Knowl Inf Syst (2011) 27:327–344
DOI 10.1007/s10115-010-0297-9

REGULAR PAPER

A system for relevance analysis of performance indicators in higher education using Bayesian networks

Antonio Fernández · María Morales · Carmelo Rodríguez · Antonio Salmerón

Received: 10 March 2009 / Revised: 8 February 2010 / Accepted: 27 March 2010 /
Published online: 1 May 2010

Abstract In this paper, we propose a methodology for relevance analysis of performance indicators in higher education based on the use of Bayesian networks. These graphical models provide, at first glance, a snapshot of the relevant relationships among the variables under consideration. We analyse the behaviour of the proposed methodology in a practical case, showing that it is a useful tool to help decision making when elaborating policies based on performance indicators. The methodology has been implemented in a software that interacts with the Elvira package for graphical models, and that is available to the administration board at the University of Almería (Spain) through a web interface. The software also implements a new method for constructing composite indicators by using a Bayesian network regression model.

Keywords Performance indicators · Bayesian networks · Relevance analysis · Higher education management · Decision support systems · Composite indicators

1 Introduction

During the last decades, the way in which the financial support provided by the administration to public Universities is determined, has been gradually moved to a system where an increasing part of the funds are obtained depending on the goals achieved by each institution. The usual way to determine to which extent an institution has achieved the compromised goals is through the so-called *performance indicators* [3]. Sometimes, the term *performance* is understood in a wide sense, assuming that a performance indicator is any institutional goal that can be objectively measured [4].

In order to design efficient policies oriented to increase the amount of public funds, the administration boards of the Universities should determine which variables, under their control, actually have an impact on the value of the performance indicators that are lately used

A. Fernández · M. Morales · C. Rodríguez · A. Salmerón (✉)
Department of Statistics and Applied Mathematics, University of Almería, 04120 Almería, Spain
e-mail: antonio.salmeron@ual.es

to compute the funds. This task requires to take into account a high number of variables of different nature (qualitative and quantitative), and which may have a complex dependence structure. In the last years, there has been an increasing interest, within the fields of Statistics and also in Artificial Intelligence, in handling scenarios in which a high number of variables take part. One of the most satisfactory solutions is based on the use of *probabilistic graphical models* and, more precisely, *Bayesian networks* [1,7]. Examples of applications of Bayesian networks in enterprise information systems can be found in the literature [17].

The main advantage of Bayesian networks is that they have a rich semantics, and they can be easily interpreted by the user with no need of a high background on Statistics. From an operational point of view, Bayesian networks provide a natural framework for relevance analysis and also they can be used for prediction tasks [11].

In this paper, we propose a methodology for relevance analysis of performance indicators in higher education based on the use of Bayesian network models. We illustrate the appropriateness of the proposed methodology for the particular case of the University of Almería (Spain). We also describe the decision support system designed to implement this methodology. The system interacts with the Elvira platform [5], and provides a web interface that guides the user through the process of determining the variables that are relevant to a given performance indicator. Furthermore, the software implements a novel procedure for constructing composite indicators, based on rankings provided by experts. Composite indicators [12] are indicators that sum up the information provided by various indicators of different nature, with the aim of describing, with a single number, the performance of an institution. Our proposal is a supervised algorithm, consisting of creating a database using the rankings provided by the experts and the corresponding values of the individual indicators. The composite indicator is constructed through a Bayesian network regression model [11] induced from the aforementioned database.

The rest of the paper is organised as follows. In Sect. 2, we explain the fundamentals of the methodology for relevance analysis using Bayesian networks. Section 3 is devoted to show the behaviour of the proposed technique in a real-world problem. The software developed to implement the methodology is described in Sect. 4. We describe the procedure to construct composite indicators in Sect. 5. The paper ends with conclusions in Sect. 6.

2 Bayesian networks and relevance analysis

A *Bayesian network* [1,7] is a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, which is defined in terms of two components:

- **Qualitative component**: A directed acyclic graph (DAG) where each vertex represents one of the variables in the model, and so that the presence of an edge linking two variables indicates the existence of statistical dependence between them.
- **Quantitative component**: A conditional distribution $p(x_i|pa(x_i))$ for each variable $X_i, i = 1, \ldots, n$ given its parents in the graph, denoted as $pa(X_i)$.

For example, the graph depicted in Fig. 1 could be the qualitative component of a Bayesian network for variables $X_1, \ldots, X_5$. According to the structure of the graph, it would be necessary to specify a conditional distribution for each variable given its parents. In this case, the distributions are $p(x_1)$, $p(x_2|x_1)$, $p(x_3|x_1)$, $p(x_4|x_2, x_3)$ and $p(x_5|x_4)$.

In the next two subsections, we will describe how the qualitative component encodes the dependencies among the variables in the model, and how the strength of the dependencies is determined by the quantitative component, i.e., the conditional distributions.

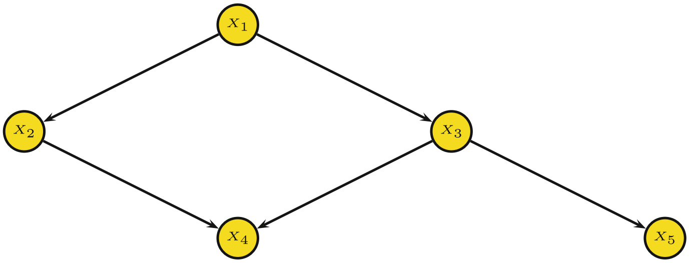
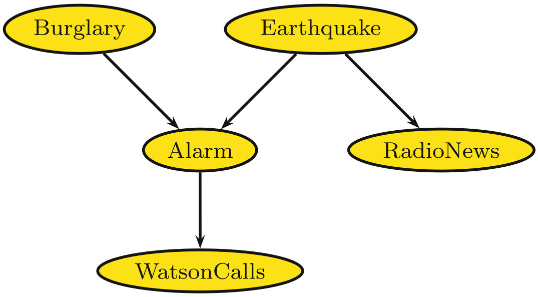


Fig. 1 An example of Bayesian network with five variables

Fig. 2 The Bayesian network for the *burglary or earthquake* example



2.1 Qualitative component of a Bayesian network

One of the most important advantages of Bayesian networks is that the structure of the associated DAG determines the dependence and independence relationships among the variables, so that it is possible to find out, with no need of carrying out any numerical calculations, which variables are relevant or irrelevant for some other variable of interest (for instance, a performance indicator). More precisely, we will illustrate how the relevance analysis is performed in Bayesian networks through the concept of *transmission of information*, so that two variables are irrelevant to each other if no information can be transmitted between them.

We will use a toy example, taken from [6], to explain the transmission of information in a Bayesian network.

Example 2.1 (Burglary or earthquake) Mr. Holmes is working in his office when he receives a phone call from his neighbour Dr. Watson, who tells him that Holmes' burglar alarm has gone off. Convinced that a burglar has broken into his house, Holmes rushes to his car and heads for home. On his way, he listens to the radio, and in the news it is reported that there has been a small earthquake in the area. Knowing that earthquakes have a tendency to turn burglar alarms on, he returns to his work.

The scenario described in example 2.1 can be represented by the Bayesian network in Fig. 2. In general, there are only three types of connections among variables in a DAG: serial, converging and diverging connections. Therefore, it is enough to explain how information flows for these three types of connections. We will use the example mentioned previously to illustrate this.

1. **Serial connections**.
   - "Burglary" has a causal influence on "Alarm", which in turn has a causal influence on "Watson calls". Therefore, information flows from "Burglary" to "Watson calls"

and vice versa, since knowledge about one of the variable provides information about the other.
- However, if we observe "Alarm", any information about the state of "Burglary" is irrelevant to our belief about "Watson calls" and vice versa, since once we have certainty about the fact that the alarm has gone off, the information provided by Watson does not change our state of belief.

2. **Diverging connections**.
   - "Earthquake" has a causal influence on both "Alarm" and "Radio news". Therefore, information flows from "Alarm" to "Radio news" and vice versa, since knowledge about one of the variable provides information about the other. For instance, if our only knowledge is that the radio news reported a small earthquake, our belief about the alarm going off would increase.
   - On the other hand, if we observe "Earthquake", i.e. we have certainty about that, any information about the state of "Alarm" is irrelevant for our belief about an earthquake report in the "Radio news" and vice versa.

3. **Converging connections**.
   - "Alarm" is causally influenced by both "Burglary" and "Earthquake". However, in this case the last two variables are irrelevant to each other: If we do not have any information about the alarm, there is no relationship between the other two variables.
   - However, if we observe "Alarm" and "Burglary", then this will effect our belief about "Earthquake": Burglary explains the alarm, reducing our belief that earthquake is the triggering factor, and vice versa.

In general, the rules for interpreting information flow given the structure of a Bayesian network are the following:

- **Serial connections** ($X \longrightarrow Y \longrightarrow Z$). Information may be transmitted through a serial connection unless the state of the variable (Y) in the connection is known.
- **Diverging connections** ($X \longleftarrow Y \longrightarrow Z$). Information may be transmitted through a diverging connection, unless the state of the variable (Y) in the connection is known.
- **Converging connections** ($X \longrightarrow Y \longleftarrow Z$). Information may only be transmitted through a converging connection if either information about the state of the variable in the connection (Y) or one of its descendants is available.

Applying these three rules, it is possible to determine the variables that are relevant to our goal variable. For instance, we could determine which are the variables over which the administration board of a university has to operate in order to change the value of a performance indicator.

2.2 Quantitative component of a Bayesian network

Though relevance analysis can be carried out simply taking into account the structure of the network, once the relevant variables for a given performance indicator have been located, it is necessary to know to which extent the changes in those variables determine the value of the performance indicator. This is achieved by using the quantitative component of the Bayesian network.

Taking into account the independencies encoded by the network structure, it holds that the joint distribution over all the variables is equal to the product of the conditional distributions attached to each node, so that

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i|pa(x_i)). \tag{1}$$

Assume X_i is the performance indicator in which we are interested, and X_E is a set of variables that can be controlled by the administration board. Then, the prediction for the value of X_i given X_E can be obtained by computing the distribution $p(x_i|x_E)$ that would provide us the likelihood of each possible value of X_i given each possible configuration of X_E. This distribution can be obtained from the joint distribution in Eq. (1). In fact, there is no need to compute the joint distribution, since there are efficient algorithms that allow to compute $p(x_i|x_E)$ taking advantage of the factorisation of the joint distribution imposed by the network structure [9,15].

2.3 Constructing a Bayesian network from a database

Once we know how to use a Bayesian network model for relevance analysis, we must consider how to obtain it. Nowadays, university administration is fully assisted by computers, so that a large amount of statistical data is available. More precisely, it is in general possible to obtain databases composed of records describing items of information that contain the value of some performance indicators together with other variables that can be controlled. For instance, we can have a record with information about a course (number of students, number of lecturers, etc.) together with some performance indicator regarding that course (the success rate, for instance).

There are several algorithms that allow the construction of Bayesian networks from databases. We will mention two of them that are commonly used: The so called K2 [2] and PC [16] algorithms. The K2 algorithm searches within the space of all Bayesian networks that contain the variables in the database, and tries to find an optimal network in terms of the likelihood of the database for each candidate network. On the other hand, the PC algorithm tries to determine the structure of the network by means of statistical tests of independence. None of the methods is absolutely superior to the other, so that in practical applications it is common to construct two networks, one with each algorithm, and then use the network for which the likelihood of the database is higher. A common feature of both algorithms is that they operate with qualitative variables; therefore, the continuous variables must be discretised beforehand. A review on discretisation methods can be found in [8].

There are free software packages that allow the construction of Bayesian networks from databases. In this work, we have used the Elvira system [5] which is available at http://leo.ugr.es/elvira.

3 Case study: Application to the analysis of performance indicators at the University of Almería

In this section, we describe a practical application of the methodology introduced in Sect. 2, consisting of the analysis of some performance indicators that are used to compute the amount of public funds received by the University of Almería (Spain).

The starting point is a database with 1,345 records and 17 variables regarding all the courses taught at the University of Almería in the different degree programs during the

Table 1 Description of the variables considered in the case study

Variable	Description
Performance rate	Ratio between the number of students that succeed in a course and the number of students that go to the exam.
Relative mark	Average of the marks obtained by each student in the course in relation to the other students' marks.
RepStudents	Percentage of students that repeat a course.
Used exam diets	Number of times a student goes to the course exam before passing.
Rate of diets used	Number of times a student goes to the course exam divided by the maximum number of trials allowed.
#StudTheLect	Number of students per classroom in theoretical lectures.
#StudPrtLect	Number of students per classroom in practical lectures.
Type of course	Whether the course is compulsory or optional
Semester	The semester, within the degree schedule, in which the course is taught.
#Lecturers	Number of lecturers in the same course.
Lecturer evaluation	Mark obtained by the lecturer in the students opinion polls.
Type of lecturer	Position of the lecturer.
PerctPhD	Percentage of lecturers in the course with PhD degree.
Degree program	The degree program in which the course is taught.
FullTimeStud	Percentage of full time students.
AvgAccessMark	Average marks of the students in the degree obtained in the high school.
P80AccessMark	80th percentile of the marks of the students in the degree obtained in the high school.

academic year 2003–2004. A description of the considered variables can be found in Table 1. The first five variables correspond to academic performance indicators.

The database has been pre-processed by discretising the continuous variables using the k-means clustering algorithm, which is one of the most popular clustering algorithms in data mining [18], establishing a number of five categories for each discretised variable. We have used the PC and the K2 algorithms, obtaining the best model, in terms of likelihood of the data, with the K2. The resulting network can be seen in Fig. 3.

Attending the structure of the network in that figure, it can be seen that there are two important variables, *Type of Course* and *Degree Program*, which play an important role in the network, since information can flow from them to all the performance indicators. We can evaluate the importance of these variables using the quantitative part of the Bayesian network. For instance, if we concentrate on variable *Type of Course*, its influence on two important performance indicators, *Performance Rate* and *Relative Mark* is clear attending to the conditional probabilities displayed in Tables 2 and 3 , respectively. The differences for the distribution of the values of both performance indicators are significant depending on whether the course is compulsory or optional. This fact suggests that a separate study of compulsory and optional courses is appropriate.

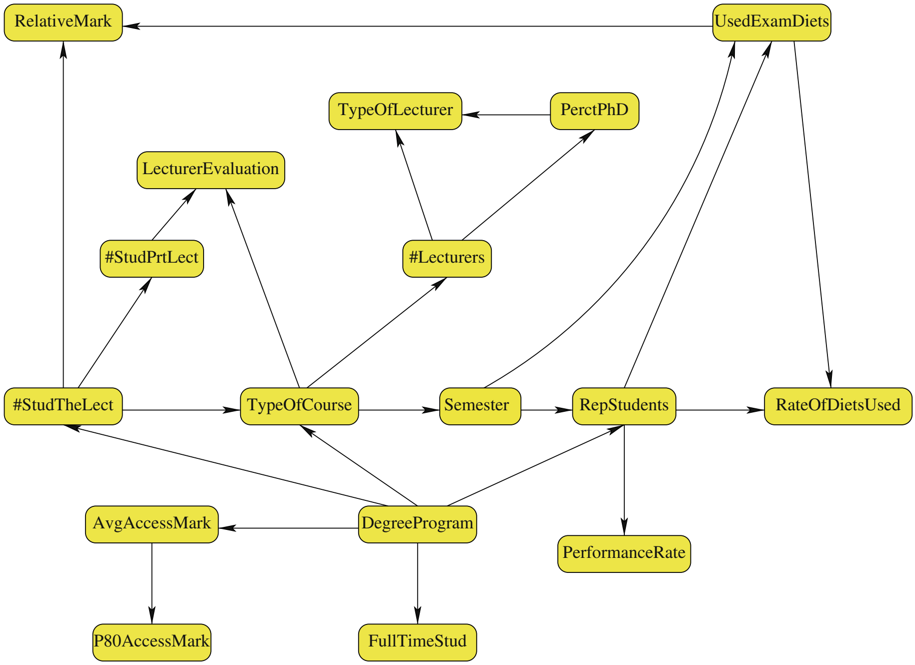


Fig. 3 Bayesian network obtained for the case study

Table 2 Conditional probabilities of performance rate given the type of course

Performance rate	Prior probability	Type of course	
		Compulsory	Optional
[0, 0.545)	0.03	0.04	0.03
[0.545, 0.735)	0.10	0.15	0.04
[0.735, 0.855)	0.17	0.25	0.09
[0.855, 0.955)	0.20	0.26	0.13
[0.955, 1]	0.49	0.31	0.71

Table 3 Conditional probabilities of relative mark given the type of course

Relative mark	Prior probability	Type of course	
		Compulsory	Optional
[0, 0.195)	0.23	0.28	0.15
[0.195, 0.315)	0.27	0.29	0.24
[0.315, 0.465)	0.24	0.21	0.28
[0.465, 0.775)	0.18	0.14	0.23
[0.775, 1]	0.08	0.07	0.09

3.1 Relevance analysis for compulsory courses

The Bayesian network obtained using the registers in the database concerning compulsory courses is displayed in Fig. 4. We can draw the following conclusions:

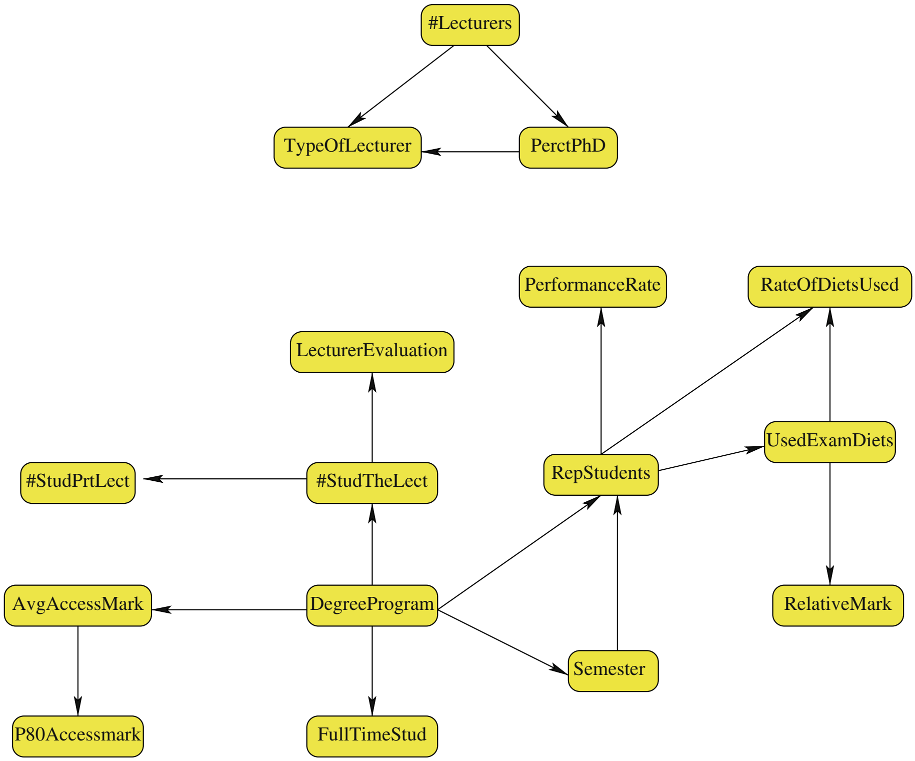


Fig. 4 Bayesian network for compulsory courses

- The structure of the Lecturer board is irrelevant to the rest of the network, since variables *number of lecturers*, *type of lecturer* and *percentage of PhDs* are disconnected from the rest.
- The evaluation obtained by a lecturer in the opinion polls is fully determined by the number of students in classrooms in theoretical lectures. It is an important conclusion, since it is common to find poor evaluation results in large classrooms, which suggests that rather than the lecturer's profile it is the size of the classroom that determines the result of the evaluation.
- Any possible information flow towards the performance indicators goes through variable *Degree Program*. It is true that the administration board cannot control the value of the degree program where a subject is included, but they actually can control some characteristics of the degree program as the access mark and number of students per classroom in theoretical and practical lectures. The effect of these variables on the performance rate is illustrated in Tables 4, 5 and 6.
- Attending to the results in Table 4, we can conclude that a good policy in order to increase the performance rate is to establish a maximum number of students per classroom not greater than 49.
- The influence of the number of students in practical lectures is not so important, as can be seen in Table 5 (the columns are rather similar).
- Finally, the access marks have little impact on the performance rate. Only increasing the 80th percentile of the access mark above 7.3 points out of 10, a slight improvement in the performance rate can be noticed.

Table 4 Performance rate versus # students in theoretical lectures for compulsory subjects

Performance rate	Prior probability	# of students per classroom in theoretical lectures				
		<25.5	[25.5, 49.5)	[49.5, 79.25)	[79.25, 114.75)	≥114.75
[0, 0.355)	0.20	0.18	0.19	0.20	0.20	0.24
[0.355, 0.535)	0.28	0.26	0.27	0.28	0.28	0.29
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.22
[0.695, 0.845)	0.18	0.20	0.19	0.18	0.18	0.15
[0.845, 1]	0.11	0.13	0.12	0.11	0.11	0.09

Table 5 Performance rate versus # students in practical lectures for compulsory subjects

Performance rate	Prior probability	# of students per classroom in practical lectures				
		<23.9	[23.9, 41.65)	[41.65, 68.5)	[68.5, 114.16)	≥114.16
[0, 0.355)	0.20	0.19	0.20	0.21	0.20	0.24
[0.355, 0.535)	0.28	0.27	0.28	0.28	0.28	0.29
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.24
[0.695, 0.845)	0.18	0.19	0.18	0.18	0.18	0.15
[0.845, 1]	0.11	0.12	0.11	0.11	0.11	0.10

Table 6 Performance rate versus student profile for compulsory subjects

Performance rate	Prior probability	Average access mark				
		[5.32, 5.92)	[5.92, 6.17)	[6.17, 6.46)	[6.46, 6.83)	[6.83, 7.71]
[0, 0.355)	0.20	0.19	0.21	0.21	0.19	0.19
[0.355, 0.535)	0.28	0.28	0.28	0.28	0.27	0.27
[0.535, 0.695)	0.23	0.24	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.18	0.17	0.17	0.18	0.19
[0.845, 1]	0.11	0.11	0.11	0.11	0.12	0.13
Performance rate	Prior probability	80th percentile of the access mark				
		[5.46, 6.30)	[6.30, 6.65)	[6.65, 7.30)	[7.30, 7.61)	[7.61, 8.5]
[0, 0.355)	0.20	0.19	0.21	0.21	0.20	0.19
[0.355, 0.535)	0.28	0.28	0.28	0.28	0.27	0.27
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.18	0.18	0.18	0.18	0.19
[0.845, 1]	0.11	0.11	0.11	0.11	0.12	0.12

3.2 Relevance analysis for optional courses

The Bayesian network obtained using the registers in the database concerning optional courses is displayed in Fig. 5. Analysing the structure of this network, we can deduce that the lecturers' profile, including the result of the evaluation, is irrelevant to the course performance indicators.

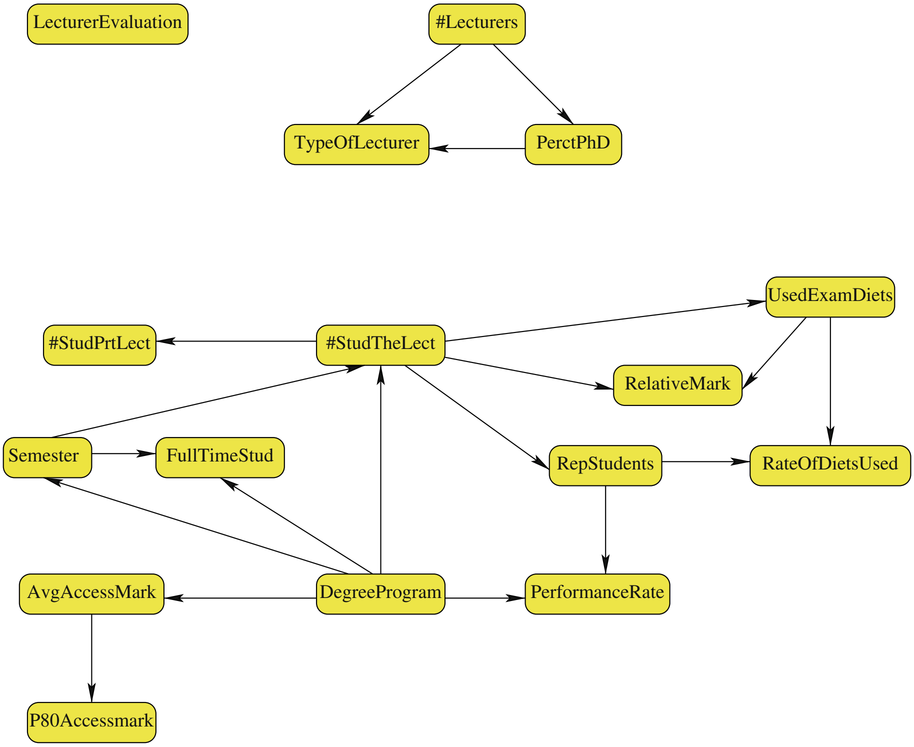


Fig. 5 Bayesian network for optional courses

The access marks are connected to the performance rate through the degree program. Their impact on this indicator is quantified in Table 7. The probabilities in that table indicate that the best performances are attained for average access marks above 6.69 points and 80th percentiles above 7.22 points.

The number of students in theoretical lectures is more relevant here that in the case of compulsory subjects, since it is connected to the relative mark, the number of diets used and the percentage of repeating students. Also, it is indirectly connected to the performance rate.

Table 8 summarises the probabilities of indicator *performance rate*, given the number of students in theoretical lectures. It can be observed that the performance rate is strongly influenced by this variable, so that low performances appear when the number of students increases. Therefore, any policy oriented to decrease the number of students per classroom conveys a significant improvement in the course performance.

Finally, it can be concluded from the probabilities in Table 9 that the influence of the number of students in practical lectures is not as important as in the case of theoretical lectures.

4 Software for relevance analysis

We have implemented the above described methodology in a software package, called *academic advisor*, which provides an intuitive web-based interface appropriate for academic staff not familiarised with Bayesian network models.

Table 7 Performance rate given access marks for optional courses

Performance rate	Prior probability	Average access mark				
		[5.32, 6.00)	[6.00, 6.20)	[6.20, 6.44)	[6.44, 6.69)	[6.69, 8.03]
[0, 0.59)	0.20	0.18	0.19	0.22	0.22	0.17
[0.59, 0.71)	0.19	0.17	0.20	0.29	0.19	0.18
[0.71, 0.82)	0.20	0.21	0.21	0.21	0.21	0.18
[0.82, 0.92)	0.20	0.23	0.21	0.19	0.19	0.19
[0.92, 1]	0.21	0.20	0.19	0.19	0.20	0.28
Performance rate	Prior probability	80th percentile of the access mark				
		[5.37, 6.28)	[6.28, 6.61)	[6.61, 6.92)	[6.92, 7.22)	[7.22, 8.57]
[0, 0.59)	0.20	0.19	0.19	0.21	0.21	0.18
[0.59, 0.71)	0.19	0.18	0.19	0.19	0.19	0.18
[0.71, 0.82)	0.20	0.21	0.21	0.21	0.20	0.18
[0.82, 0.92)	0.20	0.22	0.21	0.20	0.19	0.19
[0.92, 1]	0.21	0.20	0.20	0.20	0.21	0.26

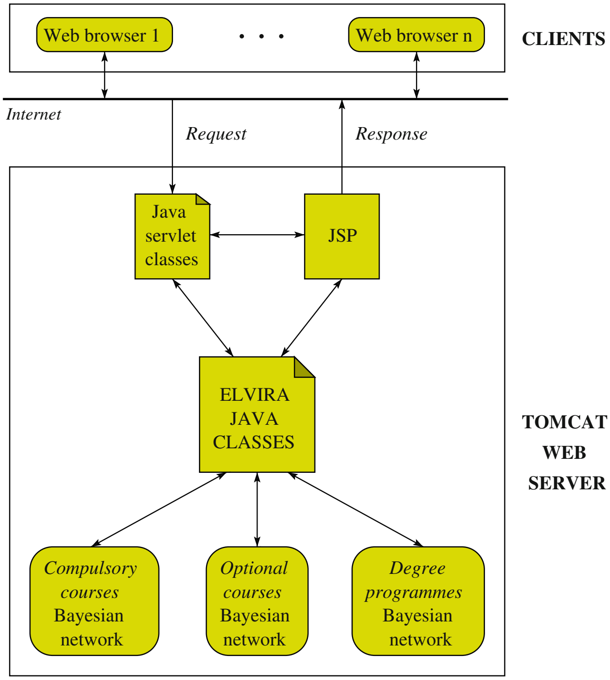


Fig. 6 Structure of the decision support system for relevance analysis

The functionality of the academic advisor is based on a client/server Web architecture (Fig. 6). In the client side, the users interact with the system using a Web browser to access an interface with data forms. The server side, contains most of the functionality of the

Table 8 Results for optional subjects given the size of the classrooms in theoretical lectures

Performance rate	Prior probability	# students per classroom in theoretical lectures				
		<10	[10, 18)	[18, 28)	[28, 51)	[51, 137]
[0, 0.59)	0.20	0.18	0.20	0.19	0.21	0.20
[0.59, 0.71)	0.19	0.15	0.19	0.19	0.20	0.19
[0.71, 0.82)	0.20	0.17	0.20	0.21	0.21	0.22
[0.82, 0.92)	0.20	0.19	0.19	0.19	0.21	0.23
[0.92, 1]	0.21	0.31	0.22	0.21	0.17	0.16

Table 9 Results for optional subjects given the size of the classrooms in practical lectures

Performance rate	Prior probability	# students per classroom in practical lectures				
		<9	[9, 17)	[17, 25)	[25, 38)	≥38
[0, 0.59)	0.20	0.18	0.20	0.20	0.20	0.20
[0.59, 0.71)	0.19	0.16	0.18	0.19	0.19	0.19
[0.71, 0.82)	0.20	0.17	0.20	0.21	0.21	0.21
[0.82, 0.92)	0.20	0.19	0.19	0.20	0.21	0.22
[0.92, 1]	0.21	0.30	0.23	0.20	0.18	0.17

application. The system uses the Web server Apache Tomcat 5.5 Servlet/JSP Container, which allows to run servlets and to generate JSPs (Java Server Pages) automatically. In addition, Java classes of the Elvira program [5] and the Bayesian networks described in Sect. 3 are stored.

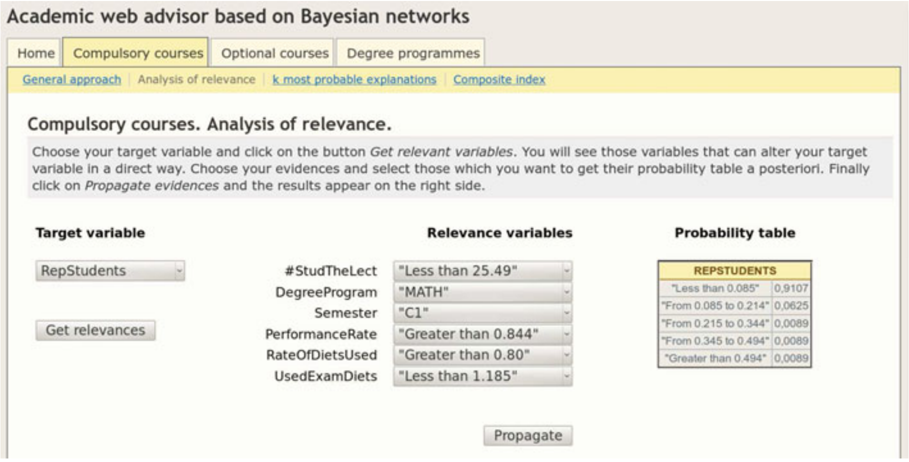
Servlets and JSPs are two methods to create dynamic Web pages under the context of a server and using the Java language. More precisely, the JSPs are HTML pages with special labels and Java code embedded using scripts, which makes possible to generate dynamic content. On the other hand, a servlet is a Java program that receives requests, processes them and generates a Web page from them.

We can justify the use of these technologies from different points of view. At first, this problem requires a constant interactivity between the application and the user. In addition, the client/server approach seems appropriate, since it makes possible the remote use of the application by different users at the same time. On the other hand, the use of Java language allows the direct interaction with the Elvira system, which is implemented using that language.

The interaction process is made as follows: the users introduce the input data using a Web form designed in HTML, or JSP in case of needing some kind of processing in Java. This request is sent to the server to activate the corresponding servlet that manages the search interacting with the underlying Elvira Java classes and Bayesian network files. Once the information has been processed, another servlet is in charge of generating a HTML/JSP page that will show the results to the user (response).

From the point of view of the users, the system can be used for four tasks: relevance analysis, probability propagation, profile extraction and construction of composite indicators.

In the *relevance analysis module*, which is depicted in Fig. 7, the user can choose a target variable and the system returns the list of variables that are directly related to it, according to the criteria for interpreting the structure of the underlying Bayesian network explained in Sect. 2. The process can be repeated in order to detect the relevant variables at a second level



Compulsory courses. Analysis of relevance.

Choose your target variable and click on the button *Get relevant variables*. You will see those variables that can alter your target variable in a direct way. Choose your evidences and select those which you want to get their probability table a posteriori. Finally click on *Propagate evidences* and the results appear on the right side.

Target variable: RepStudents

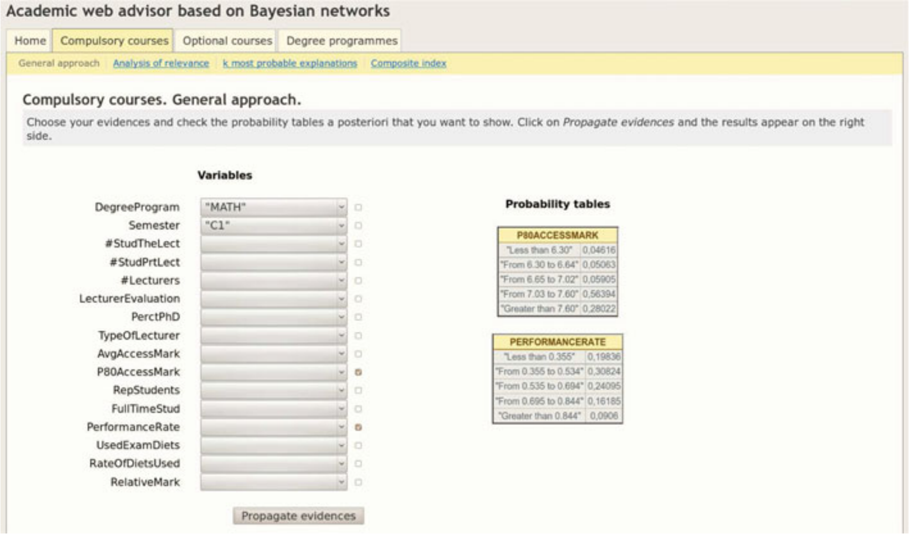
Relevance variables

#StudTheLect	"Less than 25.49"
DegreeProgram	"MATH"
Semester	"C1"
PerformanceRate	"Greater than 0.844"
RateOfDietsUsed	"Greater than 0.80"
UsedExamDiets	"Less than 1.185"

Probability table

REPSTUDENTS	
"Less than 0.085"	0,9107
"From 0.085 to 0.214"	0,0625
"From 0.215 to 0.344"	0,0089
"From 0.345 to 0.494"	0,0089
"Greater than 0.494"	0,0089

Fig. 7 The relevance analysis screen of the academic advisor



Compulsory courses. General approach.

Choose your evidences and check the probability tables a posteriori that you want to show. Click on *Propagate evidences* and the results appear on the right side.

Variables

DegreeProgram	"MATH"	□
Semester	"C1"	□
#StudTheLect		□
#StudPrtLect		□
#Lecturers		□
LecturerEvaluation		□
PerctPhD		□
TypeOfLecturer		□
AvgAccessMark		□
P80AccessMark		☑
RepStudents		□
FullTimeStud		□
PerformanceRate		☑
UsedExamDiets		□
RateOfDietsUsed		□
RelativeMark		□

Probability tables

P80ACCESSMARK	
"Less than 6.30"	0,04616
"From 6.30 to 6.64"	0,05063
"From 6.65 to 7.02"	0,05905
"From 7.03 to 7.60"	0,56394
"Greater than 7.60"	0,28022

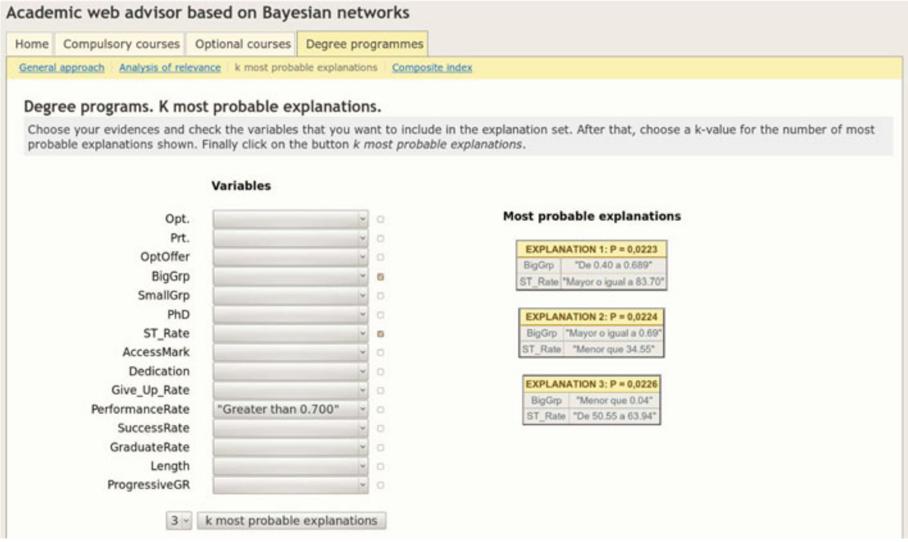
PERFORMANCERATE	
"Less than 0.355"	0,19836
"From 0.355 to 0.534"	0,30824
"From 0.535 to 0.694"	0,24095
"From 0.695 to 0.844"	0,16185
"Greater than 0.844"	0,0906

Fig. 8 The probability propagation screen of the academic advisor

in the Bayesian network, and so on. Finally, the posterior distribution of the target variable, given the relevant selected ones, can be obtained.

In the *probability propagation module*, the user can obtain the posterior probability distribution of any target variable, given an assignment of values to some other variables (see Fig. 8).

The *profile extraction module* can be seen in Fig. 9. This module allows to compute a set of explanations to a given fact. For instance, we can compute the best explanation for a given value of the success rate in terms of number of students in classroom and in terms of the rate student/teacher. This tool is very useful for descriptive purposes, as it allows to determine typical situations under given restrictions. The problem of finding a set of explanations, also



Academic web advisor based on Bayesian networks

Degree programs. K most probable explanations.

Choose your evidences and check the variables that you want to include in the explanation set. After that, choose a k-value for the number of most probable explanations shown. Finally click on the button *k most probable explanations*.

Variables

Opt. □
Prt. □
OptOffer □
BigGrp ☑
SmallGrp □
PhD □
ST_Rate ☑
AccessMark □
Dedication □
Give_Up_Rate □
PerformanceRate "Greater than 0.700" □
SuccessRate □
GraduateRate □
Length □
ProgressiveGR □

3 [k most probable explanations]

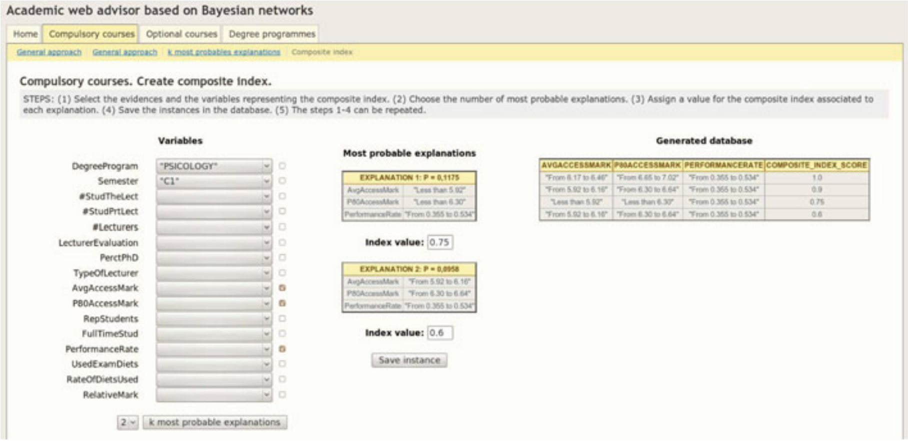
Most probable explanations

EXPLANATION 1: P = 0,0223	
BigGrp	"De 0.40 a 0.689"
ST_Rate	"Mayor o igual a 83.70"

EXPLANATION 2: P = 0,0224	
BigGrp	"Mayor o igual a 0.69"
ST_Rate	"Menor que 34.55"

EXPLANATION 3: P = 0,0226	
BigGrp	"Menor que 0.04"
ST_Rate	"De 50.55 a 63.94"

Fig. 9 The profile extraction screen of the academic advisor



Academic web advisor based on Bayesian networks

Compulsory courses. Create composite Index.

STEPS: (1) Select the evidences and the variables representing the composite index. (2) Choose the number of most probable explanations. (3) Assign a value for the composite index associated to each explanation. (4) Save the instances in the database. (5) The steps 1-4 can be repeated.

Variables

DegreeProgram "PSICOLOGY" □
Semester "C1" □
#StudTheLect □
#StudPrtLect □
#Lecturers □
LecturerEvaluation □
PerctPhD □
TypeOfLecturer □
AvgAccessMark ☑
P80AccessMark ☑
RepStudents □
FullTimeStud □
PerformanceRate ☑
UsedExamDiets □
RateOfDietsUsed □
RelativeMark □

2 [k most probable explanations]

Most probable explanations

EXPLANATION 1: P = 0,1175	
AvgAccessMark	"Less than 5.92"
P80AccessMark	"Less than 6.30"
PerformanceRate	"From 0.355 to 0.534"

Index value: 0.75

EXPLANATION 2: P = 0,0958	
AvgAccessMark	"From 5.92 to 6.16"
P80AccessMark	"From 6.30 to 6.64"
PerformanceRate	"From 0.355 to 0.534"

Index value: 0.6

[Save instance]

Generated database

AVGACCESSMARK	P80ACCESSMARK	PERFORMANCERATE	COMPOSITE_INDEX_SCORE
"From 6.17 to 6.46"	"From 6.65 to 7.02"	"From 0.355 to 0.534"	1.0
"From 5.92 to 6.16"	"From 6.30 to 6.64"	"From 0.355 to 0.534"	0.9
"Less than 5.92"	"Less than 6.30"	"From 0.355 to 0.534"	0.75
"From 5.92 to 6.16"	"From 6.30 to 6.64"	"From 0.355 to 0.534"	0.6

Fig. 10 Screen for constructing composite indicators

known as MAP problem, is solved using the implementation in Elvira, which corresponds to the method proposed in [13].

The module for constructing composite indicators is shown in Fig. 10, and described in Sect. 5.

5 Using the software to construct composite indicators

Composite indicators [12] are indicators that sum up the information provided by various indicators of different nature, with the aim of describing, with a single number, the performance of an institution. The module for constructing composite indicators implements a

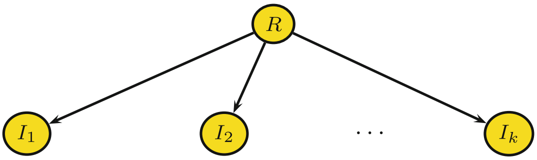


Fig. 11 Naïve Bayes structure for regression

novel methodology of supervised nature. It is supervised in the sense that there must be an expert that ranks different descriptions of the institution in terms of performance. Out of that description, the software creates a composite indicator that is computed from the values of the individual indicators for each description. We give the details of this two main tasks in the next sub-sections.

5.1 Generating the rank of descriptions

The first step to construct a composite indicator is to choose a set of individual indicators $I_1, \dots, I_k$. Then, an expert gives a set of descriptions of the institution in terms of some observable variables. For each description, the software computes a list of profiles that explain the given description, and show them on the screen. Afterwards, the expert assigns a number between 0 and 1 to each profile, according to the performance of the institution corresponding to each description, where a value close to 1 indicates high performance, and a value close to 0 means low performance. A screenshot of this procedure can be seen in Fig. 10. After this process, the software has stored a database D with variables $I_1, \dots, I_k, R$, where R is the ranking assigned to each description by the expert.

5.2 Generating the composite index from the database

Using the database D, described earlier, we construct the composite indicator by using a Bayesian network regression model [11]. Consider the set of variables $R, I_1, \dots, I_k$ in database D, where R is the composite indicator (response variable) and $I_1, \dots, I_k$ are the individual indicators (explanatory variables). Note that R is constructed from the rankings given by the expert, and therefore we use the same variable name. Regression analysis consists of finding a model g that explains the *response* variable R in terms of the *explanatory* variables $I_1, \dots, I_k$, so that given an assignment of the explanatory variables, $i_1, \dots, i_k$, a prediction about R can be obtained as $\hat{r} = g(i_1, \dots, i_k)$. We represent the joint distribution of $R, I_1, \dots, I_k$ as a Bayesian network, and use the posterior distribution of R given $I_1, \dots, I_k$ (more precisely, its expectation) to obtain a prediction for R.

In this paper, we will focus on a particular Bayesian network structure, the so-called naïve Bayes, as in [11]. The *naïve Bayes* [19] structure is an extreme case in which all the explanatory variables are considered independent given the response variable. This kind of structure is represented in Fig. 11. The reason to make the strong independence assumption behind naïve Bayes models is that it is compensated by the reduction of the number of parameters to be estimated from data, since in this case, it holds that the conditional distribution of the response variable can be factorised as

$$f(r|i_1, \dots, i_k) = f(i) \prod_{j=1}^{n} f(i_j|r), \tag{2}$$

which means that, instead of one k-dimensional conditional densities, k one-dimensional conditional densities are estimated. As we use the conditional expectation of the response variable, given the observed explanatory variables, our regression model will be

$$\hat{r} = g(i_1, \ldots, i_k) = E[R|i_1, \ldots, i_k] = \int_{\Omega_R} r f(r|i_1, \ldots, i_k) dr, \tag{3}$$

where $f(r|i_1, \ldots, i_k)$ is the conditional density of R given $i_1, \ldots, i_k$, which we assume to be of class MTE (Mixture of Truncated Exponentials). The MTE model [10] was shown to outperform state-of-the-art regression models as model trees [11].

After constructing the composite indicator as in Eq. (3), it is included in the system and can be computed from every possible combination of values of the individual indicators.

6 Conclusions

In this paper, we have introduced a methodology for relevance analysis of performance indicators in higher education. We have shown through a case study that this methodology can help the process of decision making when designing policies oriented to increase the amount of public funds when they are assigned according to some performance indicators.

The graphical nature of the used model allows drawing conclusions with no need to interpret any numerical data, since relevance analysis can be carried out just taking into account the structure of the Bayesian network. If the user is also interested in quantifying the strength of the dependencies among the variables, it can be achieved using the conditional probability distributions provided by the Bayesian network as well.

We have also introduced the *academic advisor* system, which implements the proposed methodology making use of the Elvira system, but providing a user friendly interface appropriate for academic staff not familiarised with Bayesian network models. An important novel feature of this software is that it allows to construct composite indicators in an easy way, using an expert's opinion.

The fact that the software interacts with the Elvira system makes it easy to update its knowledge base with, for instance, indicator databases corresponding to forthcoming academic years, or any other database. It is specially interesting from the point of view of data privacy, as there is no need of external intervention for updating the system. The importance of preserving data privacy during external intervention in data mining tasks is analysed in [14].

In the next future, we plan to apply Bayesian technology to construct a recommendation system for students, so that they can choose the appropriate courses in order to maximise their success chances.

We also plan to improve the module for constructing composite indicators, by following a semi-supervised approach, in which there is no need to specify the value for the composite indicator in all the records of the training database.

Acnowledgments We thank anonymous reviewers for their very useful comments and suggestions. This work has been supported by the Spanish Ministry of Science and Innovation, under project TIN2007-67418-C03-02 and by FEDER funds.

References

1. Castillo E, Gutiérrez J, Hadi A (1997) Expert systems and probabilistic network models. Springer, New York

2. Cooper G, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Mach Learn 9:309–347
3. Cuenin S (1987) The use of performance indicators in universities: an international survey. Int J Inst Manage High Educ 2:117–139
4. Dochy F, Segers M, Wijnen W (1990) Selecting performance indicators. A proposal as a result of research. In: Goedegebuure F, Maasen F, Westerheijden D (eds). Lemma B.V. Peer review and performance indicators, pp 135–153
5. Elvira Consortium (2002) Elvira: an environment for creating and using probabilistic graphical models. In: Gámez JA, Salmerón A (eds). Proceedings of the First European Workshop on Probabilistic Graphical Models, pp. 222–230
6. Jensen FV (2001) Bayesian networks and decision graphs. Springer, New York
7. Jensen FV, Nielsen TD (2007) Bayesian networks and decision graphs. Second edition. Springer, New York
8. Jin R, Breitbart Y, Muoh C (2009) Data discretization unification. Knowl Inf Syst 19:1–29
9. Madsen A, Jensen FV (1999) Lazy propagation: a junction tree inference algorithm based on lazy evaluation. Artif Intell 113:203–245
10. Moral S, Rumí R, Salmerón A (2001) Mixtures of truncated exponentials in hybrid Bayesian networks. ECSQARU'01. Lect Notes Artif Intell 2143:135–143
11. Morales M, Rodríguez C, Salmerón A (2007) Selective naive Bayes for regression using mixtures of truncated exponentials. Int J Uncertain, Fuzziness Knowl Based Syst 15:697–716
12. Nardo M, Saisana M, Saltelli A, Tarantola S (2008) Handbook on constructing composite indicators: methodology and user guide. OECD, European Commission, Joint Research Centre
13. Nilsson D (1998) An efficient algorithm for finding the M most probable configurations in Bayesian networks. Stat Comput 9:159–173
14. Qiu L, Li Y, Wu X (2008) Protecting business intelligence and customer privacy while outsourcing data mining tasks. Knowl Inf Syst 17:99–120
15. Shenoy P, Shafer G (1990) Axioms for probability and belief function propagation. In: Shachter R, Levitt T, Lemmer J, Kanal L (eds). Uncertainty in artificial intelligence 4. North Holland, Amsterdam, pp 169–198
16. Spirtes P, Glymour C, Scheines R (1993) Causation, prediction and search. Series: lecture notes in statistics, vol. 81. Springer, New York
17. Wang Z, Wang Q, Wang D (2009) Bayesian network based business information retrieval model. Knowl Inf Syst 20:63–79
18. Wu X, Kumar V, Quinlan J, Gosh J, Yang Q, Motoda H, McLachlan G, Ng A, Liu B, Yu P, Zhou Z, Steinbach M, Hand D, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37
19. Zhang J, Kang D, Silvescu A, Honavar V (2006) Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data. Knowl Inf Syst 9:157–179

Author biographies

Antonio Fernández received a M.S. degree in Computer Science from the University of Almería, Spain in 2005. From 2006 to 2009 he was a research assistant in the Data Analysis Group of the University of Almería, where he is now a PhD student in the Department of Statistics and Applied Mathematics. He has been a visiting PhD student in the Machine Intelligence Group at Aalborg University, Denmark. His research interests include data mining, probabilistic graphical models and their applications.

María Morales is currently Associate Professor in the Department of Statistics and Applied Mathematics at the University of Almería, Spain. She obtained her PhD in Statistics from the University of Almería in 2006. She has been data analyst in the Data Coordination Unit of the University of Almería, where she was the responsible of starting up the datawarehouse for computing indicators related to higher education management. Her current research interest is mainly focused on graph theory.

Carmelo Rodríguez is currently Professor in the Department of Statistics and Applied Mathematics at the University of Almería, Spain. He obtained his PhD in Statistics from the University of Granada in 1993. Professor Rodríguez has been Deputy-Rector for student affairs from 1997 to 1999 and for academic planning from 1999 to 2007. He is currently the Head of the Department of Statistics and Applied Mathematics. He was the Chairman of the 24th Spanish Conference on Statistics and Operations Research, held in 1998. His current research interests include optimal experimental design and probabilistic graphical models.

Antonio Salmerón is currently Professor in the Department of Statistics and Applied Mathematics at the University of Almería, Spain. He obtained his PhD in Artificial Intelligence from the University of Granada in 1998, and received the José Cuena award from the Spanish Association for Artificial Intelligence in 2001. Professor Salmerón has been Chairman of the doctoral programs study board at the University of Almería from 2001 to 2007. He was co-Chairman of the First European Workshop on Probabilistic Graphical Models, held in 2002. His current research interests are mainly focused on probabilistic graphical models, specially hybrid Bayesian networks and probabilistic decision graphs.