Neurocomputing 146 (2014) 83–94

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-objective model type selection

Alejandro Rosales-Pérez [a,*], Jesus A. Gonzalez [a], Carlos A. Coello Coello [b], Hugo Jair Escalante [a], Carlos A. Reyes-Garcia [a]

[a] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Computer Science Department, Luis Enrique Erro No. 1, Santa María Tonantzintla, Puebla 72840, Mexico

[b] Centro de Investigación y de Estudios Avanzados del IPN (CINVESTAV-IPN), Computer Science Department, Evolutionary Computation Group (EVOCINV), Av. IPN No. 2508, San Pedro Zacatenco, Mexico City 07360, Mexico

ARTICLE INFO

Article history:
Received 7 October 2013
Received in revised form
7 May 2014
Accepted 15 May 2014
Available online 11 July 2014

Keywords:
Model type selection
VC dimension
Multi-objective optimization
Ensemble methods

ABSTRACT

Classification is a mainstream within the machine learning community. As a result, a large number of learning algorithms have been proposed. The performance of many of these could highly depend on the chosen values of their hyper-parameters. This paper introduces a novel method for addressing the model selection problem for a given classification task. In our model selection formulation, both the learning algorithm and its hyper-parameters are considered. In our proposed approach, model selection is tackled as a multi-objective optimization problem. The empirical error, or training error, and the model complexity are defined as the objectives. We adopt a multi-objective evolutionary algorithm as the search engine, due to its high performance and its advantages for solving multi-objective problems. The model complexity is estimated experimentally, in a general fashion, for any learning algorithm, through the VC dimension. Strategies for choosing a single model or for constructing an ensemble of models from the resulting non-dominated set are also proposed. Experimental results on benchmark data sets indicate the effectiveness of the proposed approach. Furthermore, a comparative study shows that the obtained models are highly competitive, in terms of generalization performance, with other methods in the state of the art that focus on a single-learning algorithm, or a single-objective approach.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Classification is a common task in supervised learning. Its popularity is due to its use in a wide range of applications, such as medical diagnosis and text categorization. In the machine learning community, several learning algorithms to fit a model have been proposed, including decisions trees, artificial neural networks, and those based on statistical learning. However, to date there is not a universal "best" model; this is referred to as the *No Free Lunch Theorem* [47]. Moreover, many of these learning algorithms have a set of adjustable parameters, called hyper-parameters, whose fine-tuning can affect their generalization ability. Taking that into consideration, one might ask the questions: what learning algorithm should be used for a specific problem? Also, given a learning algorithm, what hyper-parameters values should be chosen? These questions are related to the issue of model selection.

In the literature, there are several studies that address the model selection problem. Among these, some have approached it as an optimization problem, differing in the search technique adopted, including gradient-based methods [1,4,6], grid-search [7], or bio-inspired meta-heuristics such as evolutionary algorithms [8,20,22,32,34], artificial immune systems [2] or particle swarm optimizers [3,18,33]. Grid-search is the simplest one, but it could be time-consuming. Although gradient-based methods tend to be more (computationally) efficient, they are very susceptible to the initial search point and they can easily get trapped in a local optimum. Evolutionary algorithms have gained popularity because of their ease of use and their ability to overcome these shortcomings. Indeed, evolutionary algorithms can be less computationally expensive than grid-search, and are less susceptible to their initial search points than gradient-based methods. Furthermore, evolutionary algorithms do not require gradient information and can be easily parallelized.

Another major issue in model selection is the criterion used for this purpose. In this direction, we can differentiate the works that consider a single-objective criterion and those that consider multiple criteria. The single-objective criterion approaches are generally based on an estimation of the generalization error through the well-known k-fold cross validation [3,18,34,43]. Attention has also been paid to considering multiple criteria.

* Corresponding author. Tel.: +52 222 2663100x3413.
E-mail address: arosales@inaoep.mx (A. Rosales-Pérez).

http://dx.doi.org/10.1016/j.neucom.2014.05.077
0925-2312/© 2014 Elsevier B.V. All rights reserved.

These works typically consider the model performance and some criterion for penalizing the model complexity [2,44]. Others have considered either to minimize the sensitivity and specificity [8,32], or different estimates of the model performance [21,22,37]. Alternatively, multiple criteria have also been approached by simplifying the objectives in a weighted linear combination of these [40] instead of simultaneously optimizing the objectives.

Despite these efforts, most of the existing studies consider a single model type (i.e., the learning algorithm is fixed a priori and the model selection task consists of choosing its hyper-parameters), which could not be the most suitable for a particular problem. To the best of the authors' knowledge, nowadays the works that address both the learning algorithm and the hyper-parameters selection are scarce (e.g., [18,22,43]), and most of them tackle the problem as a single-objective one. Notwithstanding, the disadvantages of using a single-objective approach for hyper-parameters optimization with respect to the generalization performance have been pointed out by several authors [8,21,28].

Inspired from previous ideas, we address both the problem of choosing a learning algorithm and its hyper-parameters during the model selection, which is faced as a multi-objective optimization problem. The error on training samples and the model complexity are considered as the objectives in our formulation. Unlike previous works in which the model complexity estimation depends on the learning algorithm (e.g., the number of support vectors in support vector machines), we propose to estimate it through the VC dimension (for Vapnik–Chervonensky dimension) [46].

The main contribution of this paper is a general model selection framework, whose formulation makes it applicable to any learning algorithm. Additional contributions of the paper are as follows: (i) a multi-objective approach for tackling the model type selection problem (i.e., model type plus its hyper-parameters); (ii) the use of the VC dimension in the model type selection formulation for estimating the model complexity to any model type; and (iii) since the outcome of the multi-objective optimization process is a set of solutions (models), that satisfy an optimal trade-off between the objectives from which a model should be chosen, the strategies proposed for constructing a final classification model from the non-dominated solutions set are an additional contribution. The performance of our proposed approach is assessed on several binary classification benchmark data sets widely used in the literature. The experimental results and comparisons show that our proposal is able to select highly effective classification models.

The remainder of this paper is organized as follows. In Section 2, we describe the VC dimension theory and the way in which it can be estimated in an experimental fashion. Section 3 presents our proposal, describing in detail how the model selection problem is formulated as a multi-objective one. It also describes the proposal for constructing a final model from solutions in the resulting non-dominated front. Section 4 presents the experiments performed to test the validity of our proposal using benchmark data sets, and the results obtained from these. Finally, the main conclusions and future work direction paths are presented in Section 5.

2. VC dimension estimation

Vapnik and Chervonenkis defined the VC dimension [46] as a measure of the capacity of a learning algorithm. The VC dimension is defined through the notion of "shattering", which is described as follows: if we have a set of n samples that can be separated by a set of indicator functions F (functions that map a sample to its corresponding binary label) in all 2^n possible ways, we say that the set of samples is shattered by the set of functions F. The VC dimension can be formally defined as [10]:

> A set of functions F has a VC dimension h if there are h samples that can be shattered by the set of functions F, but there are not $h+1$ samples that can be shattered by the set of functions F.

Notwithstanding that the VC dimension can be seen as a measure of the model complexity [23], exact analytic estimates of this are only known for a few classes of functions (linear models), whereas for many others it is unknown. To overcome this, Vapnik et al. [45] proposed a method to experimentally estimate the effective VC dimension of a model. This approach is based on the best fitting between an analytic formula and measurements of the maximum deviation between the error rates on two independent data sets of varying sizes. Conceptually, this approach can be applied to any learning algorithm [10].

The maximum deviation, $\xi(n)$, of the error rates between two independent labeled data sets is defined as

$$\xi(n) = \max_{\omega}(|\mathrm{err}(\mathbf{Z}_n^1) - \mathrm{err}(\mathbf{Z}_n^2)|) \tag{1}$$

where $\mathbf{Z}_n^1$ and $\mathbf{Z}_n^2$ are two independent labeled data sets of size n, $\mathrm{err}(\mathbf{Z}_n)$ is the error rate on the data set $\mathbf{Z}_n$, and ω is the set of parameters of a binary classifier.

As it is stated in [45], $\xi(n)$ is bounded as follows:

$$\xi(n) \leq \Phi(n/h) \tag{2}$$

where

$$\Phi(\tau) = \begin{cases} 1 & \text{if } \tau < 0.5 \\ a\dfrac{\log(2\tau)+1}{\tau - k}\left(1 + \sqrt{1 + \dfrac{b(\tau-k)}{\log(2\tau)+1}}\right) & \text{if } \tau \geq 0.5 \end{cases} \tag{3}$$

where $\tau = n/h$, and the values of the parameters $a=0.16$ and $b=1.2$ were empirically determined. The value of $k=0.14928$ is determined such that $\Phi(0.5)=1$.

Since the bound in Eq. (2) is tight, it can be assumed that

$$\xi(n) \approx \Phi(n/h) \tag{4}$$

The VC dimension h can be estimated from Eqs. (3) and (4). The maximum deviation $\xi(n)$ can be estimated by simultaneously minimizing the error rate on one labeled set and maximizing the error rate in the other one. This can be accomplished through the following procedure [10,45]:

1. Generate a random labeled set $\mathbf{Z}_{2n}$ of size $2n$.
2. Split the set $\mathbf{Z}_{2n}$ into two sets of size n: $\mathbf{Z}_n^1$ and $\mathbf{Z}_n^2$.
3. Flip the labels of the set $\mathbf{Z}_n^1$ to form $\overline{\mathbf{Z}}_n^1$.
4. Merge the two sets: $\overline{\mathbf{Z}} = \overline{\mathbf{Z}}_n^1 \cup \mathbf{Z}_n^2$, and train the binary classifier with the set $\overline{\mathbf{Z}}$.
5. Evaluate $\mathbf{Z}_n^1$ and $\mathbf{Z}_n^2$ with the trained classifier. Measure the difference of the error rates between the two sets: $\xi(n) = |\mathrm{err}(\mathbf{Z}_n^1) - \mathrm{err}(\mathbf{Z}_n^2)|$.

This procedure gives an estimate of $\xi(n)$ from which an estimate of h can be obtained. In order to reduce the variability in the estimation, this procedure is repeated for different data sets varying the samples sizes $n_1, \ldots, n_k$. Moreover, to reduce the variability due to the random samples, the procedure is repeated several times (m_j) for each sample set of size n_i. The average value for each experiment is taken for each n_i: $\overline{\xi}(n_1), \ldots, \overline{\xi}(n_k)$. The effective VC dimension can be estimated by finding the parameter h^* that best fits $\xi(n)$ with the theoretical formula $\Phi(n/h)$, as follows:

$$h^* = \underset{h}{\mathrm{argmin}} \sum_{i=1}^{k} [\overline{\xi}(n_i) - \Phi(n_i/h)]^2 \tag{5}$$

Download English Version:

https://daneshyari.com/en/article/406572

Download Persian Version:

https://daneshyari.com/article/406572

Daneshyari.com