Support Vector Machines and Genetic Algorithms for Detecting Unstable Angina

J Sepúlveda-Sanchis[1], G Camps-Valls[1], E Soria-Olivas[1],
S Salcedo-Sanz[2], C Bousoño-Calzón[2], G Sanz-Romero[3] and J Marrugat de la Iglesia[4]

[1]Grup de Processament Digital de Senyals (GPDS), Universitat de València (Spain)
[2]Dpto. Teoría de Señal y Comunicaciones (TSC). Universidad Carlos III de Madrid (Spain)
[3]Hospital Clínic, Barcelona (Spain)
[4]Instituto Municipal de Investigación Médica, Barcelona (Spain)

Abstract

In this communication we present a combination of two state-of-the-art machine learning methods for predicting mortality in patients with unstable angina (UA). Support Vector Machines (SVM) are used as non-linear discrimination tools. However, before building the models, selection of the best subset of variables is carried out with Genetic Algorithms (GA).

The best subset of descriptors selected by the GA was constituted by five variables from the originally 75 collected. The data was split into a training set (483 patients; 22 cases with UA) and a validation set (243 patients; 12 of cases with UA). The criterion used to select the best model was based on the sensitivity (SE), specificity (SP) and negative predictive values (NPV) in the validation data set. The final SVM model (RBF kernel) yielded good results (SE = 66.67%, SP = 79.77% in the validation set). The recognition rate was 79.12% and a high rate of NPV (97.87%) was obtained. Methods proposed have proven to be well-suited for this problem, simplifying the solution and providing excellent discrimination scores.

1. Introduction

Angina is the primary symptom of coronary artery disease and, in severe cases, of a heart attack. Angina is usually referred to as stable (predictable) or unstable (less predictable and a sign of a more serious situation).

The prognosis of patients admitted for acute myocardial infarction (AMI) and unstable angina (UA) has progressively improved in the past 30 years. However, despite advances in the treatment of these diseases, there is still a high in-hospital mortality. The introduction into clinical practice of effective treatments, such as thrombolysis, aspirin, β-blockers, and angiotensin-converting enzyme ACE inhibitors, has changed the prognosis of diseases. More aggressive interventions, such as direct percutaneous transluminal coronary angioplasty (PTCA) might, for selected patients, further decrease in-hospital mortality. Practitioners have a wide variety of reperfusion strategies to interrupt the evolving myocardial event but the efficacy of therapeutic intervention in acute ischemic cardiopathy is strongly time dependent.

The importance of risk assessment is due to the variability in mortality risk, and the time dependence of the efficacy of reperfusion therapy among patients with UA. In this context, the use of classification methods to predict prognosis, might, further decrease in-hospital mortality. To allocate every patient to the most beneficial treatment, the risk profile of every single patient should be available immediately when a patient enters the medical care system. Careful risk assessment for each patient aids clinicians in assessing prognosis and may therefore be a useful guide in management thus providing valuable information.

In this communication we present a combination of two state-of-the-art machine learning methods for detecting patients with unstable angina. We used Support Vector Machines (SVM) as non-linear discrimination tools. Before building the models, selection of the best subset of variables is carried out with Genetic Algorithms (GA).

The paper is outlined as follows. In Section II, data collection and the scope of our study are presented. In Section III methods are detailed. Results in Section IV will precede some concluding remarks.

2. Data collection and scope

The RESCATE (Recursos Empleados en el Síndrome Coronario Agudo y Tiempos de Espera) study consisted of a registry of first AMI and UA patients admitted to one hospital with, and three others without, coronary angiography facilities or coronary surgery. Patients were followed for six months after admission. All four participant hospitals were teaching institutions.

2.1. Inclusion criteria

Between May 1992 and June 1994, all primary UA patients up to the age of 80 years with no history of myocardial infarction admitted to the four participating hospitals were included.

The diagnosis of UA was made when a typical chest pain occurred in any of the following presentations: 1) progressive angina (i.e. increase in the number of *angina pectoris* attacks or progressive decrease in physical exertion in the last month); 2) angina at rest (i.e. ischemic-type chest pain at rest of less than 20 min duration); 3) prolonged angina (i.e. ischemic-type chest pain lasting more than 20 min); and 4) variant angina (i.e. ischemic-type chest pain at rest with ST-segment elevation). Any one of these four types was considered to be new-onset angina when it lasted less than one month. However, new-onset angina *per se* was not considered unstable if it did not meet the criteria for one category of the above classification. Conversely, ischemic electrocardiographic (ECG) changes during symptoms at any time of hospitalization, positive exercise test, significant lesions at coronary angiography or previous diagnosis of angina also had to be present. The diagnosis of AMI was ruled out in all patients by serial enzymatic determinations. Informed consent was obtained from all patients before their inclusion in the cohort, and the study was approved by the ethics committee of the four participating hospitals.

A total of 2661 patients with unstable angina were consecutively admitted to the participating hospitals. Of these patients 839 (31.5%) fulfilled the study inclusion criteria. In addition, only patients containing all characteristics were included and therefore, the final cohort was reduced to 726 patients.

2.2. Exclusion criteria

Exclusion criteria included previous AMI, residence outside of the catchment areas, previous inclusion in the registry or any of the following conditions: life-threatening diseases other than the index event, previous CABG or PTCA, or coronary angiography in the last six months. Patients enrolled in ongoing clinical trials were not excluded so as to reproduce more faithfully the real caring scenarios.

2.3. Primary end points

A composite primary end point included mortality or readmission within six months after the onset of UA for any of the following reasons: AMI, UA, congestive heart failure, sustained ventricular tachycardia or ventricular fibrillation.

2.4. Study variables

The following variables were prospectively recorded by a trained medical investigator at each center: demographic data, history of hypertension, diabetes, chronic obstructive pulmonary disease, peripheral vascular disease, smoking status, previous angina, acute pulmonary edema or cardiogenic shock, ECG changes during admission, presence of severe arrhythmia[1], delay from onset of symptoms to fist monitoring in an emergency room, hospital stay, exercise test, coronary angiography, PTCA and CABG.

3. Methods

Before building a model, the most significant predictors must be selected; otherwise, insignificant parameters could become noise and alter its performance, thus producing an unreasonable outcome. This is especially true when the number of available input variables is large, and exhaustive search through all combinations of variables is computationally infeasible.

3.1. Genetic Algorithms for subset selection

In order to reduce the models' complexity and to circumvent the *curse of dimensionality* we have used Genetic Algorithms (GA) to evaluate the influence of the independent variables on the risk of acute UA. Genetic Algorithms are a class of robust problem solving techniques based on the principles of genetic variation, and natural selection [1, 2].

The most general formulation of a GA performs three steps iteratively: *selection*, *crossover* and *mutation*, which are extensively discussed in the literature [1]. The final goal is to maximize a *fitness* function associated to every individual or coded input space in form of binary strings. This procedure provides a method for efficiently explore the entire solution space.

3.1.1 Solution coding

If we assume that there are n predictor variables X_i, $i = 1, ..., n$ and a response variable Y that labels the class belonging of a specific pattern. We can describe a linear classification model as:

$$Y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_{n-1} x_{n-1} + \alpha_n x_n + \varepsilon \qquad (1)$$

and each possible subset can be described as a binary string of length $n + 1$.

[1] Severe arrhythmia is defined in this study as the occurrence of at least one episode of sustained ventricular tachycardia requiring immediate medical intervention or ventricular fibrillation

3.1.2 Fitness function

Before attempting any GA procedure we must select a fitness function in order to evaluate the solution in each iteration. The individual with higher fitness function has a higher probability of being selected to propagate a new generation. In this paper we have used several information criteria as fitness functions: the Mallow's C_p criterion, the classical Akaike's Information Criteria (AIC) and the Maximum Description Length (MDL) criteria. We have transformed these the-smaller-the-better performance statistic measurements into fitness functions to be maximized by previously rescaling them. This approach is based on the work [3].

3.2. Support Vector Machines

Support Vector Machines (SVM) have been recently proposed as a method for pattern classification and non-linear regression [4]. Their appeal lies in its strong connection to the underlying statistical learning theory where an SVM is an approximate implementation of the method of structural risk minimization, which seeks to minimize an upper bound of the generalization error rather than minimizing the training error. This approach results in better generalization than conventional techniques.

Given a labeled training data set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$, and a non-linear mapping, $\boldsymbol{\phi}(\cdot)$, usually to a higher dimensional space, $\mathbb{R}^d \xrightarrow{\boldsymbol{\phi}(\cdot)} \mathbb{R}^H$ ($H > d$), the SVM method solves:

$$\min_{\mathbf{w},\xi_i,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \right\} \tag{2}$$

subject to the following constraints:

$$y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n \tag{3}$$

$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n \tag{4}$$

where $\mathbf{w}$ and b define a linear regressor in the feature space, non-linear in the input space unless $\boldsymbol{\phi}(\mathbf{x_i}) = \mathbf{x_i}$. In addition, ξ_i and C are, respectively, a positive slack variable and the penalization applied to errors. The parameter C can be regarded as a regularization parameter which affects the generalization capabilities of the classifier and is selected by the user. A larger C corresponds to assigning a higher penalty to the training errors.

A SVM is trained to construct a hyperplane $\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b = 0$ for which the margin of separation is maximized. Using the method of Lagrange multipliers, this hyperplane can be represented as:

$$\sum_i \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}) = 0 \tag{5}$$

where the auxiliary variables α_i are Lagrange multipliers. Its solution reduces to:
Maximize:

$$L_d \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j) \tag{6}$$

subject to the constraints:

$$\sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C, \tag{7}$$

Using the Karush-Kuhn-Tucker Theorem [5], the solution is a linear combination of training examples which lie closest to the decision boundary (the corresponding multipliers are non-zero). Only these examples, called *support vectors*, affect the construction of hyperplane.

The mapping ϕ must guaranty that patterns, non-linearly transformed to a high–dimensional space, are linearly separable. This formulation allows that all the ϕ mappings used in the SVM occur in the form of an inner product. Accordingly, the solution is to replace all occurrences of an inner product resulting from two mappings with the kernel function κ defined as: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j)$. Then, without considering the mapping ϕ explicitly, a non-linear SVM can be constructed by selecting the proper kernel.

4. Results

4.1. Model development

Non-linear classifiers are obtained by taking the dot product in kernel-generated spaces. Some common kernels are the following:

- Linear: $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- Polynomial: $\kappa(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
- Radial Basis Functions (RBF): $\kappa(\mathbf{x}_i,\mathbf{x}_j) = e^{-(\mathbf{x}_i \cdot \mathbf{x}_j)^2/\sigma^2}$

Note that one or more free parameters must be previously settled in the non-linear kernels (polynomial degree d, width σ) together with the trade-off parameter C, usually known as the *penalization* factor. We selected the best subset of free parameters using the four-fold cross-validation method in the training data set. Individual penalization parameter for every training sample was strictly necessary since the distribution of classes is highly unbalanced (4.45% of cases with UA).

All models were developed in MATLAB® environment (Mathworks, Inc). Since computational burden was very high, *m-files* were translated to MEX-files and programs were run in fast workstations[2].

[2]A huge number of trainings were performed in Dual K7-2.3GHz platforms, 1.5GBytes RAM.

4.2. Feature selection

Before building the models, selection of the best subset of variables was carried out with two approaches. Firstly, we selected relevant variables through conventional feature selection techniques (principal component analysis, correlation function, statistical descriptors and entropy measures). The best subset was reduced to 14 variables from the originally 75 collected. However, we continued inspecting smaller subsets with the GA approach (Fig. 1). Despite the optimal subset is formed with only two variables (relatives with schemic cardiopathy and previous diagnosis of angina), there are no significant differences between the solution of five factors (relatives with schemic cardiopathy, previous diagnosis of angina, hipercholesterolemia, habitual smoker and gender), which has revealed more robust in the validation set.

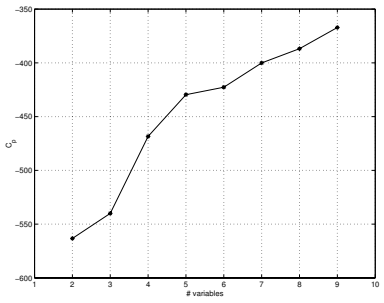

Figure 1. Evolution of Mallow's C_p for different optimal subsets of variables.

4.3. Risk stratification

Once the variables were selected, data was split into a training set (483; 22 cases with UA) and a validation set (243 patients; 12 of cases with UA). Selection of the model was subjected to the following two restrictions:

1. The negative predictive value (NPV%) must be higher than 97.5% since we must reduce as much as possible the rate of false predictions on the true positives.
2. After condition one is ensured, we must increase as much as possible the success rate (SR%). However, in order to obtain well-balanced models, we used the sum of sensitivity (SE) and specificity (SP) factors to select the best model.

The final SVM model (RBF kernel) yielded good results (SE = 66.67%, SP = 79.77% in the validation set) although regularization was a hard problem to solve. The recognition rate was 79.12% and a high rate of negative predictive values (NPV = 97.87%) was obtained.

Table 1. Results in the validation set of SVMs with different reproducing kernels in Hilbert space.

Score	RBF kernel $C = 4.5, \sigma = 9.3$	Polynomial kernel $C = 6, d = 2$	Linear kernel $C = 12$
SR	79.12%	77.55%	75.49%
SE	66.67%	66.66%	65.34%
SP	79.97%	77.56%	74.33%
NPV	97.87%	97.87%	97.87%

5. Conclusions

We have presented a combined strategy of evolutionary algorithms and state-of-the-art machine learning methods that allow for reliable prediction of angina. We can conclude that Support Vector Machines are inexpensive, quick and precise tools for assessment of risk for 6-month mortality. The use of GA has reported knowledge gain about the problem.

Acknowledgements

The authors would like to express their deepest gratitude to the Participating Institutions and Investigators for the Recursos Empleados en el Síndrome Coronario Agudo y Tiempos de Espera (RESCATE) study.

We want to express our thanks to doctors G. Sanz and M. Cardona from the Hospital Clínic i Provincial de Barcelona and to R. Masià, J. Sala, X. Alvert from the Hospital Josep.

References

[1] Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley, 1989.

[2] Michalewicz Z. Genetic algorithms + data estructures = evolution programs. Berlin: Springer-Verlag, 1992.

[3] Chen HY, Chen TC, Min D, Fischer G, Wu YM. Prediction of tacrolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients. Therapeutic Drug Monitoring Feb 1999;21(1):50–56.

[4] Vapnik VN. Statistical Learning Theory. New York: John Wiley & Sons, 1998.

[5] Fletcher R. Practical Methods of Optimization. John Wiley & Sons, Inc. 2nd Edition, 1987.

Address for correspondence:

Gustavo Camps-Valls
Grup de Processament Digital de Senyals. Universitat de València
C/ Dr. Moliner, 50. 46100 Burjassot (València). Spain
tel./fax: ++34-96-3160-197/460
E-mail: gustavo.camps@uv.es