An Overall Performance Comparative of **GA-PARSIMONY** Methodology with Regression Algorithms

Rubén Urraca-Valle, Enrique Sodupe-Ortega, Javier Antoñanzas Torres, Fernando Antoñanzas-Torres, and Francisco Javier Martínez-de-Pisón

EDMANS Group, Department of Mechanical Engineering, University of La Rioja, Logroño, Spain

edmans@dim.unirioja.es

http://www.mineriadatos.com

Abstract. This paper presents a performance comparative of **GA-PARSIMONY** methodology with five well-known regression algorithms and with different genetic algorithm (GA) configurations. This approach is mainly based on combining GA and feature selection (FS) during model tuning process to achieve better overall parsimonious models that assure good generalization capacities. For this purpose, individuals, already sorted by their fitness function, are rearranged in each iteration depending on the model complexity. The main objective is to analyze the overall model performance achieve with this methodology for each regression algorithm against different real databases and varying the GA setting parameters. Our preliminary results show that two algorithms, multilayer perceptron (MLP) with the Broyden-Fletcher-Goldfarb-Shanno training method and support vector machines for regression (SVR) with radial basis function kernel, performing better with similar features reduction when database has low number of input attributes ($\lesssim 32$) and it has been used low GA population sizes.

Keywords: Genetic Algorithm, Tuning Modeling, Feature Selection, Parsimony Criterion, Model Comparative.

1 Introduction

In real industrial and environmental applications, selecting a good overall model with a reduced number of input features has multiple advantages. Those models considering the most relevant input variables can facilitate the understanding of the problem studied. Moreover they are more robust against perturbations, noise and missing values, some important issues that are very common in these cases. Models with less attributes also imply reducing human and cost efforts on information acquiring and preprocessing. For instance, reducing input features in environmental models involves cutting down on costs in data acquisition systems as well as the time to analyze and process the information. Finally, less complex models significantly simplify future model training, tuning and exploiting stages.
Finding the most relevant model inputs and its best setting parameters is still a challenging task in machine learning that depends on multiple factors. More specifically, when trying to reduce the computational cost of these tasks, soft computing (SC) seems to be an effective alternative to other classical approaches \cite{[19][6][7][8][4][11][22]}. Several authors have reported SC strategies that combine feature selection (FS) and model parameters optimization. Huang and Chang \cite{[14]} propose genetic algorithm (GA) combined with $k$-fold cross-validation (CV) evaluation for input FS as well as Support Vector Machine (SVM) tuning in order to improve microarray classification. Ding \cite{[9]} uses particle swarm optimization for simultaneously selecting the best spectral band and optimizing SVM parameters in hyperspectral classification of remote sensing images. Winkler et al. \cite{[23]} report different evolutionary strategies to select inputs in order to optimize linear models, $k$-nearest neighbors (k-NN), artificial neural network (ANN) or SVM. Their objective is to select the best models capable of identifying tumor markers. Chen et al. \cite{[5]} also use an evolutionary approach to simultaneously optimize complexity and weights of learning vector quantification networks for bankruptcy prediction. Sanz et al. \cite{[20]} reported a novel GA-based optimization to create better overall parsimonious ANNs for predicting set points in an annealing furnace of a steel galvanizing industrial plant.

The principal idea in the afore-mentioned works is to select the best models with the lowest number of attributes (or other model complexity criteria). To this end, the majority of them combine different cost and complexity measures into the same fitness function $J$. Some examples are validation error and number of features. The resulting equations usually are similar to the following ones:

$$J = Error_{val} - wN_{FS}$$ \hspace{1cm} (1)

or

$$J = Error_{val} + w\frac{1}{N_{FS}}$$ \hspace{1cm} (2)

where $Error_{val}$ is the validation error, $N_{FS}$ is the number of model inputs and $w$ is a penalty weight which penalizes $J$ according to model complexity. However, the value of $w$ is critical, since it combines these two different measures; and in addition, its estimation is not an easy task.

Therefore, it is still challenging to develop automatic methods to select the best overall parsimonious models. To this end, in the paper published in 2013 by authors \cite{[21]} it is proposed an automatic GA-based optimization, called GA-PARSIMONY, which makes ranking models according to $k$-fold CV error and complexity in a separately way. Thus, it makes unnecessary the use of the penalty weight, $w$. The methodology has been already successfully applied for predicting set points in industrial processes, for solar energy modeling, among other applications. Generally, the resulting models show similar accuracy but with lower number of inputs than using other optimization processes.

In order to obtain more experience on applying this methodology, this paper shows preliminary results and model performance highlights obtained when applying the GA-PARSIMONY to five well-known regression methods with different population sizes and public databases.