Neurocomputing 150 (2015) 163–172

Contents lists available at ScienceDirect

Neurocomputing

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

Surrogate-assisted multi-objective model selection for support vector machines

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 23 January 2014
Received in revised form
4 July 2014
Accepted 9 August 2014
Available online 5 November 2014

Keywords:
Model selection
Multi-objective optimization
Support vector machines
Surrogate-assisted optimization

ABSTRACT

Classification is one of the most well-known tasks in supervised learning. A vast number of algorithms for pattern classification have been proposed so far. Among these, support vector machines (SVMs) are one of the most popular approaches, due to the high performance reached by these methods in a wide number of pattern recognition applications. Nevertheless, the effectiveness of SVMs highly depends on their hyper-parameters. Besides the fine-tuning of their hyper-parameters, the way in which the features are scaled as well as the presence of non-relevant features could affect their generalization performance. This paper introduces an approach for addressing model selection for support vector machines used in classification tasks. In our formulation, a model can be composed of feature selection and pre-processing methods besides the SVM classifier. We formulate the model selection problem as a multi-objective one, aiming to minimize simultaneously two components that are closely related to the error of a model: bias and variance components, which are estimated in an experimental fashion. A surrogate-assisted evolutionary multi-objective optimization approach is adopted to explore the hyper-parameters space. We adopted this approach due to the fact that estimating the bias and variance could be computationally expensive. Therefore, by using surrogate-assisted optimization, we expect to reduce the number of solutions evaluated by the fitness functions so that the computational cost would also be reduced. Experimental results conducted on benchmark datasets widely used in the literature, indicate that highly competitive models with a fewer number of fitness function evaluations are obtained by our proposal when it is compared to state of the art model selection methods.

1. Introduction

Supervised classification is the task of learning a function from a labeled dataset. This function is a model that is used to predict the response of future data points from the same problem by mapping a data point from the feature space to a class label. To date, there are many machine learning algorithms that can be used for constructing such model. Among these, support vector machines (SVMs) [5,12,43] are one of the most powerful algorithms. The popularity of SVMs relies on their theoretical background, high performance, and scalability. In spite of this, the effectiveness of SVMs depends on the fine-tuning of a set of parameters (usually called hyper-parameters), such as the kernel type and its parameters. Furthermore, other factors that can affect their performance are the way features are scaled, or the presence of irrelevant/redundant features in a dataset. Therefore, it can be beneficial for the SVMs if the data are first pre-processed. This raises the issue of model selection, which is a crucial step to obtain classifiers with a good performance.

The problem of choosing the hyper-parameters values for an SVM can be seen as an optimization one, where a search engine is used to explore the corresponding hyper-parameters space. A number of methods for tackling it have been proposed for this sake, so far. Most of these methods address this problem by fixing a priori a kernel type and they perform the selection of the hyper-parameters values for that kernel. These methods could be mainly differentiated in two aspects: by the criterion used and by the search strategy adopted for this purpose. Regarding the first aspect, the studies can be differentiated in those that consider a single-criterion and those that consider multiple criteria for guiding the search. Single-criterion approaches [1,3,8,9,27] usually adopt the well-known *k*-fold cross validation to estimate the performance of a given configuration of

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

http://dx.doi.org/10.1016/j.neucom.2014.08.075
0925-2312/

hyper-parameters. Multiple criteria approaches typically consider an estimation of the model performance and a measure of the model complexity (such as the number of support vectors) [2,40]. Others have considered the number of features and an estimation of the generalization error [23], estimates of the bias and variance of the model [36], or the minimization of the errors in positive and negative classes [10,26].

On the other hand, regarding the second aspect, the most commonly adopted techniques are grid search [8,41,42], gradient-based methods [1,7,36,9], and meta-heuristics such as evolutionary algorithms [10,23,26,27,40], artificial immune systems [2], or particle swarm optimization [3].

Grid search is the simplest method for adjusting the values of the hyper-parameters. This strategy requires to discretize the search space, which is attained by the variation of each hyper-parameter with a step size through a wide range of values and the performance of each combination is typically assessed through a k-fold cross-validation technique. Such cross-validation makes grid search a computationally expensive method which is suitable only when few hyper-parameters need to be set. The way in which the search space is discretized is another crucial issue in grid search.

Gradient-based methods are highly efficient and have been successfully applied to hyper-parameter optimization for SVMs. Notwithstanding, they still have some drawbacks. For instance, the objective function has to be differentiable with respect to the hyper-parameters and the kernel, which also needs to be differentiable. Moreover, the effectiveness of these methods highly depends on the initial point chosen for the search, which causes that they can be susceptible to getting trapped in a local optimal solution due to the multimodality of the problem.

Several studies have adopted evolutionary algorithms to alleviate the above-mentioned shortcomings, since they are more robust to local optimal solutions than gradient-based methods. Although these methods can be computationally cheaper than grid search methods, they can still be computationally expensive.

An alternative approach consists of tackling the model selection problem as a supervised learning task through meta-learning [39]. In meta-learning, a number of problems (datasets) are described by a set of features (meta-features) in conjunction with the information about the performance obtained from a set of candidate models; these constitute a meta-dataset. A meta-learner is constructed from the meta-dataset. Given a new problem, the meta-learner is used to predict a model based on its meta-features. Even when meta-learning approaches are more efficient than those based on search techniques, they have some drawbacks. The most important one is that meta-learning depends on the quality of the meta-samples, as well as on the number of problems used for generating a meta-dataset, which could be limited. Recent studies that combine meta-learning with a search strategy have been proposed [18,21,30,31,34]. The main idea behind these methods is to use meta-learning for obtaining an initial suggestion of potential models, which is then used to provide initial search points in the optimization step. Nonetheless, convergence in the optimization stage could be affected if the suggestions given by meta-learning are not good enough.

In spite of the considerable number of studies currently available on SVMs model selection, to the authors best knowledge, little effort has been devoted to considering the selection of both the pre-processing method and the feature selection method in combination with defining the parameters of the SVM. In this paper, we describe a novel approach for SVM model selection through the use of multi-objective optimization. In this case, the preprocessing stage, feature selection, and the hyper-parameters tuning for an SVM are all taken into consideration in the model selection formulation. Estimates of bias and variance of a model are defined as the objectives in our multi-objective formulation. Inspired on the ideas of meta-learning, we address the optimization stage through a surrogate-assisted multi-objective optimization approach. Unlike meta-learning approaches, in which a meta-learner is constructed to obtain an initial suggestion of models, under this formulation, a surrogate is built aiming to approximate the objective functions. The main contribution of this paper is a novel method for performing an SVM model selection (i.e., besides hyper-parameters selection for SVMs, we aim to choose both pre-processing and feature selection methods) with a reduced number of fitness functions evaluations. We assessed the performance of our proposal with a suite of benchmark datasets, widely used in the specialized literature. Our experimental results show that our proposal obtains highly competitive models in terms of generalization performance with a lower number of fitness function evaluations.

The remainder of this paper is organized as follows. In Section 2, we present the bias and variance definitions proposed for classification tasks. Section 3 describes our proposal for tackling the model selection problem for SVMs in classification problems. Next, Section 4 shows the experimental settings and experimental results that show the viability of our proposal. Finally, the main conclusions and some possible paths for future work are presented in Section 5.

2. Bias and variance decomposition in classification problems

From a statistical point of view, the expected error over a sample $\mathbf{x} \in \mathbb{R}^n$ can be decomposed into two components: the squared bias and the variance. The bias-variance decomposition was borrowed from the field of regression, using squared-loss loss function. Based on that definition, several bias-variance decompositions have been proposed in the field of classification using the 0–1 loss function, which is commonly adopted in classification tasks. Roughly speaking, square bias is a measure of the contribution to the error of the central tendency (i.e., the class with the most votes across the multiple predictions) when a model is trained with different datasets. The variance is a measure of the deviations of the central tendency when a model is trained with different datasets [44].

In order to obtain a better generalization error, both components should be minimized. Nevertheless, reducing one of them causes an increment in the other one. This is known as the bias-variance dilemma [4,17,20]. It is said that a model with low bias is too flexible and has a low training error rate, but its generalization capability is poor; this is known as the *overfitting* problem. In contrast, a model with low variance is too simple, has low complexity and does not have the ability to learn the training set and its generalization performance is also poor; this is known as the *underfitting* problem. Therefore, a good model is the one which provides a good trade-off between these two components. So, here we face the model selection task as a multi-objective optimization problem. We used as objectives the estimates of bias and variance, with the aim of selecting the model with the best trade-off between both components.

In classification tasks, different ways to estimate the bias and the variance have been proposed [15,17,22,24,25,44]. The definition proposed by Kong and Dietterich [25] measures the bias directly from the error with respect to the central tendency and the variance is defined as the difference between the error and the bias. Nonetheless, this definition applies to the noise-free cases, and it could lead to negative values for the variance. Kohavi and Wolpert define [24] the bias and variance as a quadratic function of the difference between the probabilities that a sample belongs to a class and that the model is able to predict such class. The advantage is that this definition is applicable to multi-class cases.