «A system for relevance analysis of performance indicators in higher education using Bayesian networks,» *Knowledge and Information Systems,* vol. 27, nº 3, pp. 327-344, 2011.

[5] N. Friedman, D. Geiger y M. Goldszmidt, «Bayesian network classifiers,» *Machine Learning,* vol. 29, pp. 131-163, 1997.

[6] C. Lacave, A. I. Molina, M. A. Redondo y M. Ortega, «Redes bayesianas para identificar perfiles de estudiante. Aplicación al estudio del abandono de las titulaciones de Informática en la Universidad de Castilla-La Mancha,» *ReVisión,* vol. 9, nº 3, pp. 29-37, 2016.

[7] C. Lacave, M. Luque y F. J. Díez, «Explanation of Bayesian networks and influence diagrams,» *IEEE Systems, Man and Cybernetics. Part B.,* vol. 37, pp. 952-965, 2007.

[8] M. Minsky, «Steps towards artificial intelligence,» *Computers and Thoughts,* pp. 406-450, 1963.

[9] M. Morales y A. Salmerón, «Análisis del alumnado de la Universidad de Almería mediante redes bayesianas,» de *Actas del 27 Congreso Nacional de Estadística e Investigación Operativa*, Lérida, 2003.

[10] J. Pearl, Probabillistic Reasoning in Intelligent Systems, Morgan Kaufmann Publishers Inc., 1988