Neurocomputing 271 (2018) 9–17

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Evaluation of a novel GA-based methodology for model structure selection: The *GA-PARSIMONY*

R. Urraca, E. Sodupe-Ortega, J. Antonanzas, F. Antonanzas-Torres, F.J. Martinez-de-Pison*

EDMANS Group, Department of Mechanical Engineering, University of La Rioja, Logroño 26004, Spain

ARTICLE INFO

Article history:
Received 23 November 2015
Revised 11 May 2016
Accepted 1 August 2016
Available online 6 July 2017

Keywords:
Genetic algorithms
Parameter tuning
Feature selection
Parsimony criterion
Model comparative

ABSTRACT

Most proposed metaheuristics for feature selection and model parameter optimization are based on a two-termed *Loss + Penalty* function. Their main drawback is the need of a manual set of the parameter that balances between the loss and the penalty term. In this paper, a novel methodology referred as the *GA-PARSIMONY* and specifically designed to overcome this issue is evaluated in detail in thirteen public databases with five regression techniques. It is a GA-based meta-heuristic that splits the classic two-termed minimization functions by making two consecutive ranks of individuals. The first rank is based solely on the generalization error, while the second (named *ReRank*) is based on the complexity of the models, giving a special weight to the complexity entailed by large number of inputs.

For each database, models with lowest testing RMSE and without statistical difference among them were referred as *winner* models. Within this group, the number of features selected was below 50%, which proves an optimal balance between error minimization and parsimony. Particularly, the most complex algorithms (MLP and SVR) were mostly selected in the group of *winner* models, while using around 40–45% of the available attributes. The most basic IBk, ridge regression (LIN) and M5P were only classified as *winner* models in the simpler databases, but using less number of features in those cases (up to a 20–25% of the initial inputs).

1. Introduction

The selection of a good overall model, with optimal generalization ability but with a reduced number of features, has multiple advantages for its implementation in real-world applications. The identification of the most relevant input variables facilitates the understanding of the problem being studied, and it generates more robust models against perturbations, noise and missing values. In this line, a reduction in the number of inputs has a positive impact on the human and economic efforts required for data acquisition and preprocessing. For instance, in environmental applications, it involves cutting down on costs in data acquisition systems as well as reducing the time to analyze and process the information. Finally, the development of less complex models significantly simplifies upcoming stages such as re-calibration and exploiting, and mitigates the well known overfitting issues.

One of the most frequent approaches to tackle overfitting is the use of regularization. This strategy has been included in the training stage of many machine learning algorithms, and it consists in minimizing a *Loss + Penalty* function [1]:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}}\{L(\mathbf{X},\mathbf{y},\beta)+\lambda P(\beta)\} \quad (1)$$

where $L(\mathbf{X},\mathbf{y},\beta)$ is the loss function that evaluates the performance of the model trained (β) given a set of input variables ($\mathbf{X}$) and an outcome ($\mathbf{y}$), and $P(\beta)$ is the penalty function that is related to the complexity of the model. Finally, λ is a non-negative parameter that balances cost and penalty terms in order to control the bias-variance trade-off. This type of regularization strategy is used by multiple methods such as ridge regression (L_2 penalty), LASSO (L_1 penalty), SVM (cost parameter) or ANNs (weight decay). In most of these methods, λ along with other secondary parameters are tuned with some classic optimization algorithms such as grid search (GS) or random search (RS). These optimization methods are combined with some resampling techniques such as *k*-fold Cross-Validation (CV) or Bootstrap to ensure a final model with adequate generalization ability. However, a second validation procedure is still required if other external parameters need to be optimized, such is the case of the number of features and coefficients involved in the data transformation process. This second validation procedure, performed among the best models from the first stage,

* Corresponding author.
E-mail address: fjmartin@unirioja.es (F.J. Martinez-de-Pison).

http://dx.doi.org/10.1016/j.neucom.2016.08.154
0925-2312/

must be again based on both criteria (generalization capability and complexity).

Soft computing (SC) appears as an effective alternative to reduce the computational and human cost of this task compared against the classic approaches [2–9]. In the last years, several authors have reported the use of SC strategies for the model selection process, combining feature selection (FS) and parameter tuning (PT) to generate models with good generalization capabilities [10–12]. For instance, Huang and Chang [13] combined genetic algorithms (GAs) with *k*-fold cross-validation (CV) for FS and tuning of Support Vector Machines (SVM) in order to improve microarray classification. Vieira et al. [14] used binary particle swarm optimization (PSO) to tune a wrapper approach with SVM to predict whether a patient with septic shock survived or deceased. Ahila et al. [15] modified the PSO method to perform FS and tuning of Extreme Learning Machines (ELM) in a power system disturbances classification problem. Dhiman et al. [16] designed a hybrid approach with wavelet packet decomposition and a GA-SVM scheme for FS and MPO to obtain classification models capable of detecting epileptic seizures from background electroencephalogram signals. Castillo et al. [17,18] used ant colony optimization (ACO) to adjust different membership functions of complex fuzzy controllers. Winkler et al. [19] used different evolutionary strategies to perform FS and to optimize linear models, *k*-nearest neighbors (*k*-NN), ANNs and SVM with the final purpose of identifying tumor markers. Sanz-García et al. [20] proposed a GA-based optimization method to create better overall parsimonious ANNs for predicting set points in a steel annealing furnace. Ding [21] used PSO for selecting spectral bands and optimizing SVM parameters in remote sensing.

The main objective of these works is to generate models with the lowest generalization error while maintaining the overall parsimony, which mainly concerns to the number of variables retained as inputs. However, most of these studies include an optimization via a classic two-termed *Loss* + *Penalty* function that requires to set the penalty parameter (Λ). This Λ is similar to the aforementioned λ, but here is used to compare models instead of comparing variations of the same model. Hence, its value has to be manually set prior the execution of the optimization methodology. In this context, we introduced a new GA-based optimization methodology, named *GA-PARSIMONY* [22]. Our aim is to automate the optimization process when the complexity of the model is taken into account by getting rid of the penalty parameter Λ. To do so, we break the traditional *Loss* + *penalty* optimization function by making two consecutive ranks of the individuals. First, individuals are ranked according to a loss term (*k*-fold CV error). Next, the position of individuals with no significant difference in their loss functions is modified based on the complexity of the models (process hereafter referred as *ReRank*). The complexity evaluation accounts for both, the inner complexity of the model and the number of features retained. Therefore, the methodology conducts the tuning of model parameters and feature selection at a time, while boosting the selection of parsimonious models. The methodology has been already successfully applied for predicting set points in industrial processes [20,23,24], for solar energy modeling [25–27] and for structure engineering [28] among other applications. When compared against other optimization methods, the obtained models proved to have similar generalization errors while using a lower number of inputs. The main goal of this work is to perform a more detailed analysis of the GA-PARSIMONY methodology by testing it into five well-known regression methods with different population sizes and public databases.

The remainder of this paper is organized as follows. GA-PARSIMONY methodology is presented in Section 2. The design of the experiments to evaluate the methodology is detailed in Section 3. The different regression techniques used are introduced, as well as the public databases and metrics used for evaluation. Numerical results obtained are presented and discussed in Section 4 and the conclusions drawn are shown in Section 5.

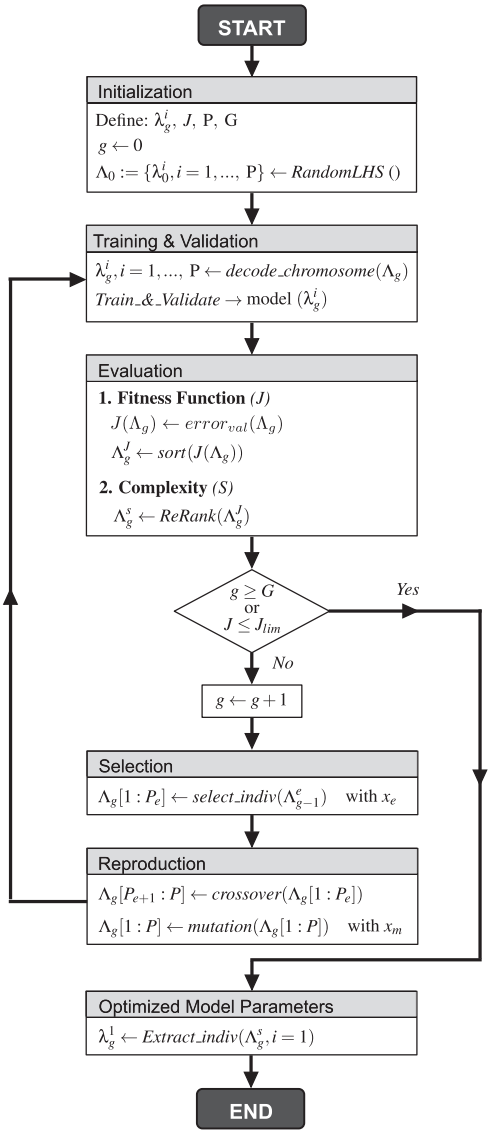

Fig. 1. GA-PARSIMONY optimization methodology.

2. *GA-PARSIMONY* methodology

The objective of the methodology is to automate the model structure selection process. Specifically, feature selection and parameter tuning are simultaneously conducted in order to obtain accurate but parsimonious models. The methodology is referred by authors as GA-PARSIMONY [22], as it combines the traditional GA structure (see Fig. 1) for FS and PT, with the selection of parsimonious models. Here, the main novelty compared to existing proposals is the elimination of the penalty parameter from the fitness function. The procedure begins with the definition of the initial population Λ_0,

$$\Lambda_0 : \{\lambda_0^1, \lambda_0^2, \ldots, \lambda_0^P\}. \quad (2)$$

Hybrid chromosomes λ_g^i are used to select features and tune model parameters. The chromosomes are composed of two different entities: a binary coded vector, with the selected features as inputs to the predictive technique, and a real coded part, with the