OSSConf 2012: 1–6

RISK ESTIMATION OF CARDIOVASCULAR PATIENTS USING WEKA

JÁN BOHÁČIK (UK, SK), DARRYL N. DAVIS (UK) AND MIROSLAV BENEDIKOVIČ (SK)

Abstract. Cardiovascular diseases remain the most prevalent cause of deaths worldwide and their prevention requires major life-style changes using limited health-care resources. Remote decision support for cardiovascular patients seems to allow them to lead a productive life and to minimize the costs of treatment. In this paper, risk estimation of cardiovascular patients on the basis of collected data used in our developing decision-making support system is described. The system makes use of some data mining techniques which are implemented in open source software tool Weka - Waikato Environment for Knowledge Analysis. The integration of Weka with our system, a description of used risk estimation models based on data mining techniques, and experimental results showing the performance of these models are also given.

Key words and phrases. Data mining, risk estimation, cardiology, telehealth.

ODHADOVANIE RISKU U KARDIOVASKULÁRNYCH PACIENTOCH S NÁSTROJOM WEKA

Abstrakt. Kardiovaskulárne choroby pretrvávajú ako najdôležitejšia príčina smrti vo svete a ich prevencia si vyžaduje značné zmeny životného štýlu s využitím obmedzených zdrojov pre zdravotnú starostlivosť. Zdá sa, že vzdialená podpora rozhodovania umožňuje kardiovaskulárnym pacientom viesť produktívny život a minimalizovať náklady na liečbu. V tomto článku je popísané odhadovanie risku u kardiovaskulárnych pacientoch použité vo vyvíjanom systéme na podporu rozhodovania na základe zhromaždených údajov. Tento systém využíva niektoré techniky na dolovanie z údajov, ktoré sú implementované v open-source softvérovom nástroji Weka - Waikato Environment for Knowledge Analysis. Rozobraná je aj integrácia nástroja Weka vo vyvíjanom systéme, popis použitých modelov na odhadovanie risku založených na technikách pre dolovanie z údajov, a experimentálne výsledky ukazujúce výkonnosť jednotlivých modelov.

Kľúčové slová. Dolovanie z údajov, odhadovanie risku, kardiológia, telezdravotníctvo.

Introduction

Cardiovascular disease (CVD) is a major cause of disability and premature death throughout the world. Three areas of prevention can be distinguished [6]: a) prevention in the total population; b) prevention in high risk groups; and c) prevention after cardiovascular events. Prevention in high risk groups and prevention after cardiovascular events require major life style changes and medication using

limited health-care resources. As participants of the BraveHealth project, we are interested in continuous and remote monitoring and real time prevention of malignant events for people already diagnosed as subjects at risk. Patients in the project are required to use a wearable unit with sensors and other devices so that regular data can be obtained about them. This is used for decision support on several levels and with several techniques. The work reported in this paper is focused on developing decision support system on the remote server in the BraveHealth project and its data mining techniques for risk estimation of cardiovascular patients. Used open source software solutions are especially discussed.

The paper is organized as follows. Our developing decision-making support system, Weka and its integration with the system are discussed in Section 1. Data mining techniques used from Weka are described in Section 2. Section 3 contains collected cardiovascular data and curried experiments with it. Section 4 concludes this paper.

1. Developing decision support system and Weka

The developing decision support system (DDSS) is intended to provide support for both the clinician and the patient. There is a remote server that receives data about patients in real time. Alert rules [1] and data mining techniques [2,3] are applied on this data. For the purpose of processing, the data is described by attributes which are created according to the classes shown with UML [8] in Figure 1. The data is held in instances created according to the classes shown in Figure 2. The instances are used by data mining techniques and alert rules implemented in the DDSS. Class PilotInstances also allows us to transform the data so that it can be used in data mining tools such as Netica^{TM} [7] and Weka [5].

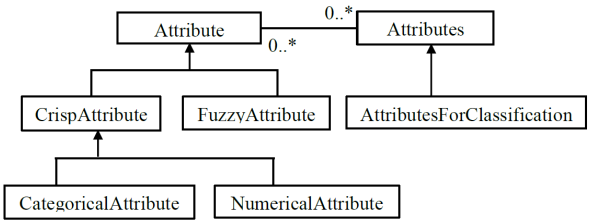

Figure 1. Class diagram for attributes.

Weka (Waikato Environment for Knowledge Learning) is an open source data mining system implemented in Java. Releasing Weka as an open source software and implementing it in Java are two factors which ensure that it remains maintainable and modifiable irrespective of the commitment or health of any particular institution or company. Its aim is to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and

practitioners alike. It can cope with preprocessing and data analysis, classification models, association models, and evaluation metrics. There are three modes of Weka operation: a) GUI, b) commandline and c) Java API [9]. Our developing system uses Java API of Weka for running some data mining techniques so that a part of risk estimation of cardiovascular patients is conducted.

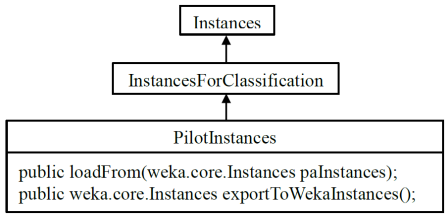

Figure 2. Class diagram for instances.

2. Data mining techniques

Given a set of cardiovascular patients (instances) $\boldsymbol{V}$ where each instance is described by attributes $\boldsymbol{A} = \{A_1; \ldots; A_k; \ldots; A_K\}$ and classified into a class $c_j \in C$, the task is to build a risk estimation model that predicts risk $c_j \in C$ of an (unseen) cardiovascular patient. Attributes in $\boldsymbol{A}$ are categorical and numerical. Four principally different data mining techniques are described here: a Bayes network classifier (BNC), a decision tree classifier (DTC), a neural network classifier (MLP) and a nearest neighbor classifier (NNC) [2].

A BNC is based on a Bayesian network which represents a joint probability distribution over a set of categorical attributes. Since it is a distribution over a set of categorical attributes, numerical attributes in $\boldsymbol{A}$ are discretized and transformed into categorical. It consists of $\langle G; \Theta \rangle$, a directed acyclic graph G consisting of nodes and arcs and conditional probability tables $\Theta = (\theta_{A_1}; \ldots; \theta_{A_K})$. The nodes represent attributes in $\boldsymbol{A}$ and attribute C whereas the arcs indicate direct dependencies. The Bayesian network allows the computation of the (joint) posterior probability distribution of any subset of unobserved assignments of values to attributes in $\boldsymbol{A}$, which makes it possible to use for determination of $c_j \in C$.

A DTC consists of a decision tree which is generated on the basis of instances in $\boldsymbol{V}$. There are two types of nodes in the decision tree: a) the root and internal nodes (associated with an attribute $A_k \in \mathbf{A}$); b) leaf nodes (associated with a $c_j \in C$). Basically, each non-leaf node has an outgoing branch for each possible value $a_{k,l} \in A_k$, $A_k \in \mathbf{A}$ is an attribute associated with the node. Numerical attributes $A_k \in \mathbf{A}$ are discretized. Value $c_j \in C$ is determined 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. Value $c_j \in C$ for the instance is simply c_j of the final leaf node.

A MLP is based on a neural network of interconnected neurons. A neuron takes positive and negative numerical values from other neurons and when the weighted sum of the stimuli is greater than a given threshold value, it activates itself. Its output value is usually a non-linear transformation of the sum of the numerical values. It can also be adapted by some continuous functions.

A NNC assumes cardiovascular patients correspond to points in space R^n. All known cardiovascular patients in $\boldsymbol{V}$ are remembered when the classifier is being made. When $c_j \in C$ for a new cardiovascular patient is being determined, k-nearest known cardiovascular patients to the new one are found and they are used with a weight. Grater points are given to closer points so that accuracy of determining $c_j \in C$ can be increased.

3. Cardiovascular data and experiments

The following cardiovascular data derived from clinical data collected at two clinical sites (the Hull site of 498 instances and the Dundee site of 341 instances) [4] is used. Describing attributes for cardiovascular patients $\boldsymbol{A}$ are defined as $\boldsymbol{A} = \{A_1; \ldots; A_k; \ldots; A_{17}\} = \{$*Age*; *Gender*; *Heart disease*; *Diabetes*; *Stroke*; *Side*; *Respiratory problem*; *Renal failure*; *ASA grade*; *Hypertension*; *ECG*; *Duration*; *Blood loss*; *Shunt*; *Patch*; *Coronary artery bypass surgery*; *Consultant*$\}$.*Age* (A_1) and *Gender* (A_2) represent the age and the gender of the patient. *Heart disease* (A_3), *Diabetes* (A_4), *Stroke* (A_5), *Renal failure* (A_8), *Hypertension* (A_10), *Shunt* (A_{14}), *Coronary artery bypass surgery* (A_16) respectively indicate if any heart disease, diabetes, a stroke, renal insufficiency, a high blood pressure, a shunt, or a coronary artery bypass surgery are present. *Side* (A_6) holds the side of surgery. *ASA grade* (A_9) is used to classify the patient according to the American Society of Anesthesiologists classification. *ECG* (A_{11}) describes electrocardiography, i.e. a transthoracic (across the thorax or chest) interpretation of the electrical activity of the heart over a period of time. *Duration* (A_12) is the duration of surgery in hours. *Blood loss* (A_{13}) represents the blood loss in surgery in milliliters. *Patch* (A_{15}) indicates which material is used for by-pass patching in the patient's surgery. *Consultant* (A_{17}) describes the particular consultant employed for the patient's treatment. Cardiovascular patients are classified into two possible categorical values low (c_1) and high (c_2) meaning risk levels. It is denoted by $C = \{c_1; c_2\}$. The values of C are generated according to the following model [4]: a cardiovascular patient is classified into high if her/his death or severe cardiovascular event appears within 30 days after an operation.

The main purpose of the experimental study is to compare data mining techniques described in the previous section, implemented in Weka [5], and used in

our developing decision making support system for risk estimation of cardiovascular patients. The performance of algorithms is measured with sensitivity = $\frac{\text{tp}}{(\text{tp+fn})}$, specificity = $\frac{\text{tn}}{(\text{tn+fp})}$, positive predictive value = $\frac{\text{tp}}{(\text{tp+fp})}$, negative predictive value = $\frac{\text{tn}}{(\text{tn+fn})}$ and accuracy = $\frac{(\text{tp+tn})}{(\text{tp+fp+fn+tn})}$. In the formulas, tp/fp/fn/tn is the number of true positives/false positives/false negatives/true negatives. "C is *low*" is considered to be negative and "C is *high*" is considered positive. Values tp, fp, fn, tn are computed during 10-fold cross-validation.

Model	SEN (%)	SPEC (%)	PPV (%)	NPV (%)	ACC (%)
BNC	7.94	97.48	35.71	85.70	84.03
DTC	4.76	98.60	37.50	85.42	84.51
MLP	15.08	89.62	20.43	85.66	78.43
NNC	15.08	90.18	21.35	85.73	78.90

Table 1. Experimental results with Weka.

The results of our experiments are given in Table 1. BNC denotes a Bayesian network which is implemented in Weka as class BayesNet. DTC is a decision tree classifier implemented in Weka as class J48. MLP is a neural network classifier implemented in Weka as class MultilayerPerception. NNC is a nearest neighbor classifier implemented in Weka as class NNGe. SEN is sensitivity, SPEC is specificity, PPV is positive predictive value, NPV is negative predictive value, ACC is classification accuracy. From the point of view of risk estimation of cardiovascular patients, sensitivity and accuracy are most important indicators. Sensitivity measures if high risk patients are not considered low risk in treatment. The best sensitivity (15.08%) is given by MLP and NNC. Classification accuracy measures the proportion of true results (both tp and tn). The best classification accuracy is achieved by DTC (84.51%), however, the other techniques achieve similar results.

4. Conclusions

The use of data mining techniques implemented in open source software Weka for risk estimation of cardiovascular patients in our developing decision support system of BraveHealth were discussed in this paper. Java API of Weka is used after data is transformed into a particular form compatible with Weka in the decision support system. Four principally different data mining techniques are used for risk estimation: a Bayes network classifier, a decision tree classifier, a neural network classifier and a nearest neighbor classifier. Experimental results where these data mining techniques are employed on collected data are also provided. It is expected these data mining techniques will be used together with other data mining techniques and alert rules implemented in the developing decision support system for alert notifications in the BraveHealth system.

Acknowledgment. This work is supported 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] BOHÁČIK, J., DAVIS, D. N.: Alert rules for remote monitoring of cardiovascular patients, *Journal of Information Technologies*, Volume: 5, Issue:. 1, Year: 2012, Pages: 16-23, ISSN: 1337-7469.

[2] BOHÁČIK, J., DAVIS, D. N.: Data mining applied to cardiovascular data, *Journal of Information Technologies*, Volume: 3, Issue: 2, Year: 2010, Pages: 14-21, ISSN: 1337-7469.

[3] BOHÁČIK, J., DAVIS, D. N.: Estimation of cardiovascular patient risk with a Bayesian network, *Proc. of the TRANSCOM 2011 : 9-th European conference of young research and scientific workers* (Published by: University of Žilina; Held in: Žilina, Slovakia), Year: 2011, Pages: 37-40, ISBN: 978-80-554-0372-4.

[4] DAVIS, D. N., NGUYEN, T. T.: *Generating and Verifying Risk Prediction Models Using Data Mining: A Case Study from Cadiovascular Medicine. Chapter of Data Mining and Medical Knowledge Management: Cases and Applications* (1st edition), Publisher: IGI Global Inc., Year: 2009.

[5] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN, I. H.: The WEKA data mining software: An update, *ACM SIGKDD Explorations Newsletter*, Volume: 11, Issue: 1, Year: 2009, Pages: 10-18, ISSN: 1931-0145.

[6] LIESHOUT, J. v., WENSING, M., GROL, R: *Prevention of cardiovascular diseases: The role of primary care in Europe* (1st edition), Publisher: Electronic book retrieved from EPA Cardio (www.epa-cardio.eu), Year: 2008, Pages: 129, ISBN: 978-90-76316-25-3.

[7] NORSYS SOFTWARE CORP.: Netica^{TM} Application [http://www.norsys.com].

[8] PAGE-JONES, M.: *Fundamentals of object-oriented design in UML* (1st edition), Publisher: Addison-Wesley Professional, Year: 1999, Pages: 480, ISBN-10: 020169946X.

[9] WITTEN, I. H., FRANK, E., HALL, M. A.: *Data Mining: Practical Machine Learning Tools and Techniques* (3rd edition), Publisher: Morgan Kaufmann, Year: 2011, Pages: 629, ISBN: 978-0-12-374856-0.

Contact addresses

Ing. Ján Boháčik, PhD., EUR ING, Department of Computer Science, Faculty of Science, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom and Department of Informatics, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic,
E-mail address: J.Bohacik@hull.ac.uk,
E-mail address: Jan.Bohacik@fri.uniza.sk

Dr. Darryl N. Davis, Department of Computer Science, Faculty of Science, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom

RNDr. Miroslav Benedikovič, Department of Informatics, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic