Genetic Algorithms to Simplify Prognosis of Endocarditis

Leticia Curiel[1], Bruno Baruque[1], Carlos Dueñas[2], Emilio Corchado[3] and Cristina Pérez-Tárrago[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 ongoing interdisciplinary research is based on the application of genetic algorithms to simplify the process of predicting the mortality of a critical illness called endocarditis. The goal is to determine the most relevant features (symptoms) of patients (samples) observed by doctors to predict the possible mortality once the patient is in treatment of bacterial endocarditis. This can help doctors to prognose the illness in early stages; by helping them to identify in advance possible solutions in order to aid the patient recover faster. The results obtained using a real data set, show that using only the features selected by employing a genetic algorithm from each patient's case can predict with a quite high accuracy the most probable evolution of the patient.

1 Introduction

Dimensionality reduction methods [1] involve processes such as feature construction, space dimensionality reduction, and sparse representations among others, which are achieved by using a wide array of techniques such as genetic algorithms [2], fuzzy systems [3] and others that investigate complex real problems in fields as medicine [4], ecology [5], engineering [6] and so on.

Infective endocarditis is a serious infection and its morbidity and mortality rate is still high, with a reported overall mortality rate ranging from 16 to 37.1% The risk of acquiring infective endocarditis is higher among patients with underlying heart diseases including valvular heart disease and congenital heart disease, among those with prosthetic cardiac valves, and among intravenous drug abusers. Substantial questions remain regarding the risk factors for infective endocarditis in bacterial infection. The changing profile of Infective Endocarditis requires continuous

epidemiological updating associated infection.Usually, the illness is caused by a growth of bacteria on the edges of a defected heart or on the surface of an abnormal valve; after the bacteria enter the blood stream most commonly from dental procedures, tonsillectomy or adenoidectomy, certain types of surgery on the respiratory passageways, but also from procedures involving the gastrointestinal or urinary tract.

The endocarditis can be diagnosed by many procedures [7, 8] such as transthoracic echocardiography, transesophageal echocardiography, Duke criteria, magnetic resonance, tomography miltislide and by embolisms, etc.

Once the illness has been diagnosed, a rapid initiation of an adequate therapeutic regimen is important to prevent complications such as arrhythmias, brain abscess, brain or nervous system changes, congestive heart failure, glomerulonephritis, jaundice, severe heart damage, stroke,.., and death.

Patients with this condition usually need to be hospitalized to begin an aggressive treatment [7, 8] based on intravenously antibiotics. Initially, the treatment is empirical and the ideal situation is encountering the specific antibiotic for the organism causing the condition. This is determined by the blood culture and the sensitivity tests, which is not an immediate process.

For all these reasons, the correct treatment of the patient in the earliest stages as possible, is considered as an interesting objective. To help achieving this objective, this research proposes the use of genetic algorithms [9] techniques to select the most important features of this illness once the patient is in treatment, helping to predict the mortality risk.

The remaining of this paper is organised as follows. Section 2 introduces the decision genetic algorithms techniques used to realize feature selection. Section 3 describes classification models; in section 4 the dataset is explained; Section 5 shows the experiments and results obtained. Finally, in Section 6, the conclusions are set out and comments are made on future lines of work.

2 Feature Selection

The objective of this study is the identification of the most important patient's characteristics or symptoms in order to determine the future evaluation of their illness. As explained in previous sections, some of those are obtained from medical tests that can take a relatively long time, so it is important to know in advance which of them must be given higher priority. This is therefore, a clear case where the application of feature selection algorithms can be of much use.

In the case of this study, a Genetic Algorithm is employed as a mean for feature selection, enabling to guide the search among the most interesting combination of attributes (or dimensions) to obtain similar results of the ones obtained by using the whole set of attributes or characteristics for each patient.

2.1. Genetic Algorithms (GAs)

These kinds of algorithms are devised to solve search and optimization problems. They were originally proposed in [9] and are based in the evolution process of the

biological species in nature. By imitating this behaviour, this family of algorithms is able to "evolve" a population of different solutions to the problem presented, until one of the generated solutions is fit enough to be considered as the final one [10].

The power of GAs comes from the fact that the basic technique is robust and can deal with a wide array of different problem statements. They are not guaranteed to find the global optimum solution for the given problem, but can achieve an "acceptably good" solution in a relatively low time [11].

In the case of the present work, this algorithm has been used as a way of performing a guided search among the different attributes that could be used to classify future evolution of the patients. This is usually known in literature as a wrapper method [12]. Each individual represents a different subset of the features chosen among the whole of them; while the fitness of each individual is the classification rate obtained by a regular machine learning classifier. In order to test the method in combination with a wider array of models, tests have been performed with three different classifiers: Support Vector Machines, ID3 Decision Trees and Naïve Bayes with Kernel Density Estimation.

3 Data Classification

3.1. Support Vector Machines

The Support Vector Machines (SVM) are supervised algorithms for the classification of multi-dimensional data samples or regression analysis. The most well-known version of the algorithm is the one proposed in [13].

It is based in the concept of hyper-planes used as decision boundaries. The algorithm is devised to find a high-dimensional plane that divides the data samples used as a training set into different classes, according to the labels provided. One of their main characteristics is that it will find the hyper-plane that accounts for the largest distance to the nearest training data points of any class, obtaining therefore the best possible generalization [14].

Mathematically expressed: if we consider the data samples $x_i \in \Re^d$ with their corresponding class labels $y_i \in \{\pm 1\}$; the SVM performs a mapping to a higher dimensional Hilbert space $\Phi : \Re^d \to H$. In that space (H) the decision rule is governed by a simple hyperplane that separates x_i into two different classes:

$$\overline{\psi} \cdot \overline{x}_i + b \geq k_0 - \xi_i, y_i = +1 \tag{1}$$

$$\overline{\psi} \cdot \overline{x}_i + b \leq k_1 + \xi_i, y_i = -1 \tag{2}$$

where ξ_i are positive slack variables introduced to handle the non-separable case and where k_0 and k_1 are typically defined to be +1 and -1 respectively.

In those cases, the Ψ is calculated by minimizing the objective function:

$$\frac{\overline{\psi}\cdot\overline{\psi}}{2}+C\left(\sum_{i=1}^{\ell}\xi_i\right)^p \tag{3}$$

subject to Eqs. (1) and (2), where C is a constant and p is usually chosen to be 2. A test vector (x_i) is then assigned a class label depending on whether $\overline{\psi}\cdot\overline{\Phi}(x)+b$ is greater or less than $(k_0 + k_1)/2$.

3.2. The Iterative Dichotomiser 3

The Iterative Dichotomiser 3 (ID3) [15] is a mathematical algorithm used to generate decision trees. 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 to separate training examples. This gain of set S on attribute A is defined as follows:

$$G(S,A) = E(S) - \sum_{v=1}^{t}\frac{|S_v|}{|S|}E(S_v) \tag{3}$$

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:

$$E(S) = -\sum p(I)\log_2 p(I) \tag{4}$$

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

3.3. Naïve Bayes with Kernel Density Estimation

The naïve Bayes classifier is also a very widespread supervised classifier, known for its simplicity and relatively good performance [16]. It is based in the probability theory, more precisely in Bayes theorem [17]. This method has the particularity that it will assume that the probability of each of the different attributes, to determine the final class of the sample, can be considered independent of the rest. That is, are conditionally independent given the class label. Although this does not always happen to be true, it is a good way to simplify the calculations. It performs its classification by calculating the *a priory* probability and the Likelihood of a sample belonging to a class by using a set of previously labeled training data.

Among many of the modifications that have been introduced to the basic algorithm, one of the most used is the inclusion of a kernel density estimator to calculate the true density of the continuous variables using kernels in the computation of the Likelihood of samples [18].

4 Data Description

The data set contains 50 cases of bacterial endocarditis extracted from the evolution of different patients that were admitted into the Complejo Hospitalario Asistencial Universitario de Burgos (Spain).

The following 14 input variables have been collected:

- Diagnostic Tool: Shows how the endocarditis has been diagnosed. The variables considered are: transesophageal echocardiography, transthoracic echocardiography, Duke criteria or autopsy.
- Clinical Time: It is the number of days passed from the appearance of first symptoms to endocarditis diagnosis.
- Patient's age: Contains the patient's age, where there are cases ranging from 15 to 89 years old.
- Patient's sex: Male or female.
- Complications: Resulting from infection during the treatment. It has been considered that the patient may suffer from heart failure, cardiogenic shock which is worse than heart failure, septic emboli and uncomplicated.
- Septic shock: It is life-threatening low blood pressure due to the introduction of bacteria into the blood stream.
- Catheter Sepsis: Indicates if sepsis is associated with intravascular catheters.
- Appropriate treatment: Once the illness is diagnosed, a rapid initiation of an adequate therapeutic regimen is required to prevent severe complications. The main treatment [7, 8] is through aggressive antibiotics. The problem is that the diagnosis of what kind of bacteria originated the infection is based on positive blood culture results with identical micro-organisms, which is not an immediate process. So, doctors in many cases have to begin the treatment before knowing the specific bacteria the patient is infected with. For this reason, sometimes, the treatment has to be changed once the blood culture results have been obtained. Then, this variable indicates whether the initial treatment is correct according to the bacteria.
- Change Time: The number of hours that the patient has been with an incorrect treatment.
- Previous valve: Indicates whether the affected patient's heart valve was working properly before being infected
- Valve type: Indicates the type of the infected heart valve. It is discriminated between native valve, prosthetic valve, pacemaker or prosthetic valve with pacemaker.
- Infected valve: Indicates the valve or valves affected. 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,…
- ICU: indicates whether the patient has been admitted to the intensive care unit.

The output to be predicted is the patient's condition 30 days after being admitted to the hospital. To simplify the problem, only the differentiation between “Alive” and “Dead” has been considered.

5 Experiments and results

The aim of the experiments is to determine the most interesting set of features to determine the future evolution of the patient. In order to validate and test the use of GAs in this study, a classification comparison is proposed. A classification has been performed only with the variables identified as most relevant by the GA, after performing a wrapped search among all the features available on the dataset; then the results are compared with a classification performed using all the variables of the dataset.

For all experiments, the initial dataset is the one described in Section 4. It is therefore, composed of 50 different cases, each corresponding to a different patient; and 14 possible variables or features. As the dataset is relatively small, all experiments have been performed using the standard 10-fold cross-validation, in order to obtain statistically significant measures.

Table 1. Parameters used in the training of the models.

SVM	**ID3**	**Naïve Bayes**
Kernel type: Dot. Kernel cache: 200 Convergence epsilon: 0.001 Maximun iterations: 1000 Complexity constant: 0	Criterion: gain ratio Minimal size for split: 4 Minimal leaf size: 2 Number of threads: 2	Estimation mode: Greedy. Bandwith: 0.1 Number of kernels: 10
Genetic Algorithm parameters		
Selection mode: Roulette wheel Population size: 5 Selection scheme: Tournament Tournament size: 0.25	Prob. initalization: 0.5 Prob. mutation: -1 Prob. crossover: 0.5 Crossover type: Uniform	

Table 2. Classification results with the different algorithms.

		Feature Selection Classification			**Classification with all features**		
		SVM	ID3	Naïve Bayes	SVM	ID3	Naïve Bayes
Class recall	Alive	94.87%	97.44%	**97.14%**	94.87%	76.92%	94.44%
	Dead	33.33%	22.22%	**42.86%**	11.11%	44.44%	12.50%
Class precision	Alive	86.05%	84.44%	**89.47%**	82.22%	85.71%	82.93%
	Dead	60%	66.67%	**75.00%**	33.33%	30.70%	33.33%
Accuracy		83%	83.50%	**88.50%**	79.50%	71%	80.50%

For the sake of comparison, three different classification algorithms have been applied both to the complete dataset and inside the GA wrapper to select the subset of features for classification. The results of the three have been compared and shown in

Table 2; while Table 3 shows the variables discriminated in each of the tests.Table 1 shows the parameters used for the training of each of the classifiers and for the Genetic Algorithm wrapper.

Table 3. Features selected by the wrapped search depending on the model.

	SVM	**ID3**	**Naïve Bayes**
Features selected	Age ICU Septic shock Complications Diagnostic Tool Previous valve Infected valve	Sex Complications Previous valve Organism Septic shock Catheter Sepsis Diagnostic Tool Infected valve Valve type	Sex AgeICU Septic shock Complications Appropriate treatment Change time Previous valve Valve type Infected valve Organism

Looking at Table 2, it seems clear that the model that best classifies the future state of the patients, when is only trained with the features extracted using the Genetic Algorithm, is the Naïve Bayes. In the experiments performed, this combination accounts for the highest classification accuracy and best recall and precision in all but one class precision. It is able to prognosticate future cases close to 89% with the features selected where other models achieve values close to 83%.

Analysing Table 3, in this case according with the doctors' expertise and previous medical publications [19]; the most interesting set of features is the one included in the third column.

In the model adjusted for clinically important variables (age, sex, health care–associated acquisition of infection, diabetes, cancer, long-term immunosuppressive therapy, Organism *(S. aureus* infection), Previous valve (paravalvular abscess, cardiac surgery, Complications (stroke, heart failure, and new conduction abnormality)), variables independently associated with higher mortality among patients with native valve endocarditis were age 60 years or older, health care–associated acquisition of infection, diabetes, *S. aureus* infection, paravalvular abscess, stroke, heart failure, and new conduction abnormality. In agreement with previous studies, the results point to the fact that advanced age and endocarditis complications (stroke, heart failure, and septic embolism) were associated with greater mortality in patients with native valve endocarditis. Independent predictors of in-hospital mortality among patients with endocarditis in the present study included increasing age, systemic embolism, heart failure, prosthetic valvular endocardidtis and clinical delay. So, it can be concluded that the mortality rate may be increased by patient factors such as age and comorbid conditions, rather than by intrinsic qualities of the organism.

It is interesting to note that precisely the model combination that obtains best classification accuracy is the one proposed as having selected the most relevant features for the problem. This even outperforms the same classification algorithm when trained with all available features. Although this situation does not really add knowledge to what doctors already know, it proves that these models can successfully

discard additional unimportant information and help to the prognosis of an illness using the patients' symptoms as they were obtained for the presented tests.

6 Conclusions and future research

The present study describes an ongoing multidisciplinary research in which an application of classical models by means of genetic algorithms to a medical problem has been presented. The features selected through the genetic algorithms presented are consistent with the medical literature found [19] and the models tested have been able to predict the mortality risk with a reasonable degree of accuracy using a relative small amount of samples.

This work proves that is possible to identify and discard the most uninteresting features for this analysis using automated learning algorithms, enabling doctors to concentrate in the remaining –most interesting– ones in the specific case of the endocarditis. In this application field, using this small amount of patients and reducing the features needed for each of them, seems as an advantageous feature; as such kind of real data are 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 alongside enough amount of different patients better results could be obtained. These results may include better prediction of mortality risk based on detailed data obtained from simple tests performed as close to the admission time of the patient as possible.

Another research line is the use of the information and experience gathered in these experiments for the development of a Case Base Reasoning system [20] 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 the evolution of more patients.

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] H. Liu and L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Educational Activities Department*, 17(4):491–502, 2005.

[2] A.C. Lorena and A.C. Ponce. Evolutionary design of code-matrices for multiclass problems. In *Soft Computing for Knowledge Discovery and Data Mining*, pages 153–184. Springer, 2008.

[3] 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.

[4] C-D. Chang, C-C. Wang, and B.C. Jiang. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert Systems with Applications*, 38(5):5507–5513, MAY 2011.

[5] B. Baruque, E. Corchado, A. Mata, and J.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.

[6] 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.

[7] B. Plicht and R. Erbel. Diagnosis and treatment of infective endocarditis. Current ESC guidelines. *HERZ*, 35(8):542–548, 2010.

[8] B. Plicht, R.A. Janosi, T. Buck, and R. Erbel. Infective endocarditis as cardiovascular emergency. *HERZ*, 51(8):987–994, 2010.

[9] J.H. Holland. *Adaptation in natural and artificial systems* . MIT Press Cambridge, USA, 1992.

[10] D.E. Goldberg. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, 1996.

[11] T. Niknam, E.T. Fard, N. Pourjafarian, and A. Rousta. An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. *Engineering Applications of Artificial Intelligence, Pergamon-Elsevier Science Ltd*, 24:306–317, 2011.

[12] Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, , 97: 273–324,1997

[13] V. Vapnik. *Statistical Learning Theory*. Springer, New York, USA, 1998.

[14] C.J.C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121â€“167, 1998.

[15] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.

[16] I. Rish. An empirical study of the naive Bayes classifier. In *Proceedings of IJCAI-01 workshop on Empirical Methods in AI In International Joint Conference on Artificial Intelligence (2001)*, pages 41–46, 2001.

[17] T. Bayes. An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53(2):370–418, 1763.

[18] P. Larrañaga, I. Inza, and A.P. Martinez. Bayesian classifiers based on kernel density estimation. *International journal of approximate reasoning*, 50(2):341–362, 2009.

[19] N. Benito, J.M. Miro, E. Lazzari, C.H. Cabell, A. Rio, J. Altclas, P. Commerford, F. Delahaye, S. Dragulescu, H. Giamarellou, G. Habib, A. Kamarulzaman, A. Sampath, F.M. Nacinovich, F. Suter, C. Tribouilloy, K. Venugopal, A. Moreno, V.G. Fowler, and the ICE-PCS (International Collaboration on Endocarditis Prospective Cohort Study) Investigators. Health Care Associated Native Valve Endocarditis: Importance of Non-nosocomial Acquisition. *Annals of Internal Medicine*, 150:586–594, 2009..

[20] A. Aamodt and E. Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *Artificial Intelligence Communications-AICom*, 7(1):39–59, 1994.