Complications detection in treatment for bacterial endocarditis

Leticia Curiel[1], Bruno Baruque[1], Carlos Dueñas[2], Emilio Corchado[3] and Cristina Pérez[2]

[1]*Department of Civil Engineering, University of Burgos, Burgos, Spain.*

[2] *Complejo Hospitalario Asistencial Universitario de Burgos (SACYL), Servicio de Medicina Interna, Burgos, Spain.*

[3] *Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, Spain.*

emails:lcuriel@ubu.es, bbaruque@ubu.es, cjdg@hgy.es, escorchado@usal.es

Abstract. This study proposes the use of decision trees to detect possible complications in a critical disease called endocarditis. The endocarditis illness could produce heart failure, stroke, kidney failure, emboli, immunological disorders and death. The aim is to obtained a tree decision classifier based on the symptoms (attributes) of patients (the data instances) observed by doctors to predict the possible complications that can occur when a patient is in treatment of bacterial endocarditis and thus, help doctors to make an early diagnose so that they can treat more effectively the infection and aid to a patient faster recovery. The results obtained using a real data set, show that with the information extracted form each case in an early stage of the development of the patient a quite accurate idea of the complications that can arise can be extracted.

1 Introduction

Machine Learning [1, 2] is a field related to tasks as recognition, diagnosis, planning, robot control, prediction, etc. These concepts involve techniques, such as algorithms for dimensionality reduction as PCA [3], artificial neural networks [4], genetic algorithms [5, 6], fuzzy systems [7] and swarm intelligence [8], which investigate complex problems to solve real problems in fields as medicine [9], ecology [10], engineering [11], industrial process [12] and so on.

Endocarditis is a term used to describe a serious infection of the endocardium that can cause severe damage to the inner lining of the heart, to any of the four valves of the heart and to other structures such us the interventricular septum, the chordae tendineae, the mural endocardium, or even on intracardiac devices. The infection can occur in any age and either sex. Usually, the illness is caused by a growth of bacteria

on one of the heart valves, leading to an infected mass called "vegetation". It could be classified in:

- Bacterial endocarditis: this is produced when bacteria enter the bloodstream.
- Fungal endocarditis: occurs in people with low resistance to infection, such as those who are taking medications that suppress the immune system.
- Noninfective endocarditis: is a heart inflammation caused by advanced step of cancer or by disorders of the immune system.

According to the American Heart Association (AHA), the infection may be contracted during brief periods of introduction of bacteria in the bloodstream, such as after dental procedures, tonsillectomy or adenoidectomy, examination of the respiratory passageways with an instrument known as a rigid bronchoscope, certain types of surgery on the respiratory passageways, the gastrointestinal tract, or the urinary tract and gallbladder or prostate surgery.

The endocarditis can be diagnosed by many procedures [13, 14] such as transthoracic echocardiography, by transesophageal echocardiography, by Duke Criteria, by autopsy, etc.

Once the illness has been diagnosed a rapid initiation of an adequate therapeutic regimen is important to prevent the patients from severe complications such as heart failure, stroke, kidney failure, septic embolism and various immunological phenomena, variety of systemic signs and symptoms through several mechanisms, including infertility or death.

The main treatment [13, 14] of the infection is through aggressive antibiotics, usually intravenously, which attack the microorganisms. The problem is that the diagnosis of what kind of bacteria originated the infection is based on positive blood culture results with identical microorganisms, which is not an immediate process. So, usually, doctors in many cases have to begin the treatment before knowing the specific bacteria the patient is infected with. Also, antibiotic treatment is sometimes not enough because the valve has been severely damaged and a surgical replacement of the valve is required.

For all these reasons the correct treatment of the patient in the earliest stage as possible is considered as an interesting objective. To help to achieve it, this research proposes the use decision tree [15, 16] techniques to recognize possible complications once the patient is in treatment, helping to identify in advance possible solutions.

The remaining of this paper is organised as follows. Section 2 introduces the decision tree learning techniques used to construct the different classifiers presented. Section 3 describes the dataset used for this analysis; Section 4 shows the experiments and results obtained. Finally, in Section 5, the conclusions are set out and comments are made on future lines of work.

2 Tree Learning Algorithms

Machine Learning [1, 2] deals with algorithms that can construct models to estimate or predict the class to which new cases belong to. One manner to do it is through decision trees [15]. A tree is a leaf node labelled with a class linked to two or more nodes, where each branching node represents a choice between different alternatives. So, to classify instances, an attribute-vector must be presented to the tree and evaluate each of its composing attributes in the corresponding node. To complete the classification process, some tests into the attributes obtained reaching one or other leaf, are carried out.

The inputs of a decision tree consist on a collection of training cases with an expected dependence between variables. Each of the training cases is included into a single class into which the problem to solve is divided. The goal of the decision trees is to learn from these training cases to be able to classify futures instances.

In the following subsections three commonly used systems for induction of decision trees for classification are described: CHAID, ID3 and C4.5.

2.1. Chi-squared Automatic Interaction Detection

Chi-squared Automatic Interaction Detection (CHAID) [17] is a decision tree method useful in exploratory analysis that relates a potentially large number of categorical predictor variables to a single categorical nominal dependent variable.

The method was proposed as a modification of the Automatic Interaction Detector method (AID) [18] for categorized dependent and independent variables

The algorithm incorporated a sequential merge and split procedure based on a chi-square test statistic and proceeds in steps as follows:

- Cross tabulate the m categories of the predictor with the k categories of the dependent variable.
- Then, find the pair of categories of the predictor which account for the least significant difference on a chi-square test and merge these two categories.
- Repeat the merging process until the chi-square test is significant according to a proposed value.
- Pick the predictor variable whose chi-square is largest and split the sample into $m \leq l$ subsets, where l is the number of categories resulting from the merging process on that predictor.
- Finally, continue splitting, until no “significant” chi-squares result.

2.2. The Iterative Dichotomiser 3

The Iterative Dichotomiser 3 (ID3) [16, 19] is a mathematical algorithm used to generate decision trees. The resulting tree is used to classify new samples. This algorithm consists of constructing a tree from a random subset of the training set. The

process must be repeated with the incorrect classifications values while the tree does not classify correctly the remaining cases of the training set.

To achieve this, the algorithm extracts the attribute that best separates the given cases into targeted classes. The algorithm uses the statistical property called "information gain" to choose which attribute is the best at separating training examples. This gain of set S on attribute A is defined as follows:

$$G(S,A) = E(S) - \sum_{v=1}^{t} \frac{|S_v|}{|S|} E(S_v) \tag{1}$$

Where $\sum$ is each value v of all possible values of attribute A; S_v represents a subset of S which attribute A has value v; $|S_v|$ and $|S|$ are the number of elements in S_v and in S, respectively; and $E(S)$ is the information entropy of the subset S expressed by:

$$E(S) = -\sum p(I) \log_2 p(I) \tag{2}$$

Where $p(I)$ is the collection of S belonging to class I.

2.3. The C4.5 Algorithm

C4.5 [20] is an algorithm used to create decision trees and is considering as the successor of the ID3 [16] algorithm developed by Ross Quinlan too. This algorithm works as the same way as its predecessor, ID3, using the information gain (Eq. (1)) to choose the test A that maximizes $G(S,A)$ (Eq. (1)). The problem of using this approach is that it can favour data sets with numerous outcomes. To avoid this, it includes a measure called the "gain ratio" (Eq. (3)) by also taking into account the potential information from the partition itself:

$$P(S,A) = -\sum_{v=1}^{t} \frac{|S_v|}{|S|} \log\left(\frac{|S_v|}{|S|}\right) \tag{3}$$

Finally, the algorithm chooses the test A that maximizes the gain ratio, expressed by:

$$H(S,A) = \frac{G(S,A)}{P(S,A)} \tag{4}$$

3 Data Description

The data set has been collected by the Complejo Hospitalario Asistencial Universitario de Burgos (Spain) and contains 50 different cases. Those cases contain medical data extracted from the evolution of 50 different patients that were admitted into the hospital and diagnosed with endocarditis.

The following input variables have been considered for the study:

- Patient's age: contains cases ranging from 15 to 89 years old.
- Patient's sex: Male or female.
- Previous valve: Indicates whether the heart valve is native, prosthetic or is a pacemaker.
- Valve type: Indicates the type of infected heart valve: It is discriminated between native valve, prosthetic valve, pacemaker or prosthetic valve with pacemaker.
- Clinical Time: Indicates the time lapse that passed from first symptoms to endocarditis diagnosis (in days).
- Organism: bacteria that causes the infection. Contains more than 10 different types and its variants; such us enterococcus faecalis, enterococcus faecium, Haemophilus parainfluenzae, staphylococcus Lugdunens, staphylococcus parasanguis,…

The output to be predicted is the complications that may occur during treatment. The following complications have been considered:

- Heart failure.
- Cardiogenic shock: worse than heart failure.
- Septic emboli.
- Uncomplicated.

4 Experiments and results

The purpose of this multidisciplinary study is the prediction of possible complications once the patient is in treatment of endocarditis.

The dataset considered has 50 different cases: 38 of those cases have been used to train the decision tree and the remaining 12 samples are used to test the model.

In order to get the most adequate classifier to this case, different decision tree learning algorithms have been applied and their results have been compared. This comparison is shown in Table 1.

Table 1. Decision trees results (CHAID, ID3 and C4.5).

		CHAID	ID3	C4.5
Class recall	Uncomplicated	100.00%	100.00%	**100.00%**
	Cardiogenic shock	0.00%	100.00%	**100.00%**
	Septic emboli	0.00%	0.00%	**100.00%**
	Heart Failure	0.00%	50.00%	**50.00%**
Class prediction	Uncomplicated	66.67%	80.00%	**88.89%**
	Cardiogenic shock	0.00%	100.00%	**100.00%**
	Septic emboli	0.00%	0.00%	**100.00%**
	Heart Failure	0.00%	100.00%	**100.00%**
Parameters		minimal size for split 4, minimal leaf size 3, minimal gain 0.1, maximal depth 10 and confidence 0.2	minimal size for split 4, minimal leaf size 2, minimal gain 0.1	minimal size for split 4, minimal leaf size 2, minimal gain 0.1, maximal depth 20, confidence 0.25
Accuracy		66.67%	83.33%	**91.97%**

As shown in Table 1, the best results are obtained with the C4.5 model. C4.5 model is able to predict future cases in a figure close to 92% when the other models that achieve values close to 84% and to 67%. Figure 1 shows the final tree decision model.

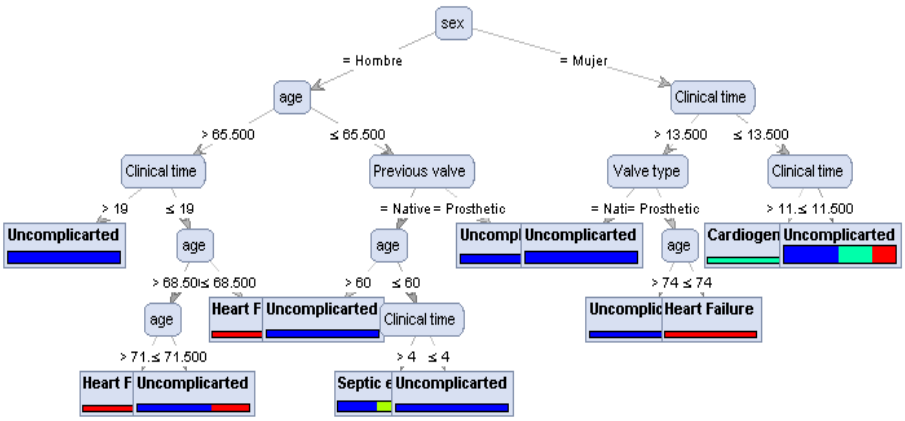

Fig. 1. The C4.5 model

The final model (Fig. 1) shows the structure of the decision tree. It can be noticed that the organism input variable does not appear in the model because it does not affect the classification in a substantial way. As it has been previously mentioned, one of the problems related to the endocarditis treatment is the ignorance of the sorts of bacteria causing the infection, so the identification of a model that is able to classify without this variable is very advantageous.

5 Conclusions and future research

The present study describes an ongoing multidisciplinary research in which an application of classical models by means of decision tree algorithms to a medical diagnosis problem has been presented. We have identified the complications with a reasonable degree of accuracy using a relatively quite small amount of samples and attributes. In this application field this means small amount of patients and a low number of medical tests and analyses; which seems as an advantageous feature, being this kind of real data so costly to acquire.

Future work will be focused on the collection and storage of more specific attributes for each patient. Results seem to point to the fact that with more detailed data the medical condition of each patient and enough amount of different patients better results could be obtained. These results may include better prediction of complications based on detailed data obtained from simple tests performed as close to the admission time of the patient as possible.

Another research line may be the use of the information and experience gathered in these experiments for the development of a Case Base Reasoning system [21] to solve tasks related to the ones presented above. These would be able to handle the incorporation of new information with the treatment and monitoring of the evolution of more patients. They also seem to be more intuitive for medical professionals, which are not used to deal with complex statistical models.

Acknowledgments.

We would like to extend our thanks to Complejo Hospitalario Asistencial Universitario de Burgos (SACYL). This research has been partially supported through projects TIN2010-21272-C02-01 from the Spanish Ministry of Science and Innovation and Grupo Antolin Ingenieria, S.A., within the framework of project MAGNO2008 - 1028.- CENIT.

References

[1] A. Abraham, E. Corchado, and J.M. Corchado. Hybrid learning machines. *Neurocomputing*, 72(13-15):2729–2730, 2009.

[2] T.M. Mitchell. The Discipline of Machine Learning. Technical Report CMU-ML-06-108, School of Computer Science, Carnegie Mellon University, 2006.

[3] K.H. Esbensen and P. Geladi. Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice. In Stephen D. Brown, R. Tauler, , and B. Walczak, editors, *Comprehensive Chemometrics*, pp. 211 – 226. Elsevier, Oxford, 2009.

[4] A. Herrero, E. Corchado, L. Sáiz, and A. Abraham. DIPKIP: A connectionist knowledge management system to identify knowledge deficits in practical cases. *Computational Intelligence*, 26(1):26–56, 2010.

[5] A.C. Lorena and A.C. Ponce. Evolutionary design of code-matrices for multiclass problems. In *Soft Computing for Knowledge Discovery and Data Mining*, pp. 153–184. Springer, 2008.
[6] M.C. Naldi, A.C. Ponce, R.J. Gabrielli, and E.R. Hruschka. Genetic clustering for data mining. Vol. 2, pp. 113–132. Springer, 2008.
[7] F.J. Berlanga, A.J. Rivera, M.J. Jesus, and F. Herrera. GP-COACH: Genetic Programming-based learning of Compact and Accurate fuzzy rule-based classification systems for High-dimensional problems. *Information Science*, 180(8):1183–1200, 2010.
[8] S. Das, A. Abraham, and A. Konar. Automatic kernel clustering with a Multi-Elitist Particle Swarm Optimization Algorithm. *Pattern Recognition Letters*, 29(5):688–699, 2008.
[9] M.Y. Lee and C.S.Yang. Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images. *Computers Methods and Programs in Biomedicine*, 100(3):269–282, 2010.
[10] Bruno Baruque, Emilio Corchado, Aitor Mata, and Juan M. Corchado. A forecasting solution to the oil spill problem based on a hybrid intelligent system. *Information Sciences*, 180(10):2029 – 2043, 2010. Special Issue on Intelligent Distributed Information Systems.
[11] J. Sedano, L. Curiel, E. Corchado, E. de la Cal, and J.R. Villar. A Soft Computing Based Method for Detecting Lifetime Building Thermal Insulation Failures. *Integrated Computer-Aided Engineering, IOS Press*, 17(2):103–115, 2010.
[12] J. Sedano, E. Corchado, L.Curiel, J.R. Villar, and P.M. Bravo. The Application of a two-step AI Model to an Automated Pneumatic Drilling Process. *International Journal of Computer Mathematics*, 86(10-11):1769–1777, 2009.
[13] B. Plicht and R. Erbel. Diagnosis and treatment of infective endocarditis. Current ESC guidelines. *HERZ*, 35(8):542–548, 2010.
[14] B. Plicht, R.A. Janosi, T. Buck, and R. Erbel. Infective endocarditis as cardiovascular emergency. *HERZ*, 51(8):987–994, 2010.
[15] J.R. Quinlan. Learning decision tree classifiers. *ACMComputing Surveys (CSUR)*, 28(1):71–72, 1996.
[16] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.
[17] G.V. Kass. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2):119–127, 1980.
[18] J.N. Morgan and J.A. Sonquist. Problems in the Analysis of Survey Data and a Proposal. *Journal of the American Statistical Association*, 58(3):415–434, 2010.
[19] A. Colin. Building Decision Trees with the ID3 Algorithm. *Dr. Dobbs Journal*, 1996.
[20] J.R. Quinlan. C4.5: Programs for Machine Learning. *Machine Learning*, 16(3):235–240, 1993.
[21] A. Aamodt and E. Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *Artificial Intelligence Communications-AICom*, 7(1):39–59, 1994.