



Estimation of cardiovascular patient risk with a Bayesian network

^{*,**} Jan Bohacik¹, ^{**} Darryl N. Davis

^{*}University of Žilina, Faculty of Management Science and Informatics, Department of Informatics,
Univerzitná 8215/1, 01026 Žilina, Slovakia, {Jan.Bohacik}@fri.uniza.sk

^{**}University of Hull, Faculty of Science, Department of Computer Science, Cottingham Road, HU6 7RX,
United Kingdom, {d.n.davis, J.Bohacik}@hull.ac.uk

Abstract. Cardiovascular decision-making support experiences increasing research interest of scientists. Ongoing collaborations between clinicians and computer scientists are looking at the application of data mining techniques to the area of individual patient diagnosis, based on clinical records. An investigation of a Bayesian network learnt according to a generated decision tree with cardiovascular data for estimation of patient risk in cardiovascular domains is presented. Promising experimental results are also provided.

Keywords: classification, cardiology, Bayesian networks, medical data mining.

1. Introduction

A major challenge facing healthcare organizations (hospitals, medical centers) is provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are cost-effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Healthcare organizations must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems [5].

The research reported in this paper considers assessing the risk of individual patients. For assessing the risk of a cardiovascular patient, computational classification models can be used. Classification models are typically used in data mining which is one of the steps of the more general knowledge discovery in databases [3]. As a database, a collected cardiovascular dataset is used by us. Knowledge discovery in the cardiovascular dataset provides knowledge and tools of use for prediction of a cardiovascular patient's risk and its improvement.

The paper is organized as follows. In Section 2, the used cardiovascular dataset and the used clinical model are described. Our designed Bayesian network as a classification model is described in Section 3. Section 4 contains the details of our experiments. Section 5 concludes this paper.

2. Cardiovascular dataset

As a dataset, a group of 839 instances (cardiovascular patients) classified into two levels of risk and described by 17 attributes A as queries about patients' symptoms, medical history, clinical findings and results of physiological measurements is used. Instances are derived from clinical data collected at the Hull site (498 instances) and at the Dundee site (341 instances). This data is noisy, contains many null values and is problematic. It was transformed into the used dataset according to [2]. Describing attributes A are defined as $A = \{ A_1; \dots; A_k; \dots; A_{17} \} = \{ \text{Age}; \text{Sex}; \text{Heart disease}; \text{Diabetes}; \text{Stroke}; \text{Side}; \text{Respiratory}; \text{Renal failure}; \text{ASA}; \text{Hypertension symptom}; \text{ECG}; \text{Duration}; \text{Blood loss}; \text{Shunt}; \text{PATCH}; \text{Coronary artery bypass surgery}; \text{Consultant} \}$. Most describing attributes are categorical with the exception of numerical *Age*, *Duration* and *Blood loss*. If A_k is a

¹ Ján Boháčik with all Slovak diacritics.

categorical attribute $A_k = \{a_{k,1}; \dots; a_{k,l}; \dots; a_{k,l_k}\}$ where $a_{k,1}; \dots; a_{k,l}; \dots; a_{k,l_k}$ are possible categorical values. $A_2 = \{a_{2,1}; a_{2,2}\} = \{female; male\}$, $A_3 = \{a_{3,1}; a_{3,2}\} = \{no; yes\}$, $A_4 = \{a_{4,1}; a_{4,2}\} = \{no; yes\}$, $A_5 = \{a_{5,1}; a_{5,2}\} = \{no; yes\}$, $A_6 = \{a_{6,1}; a_{6,2}\} = \{left; right\}$, $A_7 = \{a_{7,1}; a_{7,2}; a_{7,3}; a_{7,4}\} = \{normal; mildCOAD; modCOAD; severeCOAD\}$, $A_8 = \{a_{8,1}; a_{8,2}\} = \{no; yes\}$, $A_9 = \{a_{9,1}; a_{9,2}; a_{9,3}; a_{9,4}\} = \{one; two; three; four\}$, $A_{10} = \{a_{10,1}; a_{10,2}\} = \{no; yes\}$, $A_{11} = \{a_{11,1}; a_{11,2}; a_{11,3}; a_{11,4}; a_{11,5}; a_{11,6}; a_{11,7}\} = \{normal; qWaves; sTWaves; aFib60to90; aFibLT90; fiveEctopic; other\}$, $A_{14} = \{a_{14,1}; a_{14,2}\} = \{no; yes\}$, $A_{15} = \{a_{15,1}; a_{15,2}; a_{15,3}; a_{15,4}; a_{15,5}; a_{15,6}; a_{15,7}\} = \{armVein; legVein; otherVein; dacron; no; ptfe; stent\}$, $A_{16} = \{a_{16,1}; a_{16,2}\} = \{no; yes\}$, $A_{17} = \{a_{17,1}; a_{17,2}; a_{17,3}; a_{17,4}; a_{17,5}\}$. Class attribute C ($= Risk$) is used to classify instances into two possible categorical values c_1 and c_2 meaning risk levels (*low* and *high*, respectively). It is denoted by $C = Risk = \{c_1; c_2\} = \{low; high\}$. The values of class attribute C are generated according to the following heuristic clinical model [2]: an instance (cardiovascular patient) is classified into *high* if the patient's death or severe cardiovascular event (e.g. stroke, myocardial relapse or cardiovascular arrest) appears within 30 days after an operation.

3. Bayesian network based on a generated C.45 decision tree

Our Bayesian network is learnt on the basis of a decision tree. The decision tree is generated according to the C4.5 algorithm [6] used on collected data for all 17 describing attributes A_k , $k=1, \dots, 17$, class attribute C and 839 instances. The generated decision tree is in Fig. 1. It consists of the root node and the inner nodes (expressed as ellipses) associated with attributes, branches (expressed as lines) associated with possible values of attributes and leaf nodes (expressed as rectangles) associated with risk levels $c_j \in C$.

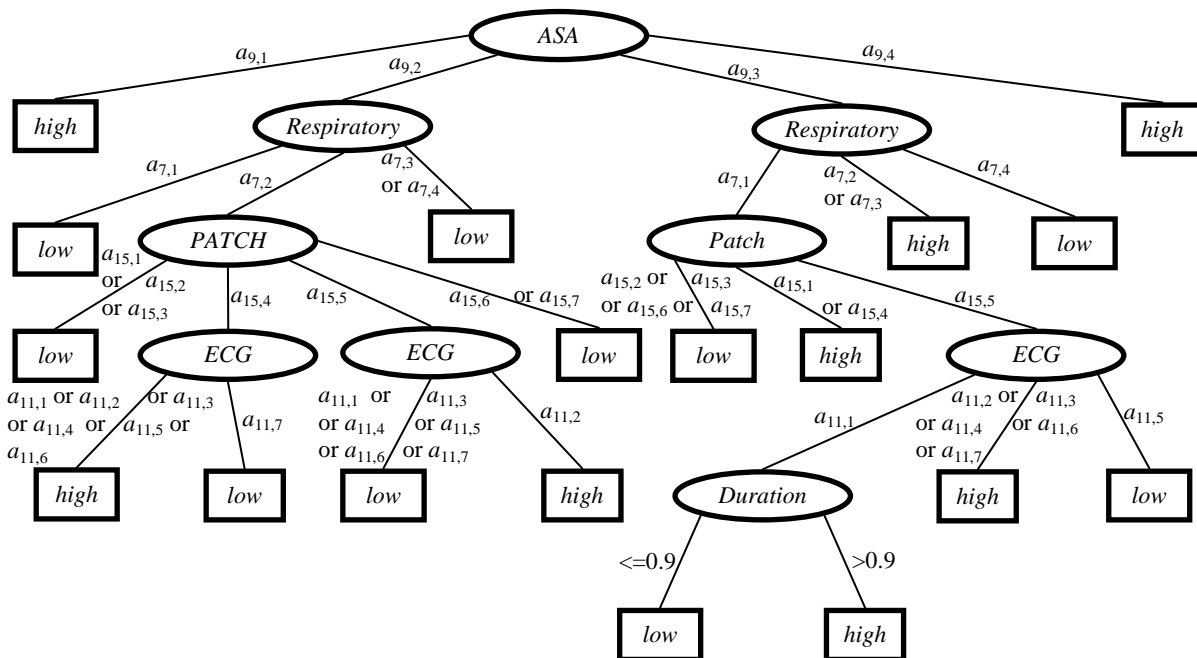


Fig. 1. Generated C4.5 decision tree on the basis of instance values for A_k , $k = 1, 2, \dots, 17$, and C .

The attributes associated with the root node and the inner nodes in Fig.1 are used as nodes in our Bayesian network in Fig. 2. Basically, a Bayesian network is a graph with arcs connecting nodes and no directed cycles (i.e., a directed acyclic graph), whose nodes represent random

variables and whose arcs represent direct dependencies. Each node has a conditional probability table (CPT), which, for each combination of values of the parents, gives the conditional probability of each of its values. If there is a branch from an attribute to another attribute in Fig. 1, there is an arc from the attribute to the other attribute in our Bayesian network.

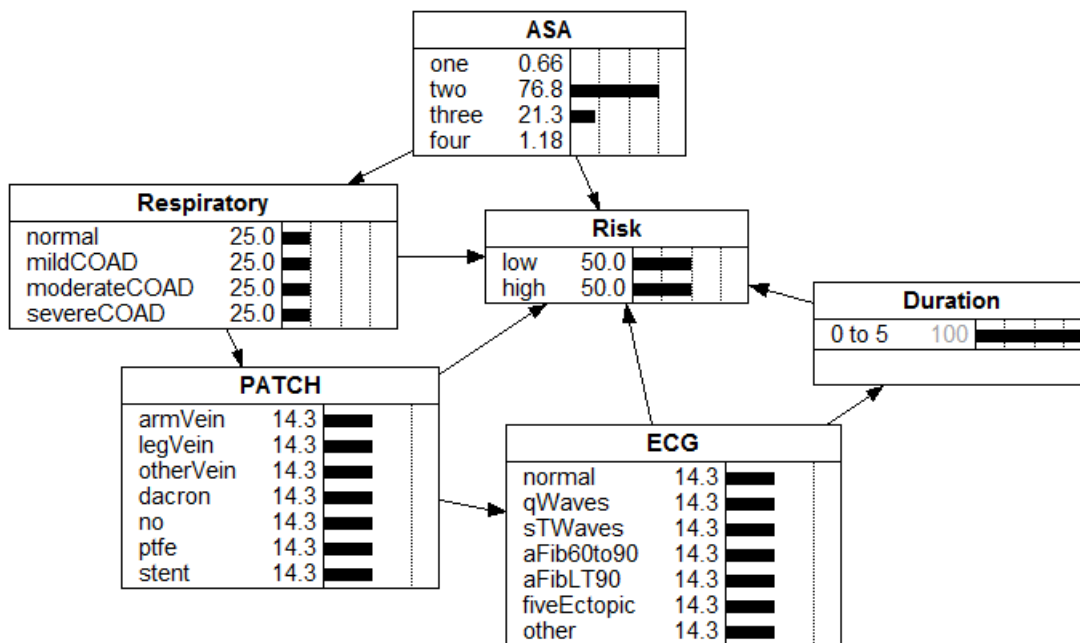


Fig. 2. Proposed Bayesian network.

The CPTs in our Bayesian network are learnt on the basis of data for attributes A_7 , A_9 , A_{11} , A_{12} , A_{15} , class attribute C and 839 instances with software library NeticaTM [4].

4. Experimental results

The main purpose of the experimental study is to compare the performance of our designed Bayesian network with other classification models on our cardiovascular dataset. Experiments were carried out with our Java software tool which is being developed with the intention of its integration into the medical decision-making support system of the BraveHealth system. The core algorithms are implemented in external libraries: NeticaTM [4] and Weka [6]. The performance of the particular classification models is measured with sensitivity = $tp/(tp + fn)$, specificity = $tn/(tn + fp)$, positive predictive value = $tp/(tp + fp)$, negative predictive value = $tn/(tn + fn)$ and accuracy = $(tp + fn)/(tp + fp + fn + tn)$ where tp is true positive, fp is false positive, fn is false negative, tn is true negative, ‘ C is low’ is negative and ‘ C is high’ is positive.

Method	SEN (%)	SPEC (%)	PPV (%)	NPV (%)	ACC (%)
TreeBayesNet	77.78	96.63	80.33	96.09	93.80
Bayes	7.94	97.48	35.71	85.70	84.03
C4.5	4.76	98.60	37.50	85.42	84.51
NNge	15.08	90.18	21.35	85.73	78.90
MLP	15.08	89.62	20.43	85.66	78.43

Tab. 1. Experimental results.

The results of our experiments are given in Tab. 1 where TreeBayesNet denotes our Bayesian network described in Section 4 and implemented with NeticaTM and Weka, Bayes denotes a Bayesian network classifier implemented in Weka as class BayesNet, C4.5 is a decision tree classifier implemented in Weka as class J48, NNge is a nearest neighbor classifier using non-tested generalized exemplars [1] implemented in Weka as class NNge, and MLP is a neural network

classifier using multilayer perceptron implemented in Weka as class MultilayerPerception. SEN is sensitivity, SPEC is specificity, PPV is positive predictive value, NPV is negative predictive value and ACC is accuracy. Since these classification models are considered to be used as a part of the medical decision-making support system of the BraveHealth system, they should avoid cases when high risk patients are labeled low risk and so sensitivity should be maximized. Our proposed classification model TreeBayesNet learnt from a generated C4.5 decision tree gives the highest sensitivity 77.78% of all classification models. Its accuracy with 93.80% is also the highest.

5. Conclusions

A classification model based on a Bayesian network learnt from a generated decision tree is proposed. It is employed together with several other essentially different classification models on data which is collected about cardiovascular patients. Our model gives considerably better results, especially with its avoidance of labeling high risk patients as low risk patients. However, further investigation, including a simultaneous use of more classification models, continues so that even fewer high risk patients are labeled as low risk ones. The aim of our study was to investigate/develop issues and software capable of being integrated into the BraveHealth system which will provide a patient centric approach for an integrated, adaptive, context aware remote diagnosis and management of cardiovascular diseases.

Acknowledgement

This work is funded by the European Commission's 7th Framework Program: BRAVEHEALTH FP7-ICT-2009-4, Objective ICT-2009.5.1: Personal Health Systems: a) Minimally invasive systems and ICT-enabled artificial organs: a1) Cardiovascular diseases.

References

- [1] BRENT, M.: *Instance-Based Learning: Nearest Neighbour With Generalization*. Master's thesis. Waikato:University of Waikato, 1995.
- [2] DAVIS, D. N., NGUYEN, T. T.: Chapter IX: Generating and verifying risk prediction models using data mining: A case study from cardiovascular medicine. *Data Mining and Medical Knowledge Management: Cases and Applications*. 1st ed. :IGI Global Inc., 2009, ISBN10: 1605662186.
- [3] FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P.: From data mining to knowledge discovery in databases. *AI Magazine*, Vol. 17, No. 3, pp. 37-54, 1996, ISSN: 0738-4602.
- [4] NORSYS SOFTWARE CORP.: *Netica™ Application* [<http://www.norsys.com/netica.html>].
- [5] PALANIAPPAN, S., AWANG, R.: Intelligent heart diseases prediction system using data mining techniques. *Int. Journal of Computer Science and Network Security*, Vol. 8, No. 8, pp. 343-350, 2008, ISSN: 1738-7906.
- [6] WITTEN, I. H., FRANK, E., HALL, M. A.: *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. :Morgan Kaufmann, 2011, ISBN: 978-0-12-374856-0.