Expert Systems with Applications 40 (2013) 5004–5010

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Expert system for predicting unstable angina based on Bayesian networks ☆

Joan Vila-Francés [a,*], Juan Sanchís [b], Emilio Soria-Olivas [a], Antonio José Serrano [a], Marcelino Martínez-Sober [a], Clara Bonanad [b], Silvia Ventura [b]

[a] Intelligent Data Analysis Laboratory, University of Valencia, Avd Universitat s/n, 46100 Burjassot, Valencia, Spain
[b] Servicio de Cardiología, Hospital Clinic Universitari de Valencia, Spain

ARTICLE INFO

Keywords:
Bayesian networks
Expert systems
Medical applications

ABSTRACT

The use of computer-based clinical decision support (CDS) tools is growing significantly in recent years. These tools help reduce waiting lists, minimise patient risks and, at the same time, optimise the cost health resources. In this paper, we present a CDS application that predicts the probability of having unstable angina based on clinical data. Due to the characteristics of the variables (mostly binary) a Bayesian network model was chosen to support the system. Bayesian-network model was constructed using a population of 1164 patients, and subsequently was validated with a population of 103 patients. The validation results, with a negative predictive value (NPV) of 91%, demonstrate its applicability to help clinicians. The final model was implemented as a web application that is currently been validated by clinician specialists.

1. Introduction

The use of computer-based clinical decision support (CDS) tools is growing in recent years due to different reasons (Steyerberg, 2009):

- Helps the clinicians making decisions, thus reducing the clinical errors.
- Improves the time to get a diagnostic reducing waiting time.
- Optimizes health resources reducing unnecessary medical tests.

Those advantages have expanded the use of these tools in clinical practice in the form of web services, desktop programs or applications for mobile phones and tablets. The use of these tools in different clinical areas, supported by the increasing amount of patient data and the ability to analyse and process it through Big Data techniques, is foreseen as a huge boost in health care (Manyika et al., 2011). Currently, one of the areas that demands more resources is cardiology. Nowadays cardiovascular diseases are the main cause of death in the developed world (Escaned et al., 2008). Modern lifestyle that leads us to many stressful situations, poor diet and little exercise have made heart disease the cause of many deaths. One of the symptoms of possible heart failure angina is characterised by severe chest pain. When suffering from that pain for more than 15 min, it is highly recommended requiring clinical attention, as it can be the initial phase of a myocardial infarction. This pain occurs when the demand for oxygen by the heart muscle is not served (because there is an interruption of the blood supply to a part of the heart muscle). This pain is in fact one of the most frequent causes of admission in the Emergency Services of the hospitals. Some of these patients suffer an acute coronary syndrome that is diagnosed by electrocardiogram (ECG) findings or alteration of biomarkers of myocardial damage (troponin). However, in some patients the ECG is nonspecific and troponin is normal. This population suffers form a chest pain of uncertain origin. They are mostly low-risk patients without any heart disease, but we can not reject an acute coronary syndrome in some of them. The key challenge is to identify those patients at risk for suffering an acute coronary syndrome with normal troponin (unstable angina), within a population that is generally at low risk. The tools currently available in the emergency room of a hospital do not work, because the ECG is nonspecific, and troponin is normal. As a result, the final decision of admission or release is postponed until a treadmill stress test is carried out (usually the next morning) Sanchis et al. (2006). This strategy is suboptimal because the patient must wait for several hours, many patients can not run on the treadmill, and sometimes the results are inconclusive.

This article proposes the development of a Clinical Decision Support System, for being use in the emergency units of a hospital in order to determine the probability of unstable angina within

☆ This work was supported by the Spanish Ministry of Education and Science under Grant Instituto Carlos III (FEDER), Red HERACLES 06/0009.
* Corresponding author. Tel.: +34 963543398.
E-mail address: joan.vila@uv.es (J. Vila-Francés).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.029

24 h of patient entry into the hospital. The inputs of the system are the clinical data that are routinely collected in the emergency room of a hospital. Almost all of those data have a binary response (e.g. – patient gender, smoker, etc.). With this kind of inputs, the most appropriate machine learning models are decision trees and Bayesian networks (Alpaydin, 2009). The results obtained from decision trees were not good enough (deep trees were needed, and therefore generalisation was bad). Moreover, a system whose parameters could be updated continuously when new information was available was required, that is why Bayesian networks were used. The advantages of these models for using them in clinical problems are Lucas et al. (2004):

- The Bayesian model can be interpreted by the clinician, as the relationships among variables are clearly represented by a graph (directed graph in our model).
- The clinician can provide expertise knowledge to establish new relationships between variables that might not have been reflected in the case of using an automatic learning algorithm for the Bayesian network structure.
- Adding new knowledge is a straight forward process, which can be automated by updating the frequency tables of each input variable.
- It is not necessary to know the values of all inputs to the model to obtain a valid output. Thus, if the inputs are obtained in a sequential way, as it happens in an emergency unit, or some information about the patient is missing, the system may be updating the probability as soon as new information of clinical tests is available.

Once the Bayesian network was validated, it was implemented into a Decision Support System which allowed clinicians to access it remotely, in an easy and simple way. The obvious solution was to implement the expert system as a web application. This way it is possible to centralise the data at a single location while access can be granted from any computer without requiring dedicated software (only a web browser and an Internet connection). Moreover, it is possible to implement a user-based management to control the people that use the tool and access the data. The rest of this paper is organised as follows. Section 2 explains the Bayesian models used. Section 3 discusses the data used and results obtained. Section 4 explains the Web tool development. Finally, Section 5 summarizes the conclusions of the present work.

2. Bayesian networks

A Bayesian network (BN) is a probabilistic graphical model composed of two different parts: on one hand is the graphical structure (directed acyclic graph) that defines the relationship between variables and, on the other hand, the probabilities established between these variables (Koller and Friedman, 2009; Korb and Nicholson, 2011). The elements of a Bayesian network are as follows Rusell and Norvig (2009):

- A set of variables (continuous or discrete) forming the network nodes.
- A set of directed links that connect a pair of nodes. If there is a relationship with direction $X \rightarrow Y$ is said that X is the parent of Y.

The network fulfils the following facts:

- Each node X_i is associated with a conditional probability function $P(X_i|Parents(X_i))$ that takes as input a particular set of values for the node's parent variables, and gives the probability of the variable represented by the node X_i.
- The graph has no directed cycles.

The knowledge is reflected by the relationships established in the graph nodes, and the conditional probabilities of the variables represented in each node. Those probabilities are estimated using the dataset. In this paper, the most widely used Bayesian classifiers have been applied: Naïve Bayes, FAN (Forest Augmented Network) and TAN (Tree Augmented Network). The following subsections explain each of these approaches.

2.1. Naïve Bayes

The hypothesis of this approach is to assume that the predictive variables are conditionally independent given the variable to classify. Although this assumption is quite restrictive, this classifier is one of the most used, and several studies show that the results are as good as the ones obtained with other techniques (neural networks, decision trees etc.) Korb and Nicholson (2011). The big problem with this model arises when the conditional independence among predictive variables is not fulfilled. This happens when predictive variables are redundant or are highly correlated with each other. Assuming the conditional independence, the graph of a Naïve Bayes is shown in Fig. 1.

In Naïve Bayes approach, the conditional probability $P(class|X_1, X_2,\ldots,X_N)$ is factorized as $P(class|X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(class|X_i)$. This factorisation is more easily obtained from experimental data, in addition to being easier to analyze and, later on, to obtain inferences from available information.

2.2. Tree Augmented Network, TAN

This structure is an extension of a Bayesian classifier in which each variable is allowed to have another parent outside the class node. The idea is to build a Bayesian network tree for all predictive variables and complete the model with a Naïve Bayes. TAN algorithm forms a tree with the predictive variables and then add edges to the class node. Fig. 2 is a conceptual illustration of how the model works. It is seen that each predictive variable, X_i, can have up to two parents.

2.3. Forest Augmented Network, FAN

An important limitation that TAN model may have, *a priori*, is that some arcs of the tree, formed between the descriptive variables, can introduce noise into the classification if such relationships do not exist. This model proposes the formation of disjoint trees with predictive variables. So this approach creates a tree structure with all the variables. Obviously many of the dependencies are enforced by the method of construction and do not exist. These dependencies are discarded during the process of creating the tree; an edge is discarded if it is independent (for example, using a statistical χ^2test.). Fig. 3 shows a model built from two disjoint tree predictors.

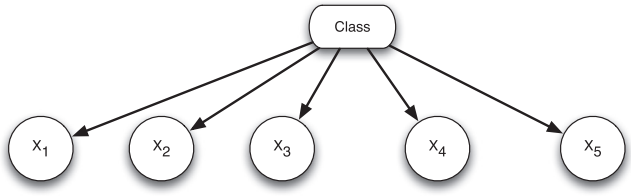

Fig. 1. Structure of a Naïve network.

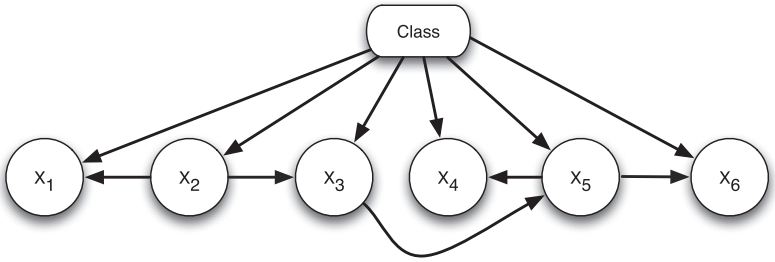

Fig. 2. Structure of a TAN network.

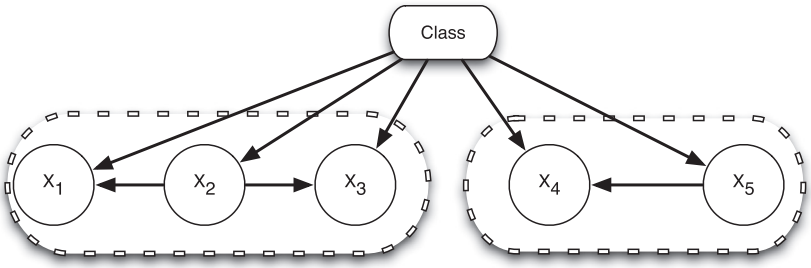

Fig. 3. Structure of a FAN network.

3. Data used, methodology and results

The implemented Clinical Decision Support System is based on a Bayesian Network model trained with a dataset of 1164 cases.

3.1. Dataset

The dataset was collected from patients incoming to the emergency unit of the Hospital Clínic from Valencia, Spain. Each case of the dataset is formed by 17 input variables corresponding to different characteristics and symptoms of the patients, and an observed (output) variable, which corresponds to the fact of suffering an unstable angina. The input variables were: age and gender of the patient; Patient Symptoms: effort related pain (EffortPain), two or more pain episodes in the last 24 h (x2pain24); and the Risk Factors of the patient: creatinine level, smoker, arterial hypertension (HTA), history of hypercholesterolemia (antcolest), diabetes mellitus (DM), family history of ischaemic heart disease (antfam), history of myocardial infarction (antiam), previous Coronary Stenosis (estenosisc), previous admission due to Heart failure (anticc), prior coronary angioplasty (antactp), prior coronary surgery (antcir), peripheral arteriopathy (antperif), and cerebral ictus. Most of the input variables are binary, and hence they are converted to discrete nodes of the Bayesian Network directly. Only two variables are continuous (age and creatinine) and therefore have been converted to discrete nodes before being used by the model. The age has been discretised into intervals of ten years (less than 10 years, from 10 to 19 years, and so on) and the creatinine level has been discretised into two values (less than 1.7 mg/dl and greater than this value), following a clinical criterion. The observed variable of the model is the event of a heart attack within 30 days (event30). The distribution of the values on the training dataset is summarised in Tables 1 and 2. All the cases were used for training the BN model.

Table 1
Distribution of discrete variables.

Variable	Positive (%)	Negative (%)
EffortPain	41.75	58.25
x2pain24	62.80	37.20
var_n	33.51	66.49
Smoker	76.29	23.71
HTA	40.64	59.36
antcolest	46.74	53.26
DM	73.28	26.72
antfam	89. 52	10.48
antiam	76.98	23.02
estenosisc	77.92	22.08
anticc	98.11	1.89
antactp	88.92	11.08
antcir	93.30	6.70
artperif	95.27	4.73
ictus	93.38	6.62
event30 (observed variable)	80.24	19.76

Table 2
Distribution of continuous variables.

Variable	Mean	Std	Min	Max
Age (years)	63.6735	11.4010	30	92
Creatinine (mg/dl)	1.0623	0.5652	0.0	11.1

3.2. Methodology and results

The above-mentioned Bayesian models (Naïve Bayes, TAN structure and FAN structure) were developed with the BNT Toolbox for MATLAB (Murphy, 2007). This toolbox can learn automatically the network structure from the data and construct a directed acyclic graph (DAG), which is represented as an adjacency matrix. Afterwards, the automatically created structures were fine-tuned by the expert, who added some arcs to the graph: a relationship from *HTA* to *anticc* and *ictus*, and another one from *DM* to *creatinine*. The resulting structure, named Medical FAN, is shown in Fig. 4. Finally, the four models were compared in order to find the best fit to the data. The Naïve and TAN models were proved to generate the same results. So that, only the Naïve, FAN and Medical FAN analysis are shown here. First of all, we compare in Table 3 the correlation between the outputs from the three models by measuring the correlation coefficient R^2 (it is worth noting that

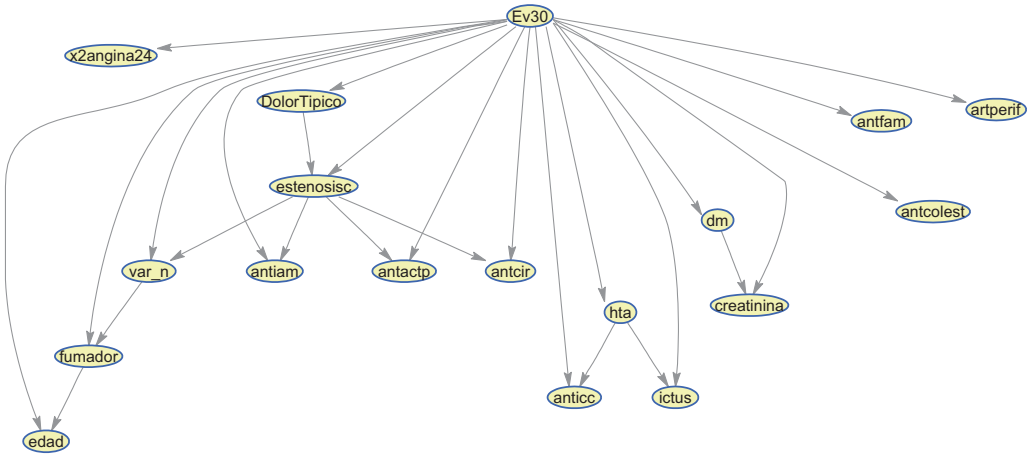

Fig. 4. Bayesian Network to predict unstable angina probability.

Table 3
Correlation between the values predicted by the implemented Bayesian Network Structures.

R^2	FAN	Medical FAN
Naïve Bayes	0.8571	0.8400
FAN		0.9837

Table 5
Area Under the Curve (IC 95%). Bold indicates the network with the best performance.

	AUC
Naïve Bayes	0.7298 ± 0.034
FAN	0.7473 ± 0.034
Medical FAN	**0.7498 ± 0.034**

the output of the models is a probability value). Another measure of the model performance, shown in Table 4, is the Cohen kappa coefficient; it is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative (categorical) items (Carletta, 1996).

The last measure of the model performance taken into account is the Area Under the Curve (AUC), which reflects the percentage of correct classification (an AUC of 1 corresponds to a perfect classification for all the cases, while a AUC of 0.5 indicates a random classification, with a 50% chance of being correct for each class) Fawcett (2004). Table 5 shows the AUC of the three models. The table reflects a similar behaviour of the three models regarding this parameter.

After considering the similar performance of the three models, we chose the medical FAN model for deploying the final Decision Support System. Then, the directed acyclic graph (DAG) of the medical FAN model was converted to the Netica DNET file format for expressing BN structures. Finally, the Conditional Probability Distribution (CPD) of each node was learned from the dataset using a C program and included in the DNET file for the model. The performance of the final model over the training is summarised in Table 6. The implemented CDSS uses this BN model to calculate an output for each given set of evidences. The output of the model is a continuous probability value, which is discretised into a binary value (high risk/low risk) by using a threshold that optimises the Negative Predictive Value.

The model has been validated with a new dataset formed by 103 new cases, collected in the same emergency unit on December of 2011. The performance of the model for this new dataset is summarised in Table 7.

Table 6
Performance of the BN model over the training data.

Sensitivity	0.6957
Specificity	0.6799
Positive predictive value	0.3486
Negative predictive value	0.9007
Accuracy	0.6830

Table 7
Performance of the BN model over the validation data.

Sensitivity	0.8710
Specificity	0.6111
Positive predictive value	0.4909
Negative predictive value	0.9167
Accuracy	0.6893

Table 4
Bayesian Network decision agreement. Table values correspond to the classification threshold of each model, using the same classification criteria.

Kappa	FAN (cut 0.20631)	Medical FAN (cut 0.2084)
Naïve Bayes (cut 0.1971)	0.8056 ± 0.0480	0.8139 ± 0.0482
FAN		0.9800 ± 0.0514

4. Web application

The authors have deployed a Clinical Decision Support System based on Bayesian networks as a web application. This tool evaluates the risk of heart attack on incoming patients with chest pain in an emergency unit, based on the evidences from the Patient Symptoms and his clinical history. The system consists of a web front-end with an input form for introducing and evaluating new clinical cases, a probabilistic inference motor based on a Bayesian Network (BN), and a web administration panel for case reviewing and validation. The tool is used as follows: The user (a clinician) fills in the form with the evidences from the incoming patient (symptoms and clinical history); then the application processes the data

using the developed BN model and infers a probability of suffering a heart attack, which is classified and then presented to the user as "low", "moderate" or "high". Additionally, registered users on the application can save the input case data in order to validate it afterwards (i.e., indicate if the prediction was correct or wrong). Newly validated cases can be eventually used to update the BN beliefs (by updating the CPD of the input nodes).

The tool has been implemented using standard HTML technologies (versions XHTML 1.0 and CSS 3). Dynamic behaviour has been programmed in PHP language on the server side and Javascript on the client side (using the JQuery library). Data is stored on a MySQL database on the server. The inference motor uses a Bayesian Network model: the graph structure of the network has been developed using the BNT library for MATLAB, and the parameter learning and online probabilistic inference is done through a command line program written in C language using the Netica API for BN (Norsys Software Corp., Vancouver, BC, Canada). This subsection details the implementation of the web application, divided into three blocks: application front-end or user interface, application back-end or server-side logic, and probabilistic inference motor.

4.1. Application front-end

The web application that implements the CDSS presents two views: the input case view and the administration panel view. Through the input case view, any user can evaluate an incoming patient by filling in an online form and sending the data to the server. The server calculates the probability of heart attack with the inference motor, and returns the classified risk to the webpage. The administration panel view is used by registered users to review and validate past cases. The tool tracks all the introduced cases by assigning a random alphanumeric reference to each case. The layout of the input case view consists of a collapsible login form on the top and the main input form for evaluating the cases (Fig. 5). The main form requests some evidences from the patient–corresponding to the nodes of the BN graph described in Section 3.1 – grouped into three different sections: *Patient data*,

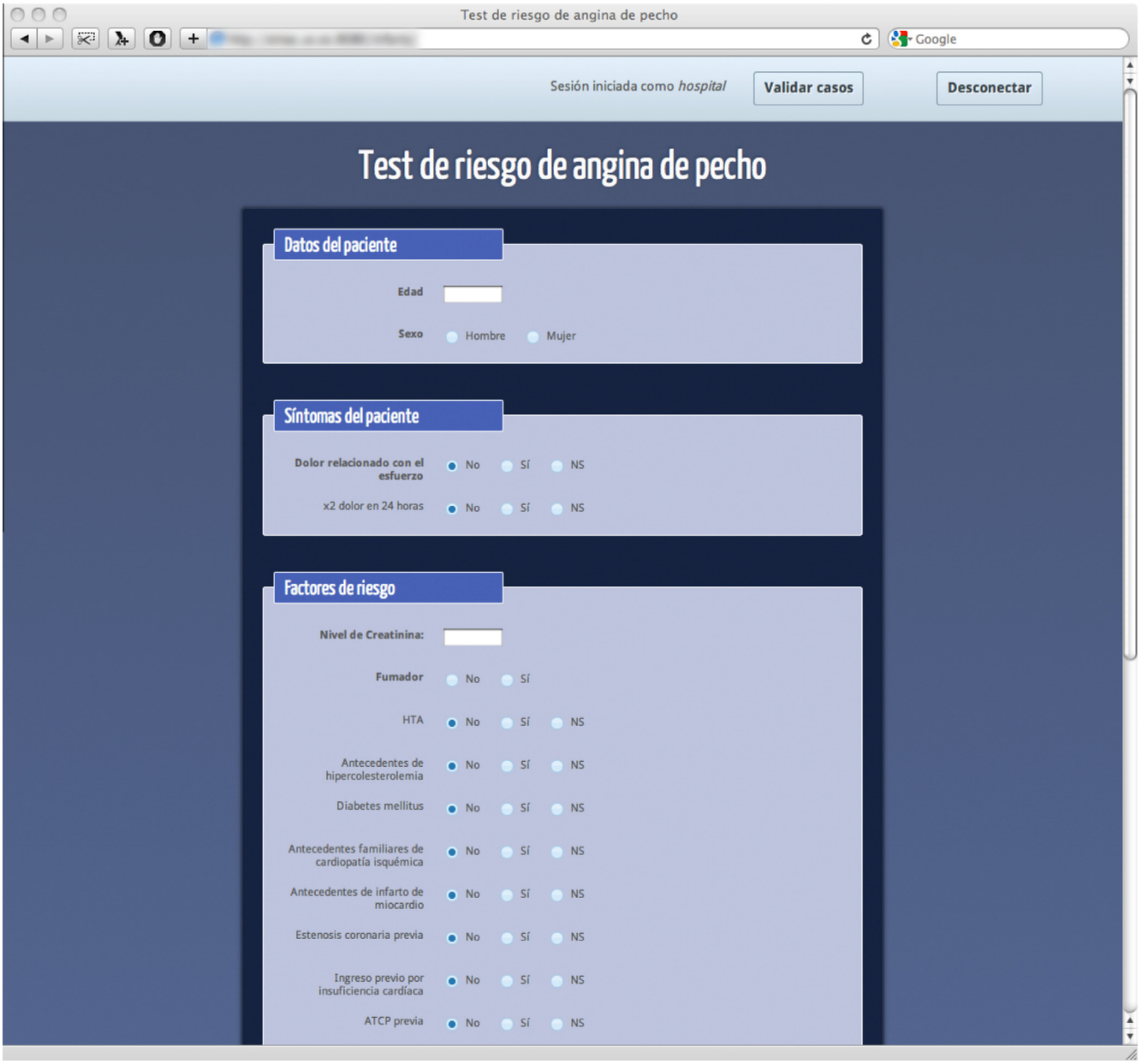

Fig. 5. Application front-end: user input form.

Patient Symptoms, and *Risk Factors*. Only a few fields, which are the most relevant for the BN model, are compulsory (age, gender, effort-related pain, creatinine level and smoker). The rest of the fields can be left empty because the probabilistic inference motor can deal with sparse data (partially empty input data patterns). Once the form is filled in, it is submitted to the server through an AJAX call, and the prediction estimated by the inference motor is shown without reloading into a section of the webpage which lays below the main form.

The administration panel can be accessed only by registered users (Fig. 6). The system allows three different user roles: administrator, coordinator and clinician. Depending on the role of the user, the panel allows different administration actions. Clinicians are assigned to a single centre (i.e., an Emergency Unit of a Hospital), and they can only review their own patient cases. There is one coordinator per centre, who can validate cases and create new users for his centre. Finally, the administrator can create users and centres, and validate cases in all the centres.

4.2. Application back-end

The main application logic is written in PHP language. This logic processes the web forms, stores and retrieves the cases into/from the MySQL database, manages the application users and interfaces with the inference motor to calculate the probability of each case. All the interactions between the application front-end (web page) and the application logic are AJAX-based, which means that no page reloading is needed for the whole process. The probability inference is carried out by a command-line program on the server, which is executed from a PHP script passing the evidences as arguments.

4.3. Probabilistic Inference motor

The tool uses a Bayesian Network to evaluate the risk of suffering a heart attack. The BN is defined by its graph structure and the parameters of the model, defined as the Conditional Probability

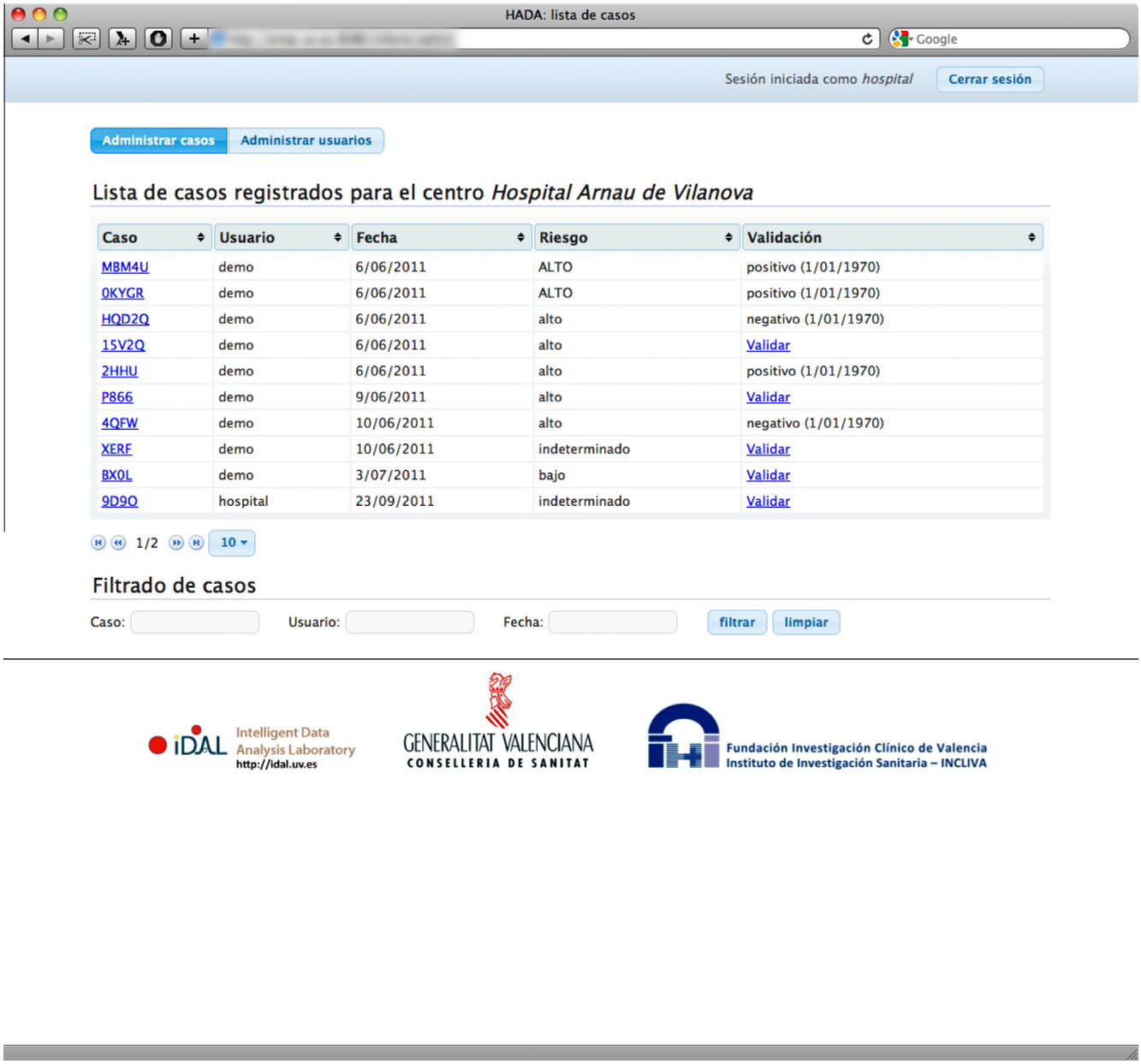

Fig. 6. Application front-end: administration panel.

Distribution (CPD) of each node. In this model, all the variables are discrete, and therefore the CPD are represented by a table of Conditional Probabilities (CPT). Both the structure of the BN and the CPT of each node is stored on a single model file using the Netica DNET format. The tool runs a probabilistic inference through this model from a command-line program executed on the server, which is written in C language using the Netica API library for BN. The application back-end, upon a evaluation request from the user, calls the inference motor as a PHP command line script, passing the evidences as arguments, and receives the estimated probability as an output from the command. This probability is classified into a risk level using the optimal threshold. Eventually, the application back-end returns a pre-formatted web snippet to the front-end showing the evaluation result.

5. Conclusions

This paper presents a Clinical Decision Support System (CDSS) that helps the clinicians in the evaluation of incoming emergency patients with unspecific chest pain that are on risk of a heart attack. A correct diagnosis of these patients, which is difficult to achieve with the standard evaluation procedure (ECG and troponin measurements), could reduce the number of unnecessary hospitalisations. The CDSS is based on a Bayesian Network (BN) model that takes into account 17 different characteristics of the patient. The model uses a FAN network structure with discrete nodes. The BN structure was created in MATLAB while the CPD tables of the nodes were learned using the Netica API for BN. A dataset of 1164 cases was used to train the model, which was validated with a new dataset of 103 cases. The model has been optimised to maximise the negative predictive value, reducing therefore the cases of undetected risk situations. In training, the model achieves a 90.07% of NPV, which increases to a 91.67% for the validation dataset.

The CDSS runs as a web application with role-based access conatrol, with three different roles with decreasing privileges (administrator, coordinators and clinicians). The application consists on an input form and an administration panel. The form is used to introduce the characteristics of the patient under evaluation and returns a classified risk of heart attack from low to high. The administration panel is used to review and validate the registered cases in a per-hospital basis.

References

Steyerberg, E. (2009). *Clinical prediction models: A practical approach to development, validation and updating*. Springer.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity, Tech. rep., McKinsey Global Institute.

Escaned, J., Roig, E., Chorro, F., Teresa, E. D., Jiménez, M., de Sá, E. L., et al. (2008). Ámbito de actuación de la cardiología en los nuevos escenarios clínicos. documento de consenso de la sociedad espa nola de cardiología. *Revista Espa nola de Cardiología, 61*, 170–184.

Sanchis, J., Bodı, V., Bertomeu, V., Gómez, C., Consuegra, L., Bosch, M., et al. (2006). Usefulness of early exercise testing and clinical risk score for prognostic evaluation in chest pain units without preexisting evidence of myocardial ischemia. *American Journal of Cardiology, 97*, 633–635.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT Press.

Lucas, P., Gaag, L., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine, 30*, 201–214.

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian artificial intelligence*. CRC Press.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Rusell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Prentice Hall.

Murphy, K. (2007)The Bayes net toolbox for matlab. *Computing Science and Statistics, 33*. URL: <http://www.cs.ubc.ca/murphyk/Software/BNT/bnt.html>.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers, Tech. rep., HP Laboratories.