Alternating decision tree applied to risk assessment of heart failure patients

Jan Bohacik[1], Department of Computer Science at the University of Hull, UK and Department of Informatics at the University of Žilina, Slovakia

C. Kambhampati, Department of Computer Science at the University of Hull, UK

Darryl N. Davis, Department of Computer Science at the University of Hull, UK

John G. F. Cleland, Department of Cardiology at the University of Hull, UK

Abstract: About 50% of the patients diagnosed with heart failure die within four years. At the same time, a rise in home telemonitoring of these patients can be observed. For its successful deployment, predicting if a heart failure patient could die within a certain period of time is an important task. An investigation of an alternating decision tree employed for this type of prediction is presented here. Experimental results are provided showing its performance on a database which contains data about 2032 patients with heart failure. Minimizing life-threatening situations while maintaining the costs of treatment are especially targeted.

Key words: alternating decision tree, classification, heart failure, cardiology

1. Introduction

In basically all European countries, the number of heart failure patients increases. An increased prevalence of heart risk factors such as aging population, obesity and diabetes can be seen and so a new increase is likely as well (8). These are followed by escalation in healthcare costs. To be more precise, there are more than 3.5 million newly diagnosed people in Europe every year. Simultaneously, patients with known heart failure are very likely to be readmitted (11). For this reasons, improvement of care at home instead of hospital care is emphasised more and more and new technologies such as remote monitoring devices are being developed. Remote monitoring devices enable patients with serious problems to record their own health measures and send them electronically to clinicians (12).

[1] Ján Boháčik with all Slovak diacritics.

When care at home is established for heart failure patients, it is important to predict if a patient could die within some time so that a prevention might be conducted. An issue is a lack of methods that could do this prediction. Some clinical methods such as EFFECT Risk Scoring system (6), Emergency Heart Failure Mortality Risk Grade (EHMRG) (7) or Seattle Heart Failure Model (SHFM) (5) are considerable for this type of task. At the same time, more and more data is collected by hospitals due to application of new information technologies such as monitoring devices there. The data can be analysed using data mining methods such as classifiers. There are several examples of medical data mining (1)(9)(10). In this paper, an alternating decision tree as a data mining method based on (4) is applied on prediction of death within six months for heart failure patients.

The organization of the paper is as follows. The used heart failure database is described in Section 2. In Section 3, the process of knowledge discovery in databases and the application of the alternating decision tree in it are described. Our experimental results achieved with the alternating decision tree are in Section 4. Section 5 concludes the paper.

2. Heart failure database

A group of 2032 patients with diagnosed heart failure $\boldsymbol{p} \in \boldsymbol{V}$ classified into two possible predictions and described by nine attributes $\boldsymbol{A}$ is used as a heart failure database. The particular patients are derived from Hull LifeLab - a large, epidemiologically representative, information-rich clinical database (2). Details of the heart failure database are in Table. 1. Describing attributes $\boldsymbol{A}$ are defined as $\boldsymbol{A} = \{A_1 ; \ldots ; A_k; \ldots ; A_9\}$. If A_k is a categorical attribute, $A_k = \{a_{k,1}; \ldots; a_{k,l}; \ldots; a_{k,l_k}\}$ where $a_{k,1}; \ldots; a_{k,l}; \ldots; a_{k,l_k}$ are possible categorical values. Class attribute C is used to classify heart failure patients into two possible predictions (class values) c_1 and c_2. It is denoted by $C = \{c_1 ; c_2\}$.

Table. 1: Heart failure database.

Attribute	**Data Type**	**Values**
Pulse Rate (A_1)	Numerical	38 - 150
NT-proBNP Level (A_2)	Numerical	0.89 - 18236
Blood Sodium Level (A_3)	Numerical	123 - 148

Age (A_4)	Numerical	27 - 96
Blood Uric Acid Level (A_5)	Numerical	0.11 – 1.06
Weight (A_6)	Numerical	29.80 – 193.80
Blood Creatinine Level (A_7)	Numerical	37 - 1262
Height (A_8)	Numerical	1.2 – 1.96
Sex (A_9)	Categorical	*female* ($a_{9,1}$)
		male ($a_{9,2}$)
Prediction (C)	Categorical	*alive* (c_1)
		dead (c_2)

Pulse Rate (A_1) is the rate of the patient's pulse measured by tactile on the outside of an artery in beats per minute. *NT-proBNP Level* (A_2) denotes the amount of the N-terminal prohormone of brain natriuretic peptide (NT-proBNP) in picograms per milliliter of the patient's blood. *Blood Sodium Level* (A_3) denotes the amount of sodium in millimoles per litre of the patient's blood. *Age* (A_4) is the age of the patient in years. *Blood Uric Acid Level* (A_5) represents the amount of uric acid in millimoles per liter of the patient's blood. *Weight* (A_6) is the patient's weight in kilograms. *Blood Creatinine Level* (A_7) denotes the amount of creatinine in micromoles per liter of the patient's blood. *Height* (A_8) represents the patient's height in meters. *Sex* (A_9) indicates if the patient is female or male. *Prediction* (C) denotes if the patient dies within six months (*dead*) or not (*alive*).

3. Alternating decision tree in knowledge discovery

Data mining, the process of knowledge discovery in databases (KDD), and the used alternating decision tree are described in more detail in this chapter. *Data mining* is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data (3). Here, data are a set of facts (for example, patients described in a database), and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, extracting a pattern also designates fitting a model to data; finding structure from data; or, in general,

making any high-level description of a set of data. In our case, the model or a high-level description is an alternating decision tree. The *KDD* itself is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The term process implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations. The steps are as follows: 1) Selection; 2) Preprocessing; 3) Transformation; 4) Data mining; 5) Interpretation/Evaluation. By nontrivial, it is meant that some search or inference is involved; that is, it is not a straightforward computation of predefined quantities such as computing the average value of a set of numbers.

A formal definition of alternating decision trees can be given through weighted votes of simple rules (4). Suppose a *base condition* is a boolean predicate over heart failure patients, AND is conjunction, NOT is negation, **T** denotes the constant predicate that is always true, and $\boldsymbol{S}$ denotes a set of conditions. A *precondition* is a conjunction of base conditions and negations of base conditions. A *base rule* r is a mapping of patients to real numbers which is defined in terms of a precondition *prec*, a base condition *bcon*, and two real numbers a and b. The base rule maps each instance to a *prediction* that is defined to be a if *prec* AND *bcon*, b if *prec* AND (NOT *bcon*), and 0 if NOT *bcon*. An *alternating decision tree* is a mapping of heart failure patients to real numbers which is defined in terms of a set of base rules. The set of base rules must obey the following conditions:

1. The set must include a base rule for which both the condition and precondition are **T**. The a value of this rule is the prediction associated with the root of the tree;
2. A base rule r with precondition $prec_r$ can be in the set only if the set includes a rule r' with precondition $prec_{r'}$ and condition $bcon_{r'}$ such that $prec_r = prec_{r'}$ AND $bcon_{r'}$ or $prec_r = prec_{r'}$ AND (NOT $bcon_{r'}$). $prec_r$ corresponds to the prediction node that is the direct parent of r.

The alternating decision tree maps each heart failure patient to a real valued *prediction* which is the sum of the predictions of the base rules in its set. The classification of a heart failure patient is the sign of the prediction.

There are several differences of alternating decision trees in comparison to standard decision trees. During building, in standard decision trees only leaf nodes can be split. In alternating decision trees, each part can be split multiple times. Also, decision nodes can be added at any

location in the tree, not just the leaves. The splitting criterion is different as well, it is the weighted error of the added rule, rather than the GINI index or information gain. In alternating decision trees, a heart failure patient defines a set of paths. As in standard decision tress, when a path reaches a decision node it continues with the child which corresponds to the outcome of the decision associated with the node. However, when reaching a prediction node, the path continues with all of the children of the node. More precisely, the path splits into a set of paths, each of which corresponds to one of the children of the prediction node. The sign of the sum of all the prediction nodes is the prediction (classification) for a particular heart failure patient. The main argument for the interpretability of alternating decision trees rests on the fact that the contribution of each decision node can be understood in isolation. Summing these contributions generates the prediction and the classification. After the meaning of each decision node in isolation is analysed, the interactions of the nodes can be analysed. Parallel decision nodes represent little or no interaction. If the conditions given in the tree are tested serially, evidence for or against the death of the heart failure patient is accumulated as the process is proceeded (a particular number is added to the total sum). The absolute value of the total sum can be considered a measure of confidence of the classification.

4. Experimental results

Our experimental analysis was conducted with our Java software tool. The core algorithm of the alternating decision tree was implemented in Weka (13) as class ADTree. The performance is measured with sensitivity = $\frac{tp}{tp + fn}$, specificity = $\frac{tn}{tn + fp}$, positive predictive value = $\frac{tp}{tp + fp}$, negative predictive value = $\frac{tn}{tn + fn}$, and accuracy = $\frac{tp + tn}{tp + fp + fn + tn}$. In the formulas, tp/fp/fn/tn is the number of true positives/false positives/false negatives/true negatives. "C is *alive*" is considered negative and "C is *dead*" is considered positive. Values tp, fp, fn and tn are computed during 10-fold cross-validation where the database is partitioned into 10 folds of patients. Of the 10 folds, a single fold is retained as the testing database for evaluation, and the remaining 9 folds are used as the learning database. The learning database is used for building of an alternating decision tree. The cross-validation process is repeated 10 times, with each of the 10 folds used exactly once as the testing database. It is crucial to avoid classification of dead patients as alive (which would lead to life-threatening situations) and classification of alive patients as dead (which would increase

the running costs of the treatment). The former is measured by sensitivity and the latter is measured by specificity. As a consequence, the sum of sensitivity and specificity is an important measure as well.

Table. 2: Experimental results for the alternating decision tree.

Measure (%)	**Value**
Sensitivity	37.3077
Specificity	91.5344
Sensitivity + Specificity	128.8421
Positive Predictive Value	60.2484
Negative Predictive Value	80.9357
Accuracy	77.6575

The achieved experimental results are presented in Table. 2 where the values for particular measures are expressed in percentages. For the alternating decision tree applied on the heart failure database described in Section 2, sensitivity is 37.3077%, specificity is 91.5344% and the sum of sensitivity and specificity is 128.8421%.

5. Conclusion

An alternating decision tree was employed on a heart failure database (the Hull LifeLab data) within the process of knowledge discovery in databases. The heart failure database consisted of 2032 patients with heart failure and the patients were described by 9 attributes and classified into alive and dead. The alternating decision tree was used for prediction of death within six months for a patient diagnosed with heart failure. The performance of the alternating decision tree was evaluated in 10-fold cross-validation where several measures were computed. Emphasis on avoiding life-threatening situations (measured by sensitivity) and decreasing the running costs of treatment (measured by specificity) was employed. The achieved sensitivity was 37.3077%, the achieved specificity was 91.5344%, and the sum of sensitivity and specificity was 128.8421%. A higher sensitivity could be achieved with

improvements of the alternating decision tree, for example, uncertainties that exist in the heart failure database could possibly be addressed through the adoption of fuzzy logic.

6. Bibliography

(1) Candelieri A., Conforti D., Perticone F., Sciacqua A., Kawecka-Jaszcz K., Styczkiewicz K.: Early detection of decompensation conditions in heart failure patients by knowledge discovery: the HEARTFAID approaches, Proceedings of Computers in Cardiology, Pages: 893-896, Year: 2008.

(2) Clinical Effectiveness and Evaluation Unit of the Royal College of Physicians: Managing chronic heart failure: learning from best practice (Published by: The Lavenham Press Ltd; In: United Kingdom), Pages: 48, Year: 2005.

(3) 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.

(4) Freund Y., Mason L.: The Alternating Decision Tree Algorithm, Proceedings of the 16th International Conference on Machine Learning, Pages: 124-133, Year: 1999.

(5) Ketchum E. S., Jacobson A. F., Caldwell J. H., Senior R., Cerqueira M. D., Thomas G. S., Agostini D., Narula J., Levy W. c.: Selective improvement in Seattle Heart Failure Model risk stratification using iodine-123 metaiodobenzylguanidine imaging, Journal of Nuclear Cardiology (Volume: 19; Number: 5), Pages: 1007-1016, Year: 2012.

(6) Lee D. S., Austin P. C., Rouleau J. L., Liu P. P., Naimark D., Tu J. V.: Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model, JAMA (Volume: 290; Number: 19), Pages: 2581-2587, Year: 2003.

(7) Lee D. S., Stitt A., Austin P. C., Stukel T. A., Schull M. J., Chong A., Newton G. E., Lee J. S., Tu J. V.: Prediction of heart failure mortality in emergent care: a cohort study, Annals of Internal Medicine (Volume: 156; Number: 11), Pages: 767-775, Year: 2012.

(8) Lopez-Sendon J.: The heart failure epidemic, Medicographia (Volume: 33; Number: 4), Pages: 363-369, Year: 2011.

(9) Phillips K. T., Street W. N.: Predicting outcomes of hospitalization for heart failure using logistic regression and knowledge discovery methods, Proceedings of AMIA Annual Symposium, Pages: 1080, Year: 2005.

(10) Trnka, A., Kovarova, M., Doci, I.: Occurrence of sexual disorders in Slovakia - a five-year analysis with Data Mining, Forum Staticum Slovacum (Volume: 7; Number: 2), Pages: 185-192, Year: 2011.

(11) Vinson J. M., Rich M. W., Sperry J. C., McNamara T. C.: Early readmission of elderly patients with congestive heart failure, Journal of the American Geriatrics Society (Volume: 38; Number: 12), Pages: 1290–1295, Year: 1990.

(12) West D.: How mobile devices are transforming healthcare, Issues in Technology Innovation, Year: 2012.

(13) Witten I. H., Frank E., Hall M. A.: Practical machine learning tools and techniques (Third Edition) (Published by: Morgan Kaufman; In: United States of America), Pages: 664, Year: 2011.

7. Addresses of the authors:

Eur Ing Dr Jan Bohacik
Department of Computer Science/Department of Informatics
Faculty of Science and Engineering/Faculty of Management Science and Informatics
University of Hull/University of Žilina
HU6 7RX/010 26
Hull/Žilina
United Kingdom/Slovakia
J.Bohacik@hull.ac.uk/Jan.Bohacik@fri.uniza.sk

Dr C. Kambhampati
Department of Computer Science
Faculty of Science and Engineering
University of Hull
HU6 7RX
Hull
United Kingdom
C.Kambhampati@hull.ac.uk

Dr Darryl N. Davis
Department of Computer Science
Faculty of Science and Engineering
University of Hull
HU6 7RX
Hull
United Kingdom
D.N.Davis@hull.ac.uk

Professor John G. F. Cleland
Department of Cardiology
Castle Hill Hospital
University of Hull

HU16 5JQ
Hull
United Kingdom
J.G.Cleland@hull.ac.uk

Acknowledgements

This work was supported by a HEIF-5 funded project. The authors would like to thank the University of Hull, UK for its support. The heart failure database was provided by the Hull York Medical School, University of Hull, UK as a part of Hull LifeLab.

The paper was accepted for publication in August 2013 by the publisher of the Journal of Information Technologies (Vol. 6, No. 1, ISSN: 1337-7467), i.e. by the Department of Applied Informatics and Mathematics, Faculty of Natural Sciences, University of Ss. Cyril and Methodius in Trnava, Slovakia.