

Data mining applied to cardiovascular data

*Jan Bohacik**, Department of Computer Science FS UH in Hull, United Kingdom and
Katedra informatiky FRI ŽU in Žilina, Slovakia

Darryl N. Davis, Department of Computer Science FS UH in Hull, United Kingdom

Abstract: Medical decision support is one area of increasing research interest. Ongoing collaborations between cardiovascular 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 four different classification models on cardiovascular data for estimation of patient risk in cardiovascular domains is presented. Experimental results are provided showing the performance of particular models.

Key words: classification, medical data mining, cardiology

1. Introduction

A major challenge facing healthcare organizations (hospitals, medical centers) is the 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 (8).

The research reported in this paper considers assessing patient risk in the domain of cardiovascular medicine. No gold standard exists for 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 (4). 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.

*Ján Boháčik with all Slovak diacritics.

The paper is organized as follows. In Section 2, cardiovascular dataset and clinical models are described. Classification models employed in our experiments are 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 (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 (3). The description of instances and their summary can be seen in Table. 1. Describing attributes A are defined as $A = \{A_1; \dots; A_k; \dots; A_{17}\}$. If A_k is a 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. Class attribute C is used to classify instances into two possible categorical values c_1 and c_2 meaning risk levels (*high* and *low*). It is denoted by $C = \{c_1; c_2\}$.

Table. 1: Cardiovascular dataset

Attribute	Data Type	Value Range	Frequency
Age (A_1)	Numerical	38 - 93	N/A
Sex (A_2)	Categorical	<i>female</i> ($a_{2,1}$)	332
		<i>male</i> ($a_{2,2}$)	507
Heart disease (A_3)	Categorical	<i>no</i> ($a_{3,1}$)	488
		<i>yes</i> ($a_{3,2}$)	351
Diabetes (A_4)	Categorical	<i>no</i> ($a_{4,1}$)	749
		<i>yes</i> ($a_{4,2}$)	90
Stroke (A_5)	Categorical	<i>no</i> ($a_{5,1}$)	567
		<i>yes</i> ($a_{5,2}$)	272

<i>Side (A₆)</i>	Categorical	<i>left (a_{6,1})</i>	458
		<i>right (a_{6,2})</i>	381
<i>Respiratory disease (A₇)</i>	Categorical	<i>normal (a_{7,1})</i>	729
		<i>mild COAD (a_{7,2})</i>	92
		<i>mod COAD (a_{7,3})</i>	18
		<i>severe COAD (a_{7,4})</i>	2
<i>Renal failure (A₈)</i>	Categorical	<i>no (a_{8,1})</i>	827
		<i>yes (a_{8,2})</i>	12
<i>ASA grade (A₉)</i>	Numerical	1 - 4	N/A
<i>Hypertension symptom (A₁₀)</i>	Categorical	<i>no (a_{10,1})</i>	384
		<i>yes (a_{10,2})</i>	455
<i>ECG (A₁₁)</i>	Categorical	<i>normal (a_{11,1})</i>	604
		<i>q waves (a_{11,2})</i>	74
		<i>afib 60-90 (a_{11,3})</i>	16
		<i>other abnormal rhythm (a_{11,4})</i>	16
		<i>st/t wave changes (a_{11,5})</i>	35
		<i>>= 5 ectopics/min (a_{11,5})</i>	2
		<i>afib <90 (a_{11,6})</i>	7
<i>Duration (A₁₂)</i>	Numerical	0.7 - 5	N/A
<i>Blood loss (A₁₃)</i>	Numerical	0 - 2000	N/A
<i>Shunt (A₁₄)</i>	Categorical	<i>no (a_{14,1})</i>	338

		<i>yes</i> ($a_{14,2}$)	501
<i>Patch</i> (A_{15})	Categorical	<i>arm vein</i> ($a_{15,1}$)	3
		<i>dacron</i> ($a_{15,2}$)	185
		<i>leg vein</i> ($a_{15,3}$)	4
		<i>no</i> ($a_{15,4}$)	325
		<i>other vein</i> ($a_{15,5}$)	150
		<i>ptfe</i> ($a_{15,6}$)	171
		<i>stent</i> ($a_{15,7}$)	1
<i>Coronary artery bypass surgery</i> (A_{16})	Categorical	<i>no</i> ($a_{16,1}$)	787
		<i>yes</i> ($a_{16,2}$)	52
<i>Consultant</i> (A_{17})	Categorical	<i>a</i> ($a_{17,1}$)	237
		<i>b</i> ($a_{17,2}$)	114
		<i>c</i> ($a_{17,3}$)	102
		<i>d</i> ($a_{17,4}$)	383
		<i>e</i> ($a_{17,5}$)	3
<i>Risk1/Risk2</i> (C)	Categorical	<i>high</i> (c_1)	126/139
		<i>low</i> (c_2)	713/700

The values of class attribute C are generated according to two heuristic clinical models (3). Class attribute *Risk1* is used for one; *Risk2* is used for the other. For *Risk1*, an instance (patient) is classified into *high* if this patient dies within 30 days after an operation. Otherwise, the instance is classified into *low*. For *Risk2*, an instance (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. Classification models

Given a cardiovascular dataset of instances (patients) V where each instance is described by attributes $A = \{A_1; \dots; A_k; \dots; A_{17}\}$ and classified into a class $c_j \in C$, the task is to build a classification model that predicts class $c_j \in C$ of an (unseen) instance. Four different classification methods, each based on a different principle, were developed for building classification models: Bayes Network Classifier; Decision Tree Classifier; Nearest Neighbor Classifier; and Neural Network Classifier.

A Bayes Network Classifier is based on a Bayesian network which represents a joint probability distribution over a set of categorical attributes (suppose attributes in A are categorical or discretized). It consists of two parts $B = \langle G, \theta \rangle$, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables $\theta = (\theta_{A_1}, \dots, \theta_{A_{17}})$. The nodes represent attributes in A and attribute C whereas the arcs indicate direct dependencies. The graph G then encodes the independence relationships of the domain. The Bayesian network allows the computation of the (joint) posterior probability distribution of any subset of unobserved assignments of values to attributes in A . This functionality makes it possible to use for determination of $c_j \in C$ by applying the winner-takes-all rule to the posterior probability distribution for the (unobserved) node representing C (1).

A Decision Tree Classifier consists of a decision tree generated on the basis of instances in V . The decision tree has two types of nodes: a) the root and the internal nodes, b) the leaf nodes. The root and the internal nodes are associated with attribute $A_k \in A$. Leaf nodes are associated with class $c_j \in C$. Basically, each non-leaf node has an outgoing branch for each possible value $a_{k,l} \in A_k$ where $A_k \in A$ is an attribute associated with the node. To determine $c_j \in C$ for a new instance using a decision tree, beginning with the root, successive internal nodes are visited until a leaf node is reached. At the root node and at each internal node, a test is applied. The outcome of the test determines the branch traversed, and the next node visited. The class for the instance is simply class $c_j \in C$ of the final leaf node (5).

A Nearest Neighbor Classifier assumes all instances correspond to points in the n -dimensional space R^n . During learning, all instances (points) in V are remembered. When a new point is classified, the k -nearest points to the new point are found and are used with a weight for determining the class value $c_j \in C$ of the new point. For the sake of increasing accuracy, greater weights are given to closer points (6). For large k , the algorithm is computationally

more expensive than a decision tree or a neural network and requires efficient indexing. When the cardinality of V is high, a reduction of its number is required.

A Neural Network Classifier is based on neural networks consisting of interconnected neurons. From a simplified perspective, a neuron takes positive and negative stimuli (numerical values) from other neurons and when the weighted sum of the stimuli is greater than a given threshold value, it activates itself. The output value of the neuron is usually a non-linear transformation of the sum of stimuli. In more advanced models, the non-linear transformation is adapted by some continuous functions. Neural networks are good for classification. They are an alternative to decision trees when understandable knowledge is not required; this can be a disadvantage in medical domains where decisions need to be validated. A further disadvantage is that it is sometimes difficult to determine the optimal number of neurons (7).

4. Experimental results

The main purpose of the experimental study is to compare the performance of different classification models on our cardiovascular dataset. Classification models are implemented in object-oriented Java software tool Weka (9). 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 = $(tp + tn)/(tp + fp + fn + tn)$ where tp is true positive, fp is false positive, fn is false negative, tn is true negative, C is *high* is positive and C is *low* is negative.

Table. 2: Experimental results

Method	Output	SEN (%)	SPEC (%)	PPV (%)	NPV (%)	ACC (%)
Bayes	<i>Risk1</i>	7.94	97.48	35.71	85.70	84.03
	<i>Risk2</i>	7.91	96.86	33.33	84.12	82.12
C4.5	<i>Risk1</i>	4.76	98.60	37.5	85.42	84.51
	<i>Risk2</i>	2.16	99.14	33.33	83.61	83.08
NNge	<i>Risk1</i>	15.08	90.18	21.35	85.73	78.90
	<i>Risk2</i>	12.95	89.00	18.95	83.74	76.40

MLP	<i>Risk1</i>	15.08	89.62	20.43	85.66	78.43
	<i>Risk2</i>	12.95	88.00	17.65	83.58	75.57

The results of our experiments are given in Table. 2 where 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-nested generalized exemplars (2) implemented in Weka as class NNGe, MLP is a Neural Network Classifier using multilayer perceptron implemented in Weka as class MultilayerPerceptron. Column Output distinguishes the results for two clinical models described in Section 2. SEN is sensitivity, SPEC is specificity, PPV is positive predictive value, NPV is negative predictive value and ACC is accuracy. Since these algorithms are considered to be used as a part of a medical decision support system which should avoid cases when high risk patients are labeled low risk, sensitivity should be maximized. NNge and MLP with sensitivity 15.08% (12.95%) give the best results. On the other hand, the best accuracy 84.51% (83.08%) is achieved with classification model C4.5.

5. Conclusions

Several essentially different classification models were employed on cardiovascular data. These models are useable, however, among other problems, labeling high risk patients as low risk patients in many cases should be avoided. Further investigation, including a simultaneous use of more classification models, continues so that this can be achieved. 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.

6. Bibliography

- (1) Baesens, B., Egmont-Petersen, M., Castelo, R., Vanthienen, J.: Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search, *Proceedings of International Congress on Pattern Recognition* (Published by: IEEE Computer Society; In: Montreal, Canada), Pages: 49-52, Year: 2002.
- (2) Brent, M.: *Instance-Based learning : Nearest Neighbor With Generalization* (Master's Thesis at the University of Waikato, New Zealand), Pages: 76, Year: 1995.

- (3) Davis, D. N., T.T. Nguyen, T.T. T.: Generating and Verifying Risk Prediction Models Using Data Mining: A Case Study from Cardiovascular Medicine, *Chapter of Data Mining and Medical Knowledge Management: Cases and Applications* (Published by: IGI Global Inc.), Year: 2009, ISBN10: 1605662186.
- (4) Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases, *AI Magazine* (Volume: 17, Number: 3), Pages: 37-54, Year: 1996, ISSN: 0738-4602.
- (5) Garofalakis, M., Hyun, D., Rastogi, R., Shim, K.: Building Decision Trees with Constraints, *Data Mining and Knowledge Discovery* (Volume: 7, Number: 2), Pages: 187 – 214, Year: 2003, ISSN: 1384-5810.
- (6) Mitchell, T. M.: *Machine Learning* (Published by: McGraw-Hill Companies; In: USA), Pages: 414, Year: 1997, ISBN: 0070428077.
- (7) Nilson, N. J.: *Introduction to Machine Learning* (Published by: unpublished draft; In: Stanford University, USA), Year: 1996.
- (8) Palaniappan, S., Awang, R.: Intelligent Heart Diseases Prediction System Using Data Mining Techniques, *International Journal of Computer Science and Network Security* (Volume: 8, Number: 8), Pages: 343-350, Year: 2008, ISSN: 1738-7906.
- (9) Witten, I. H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)* (Published by: Morgan Kaufmann), Pages: 525, Year: 2005, ISBN: 0-12-088407-0.

7. Address(es) of the author(s):

Dr Jan Bohacik (Ján Boháčik)
 Department of Computer Science
 Faculty of Science
 University of Hull
 HU6 7RX
 UNITED KINGDOM
 J.Bohacik@hull.ac.uk

Katedra informatiky, FRI
 Žilinská univerzita
 Univerzitná 8215/1
 010 26 Žilina
 SLOVAKIA
 Jan.Bohacik@fri.uniza.sk

Dr Darryl N. Davis
 Department of Computer Science
 Faculty of Science
 University of Hull
 HU6 7RX
 UNITED KINGDOM

d.n.davis@hull.ac.uk

Acknowledgments

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.

Accepted for publication in November 2010 by the publisher of the Journal of Information Technologies (Vol. 3, No. 2, ISSN 1337-7469), i.e. by Katedra aplikovanej informatiky FPV UCM in Trnava, SLOVAKIA.