Applied Soft Computing 29 (2015) 357–370

Contents lists available at ScienceDirect

Applied Soft Computing

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

Efficient multi-criteria optimization on noisy machine learning problems

Patrick Koch[a,*], Tobias Wagner[b], Michael T.M. Emmerich[c], Thomas Bäck[c], Wolfgang Konen[a]

[a] Cologne University of Applied Sciences, Germany
[b] Institut für Spanende Fertigung, TU Dortmund, Germany
[c] Leiden Institute of Advanced Computer Science, Netherlands

ARTICLE INFO

Article history:
Received 18 April 2013
Received in revised form 30 August 2014
Accepted 6 January 2015
Available online 20 January 2015

Keywords:
Machine learning
Multi-criteria optimization
Efficient Global Optimization
Kriging
Hypervolume indicator

ABSTRACT

Recent research revealed that model-assisted parameter tuning can improve the quality of supervised machine learning (ML) models. The tuned models were especially found to generalize better and to be more robust compared to other optimization approaches. However, the advantages of the tuning often came along with high computation times, meaning a real burden for employing tuning algorithms. While the training with a reduced number of patterns can be a solution to this, it is often connected with decreasing model accuracies and increasing instabilities and noise. Hence, we propose a novel approach defined by a two criteria optimization task, where both the runtime and the quality of ML models are optimized. Because the budgets for this optimization task are usually very restricted in ML, the surrogate-assisted Efficient Global Optimization (EGO) algorithm is adapted. In order to cope with noisy experiments, we apply two hypervolume indicator based EGO algorithms with smoothing and re-interpolation of the surrogate models. The techniques do not need replicates. We find that these EGO techniques can outperform traditional approaches such as latin hypercube sampling (LHS), as well as EGO variants with replicates.

1. Introduction

In machine learning (ML), and in particular in classification, the quality of a learning algorithm can be measured by the number of correctly predicted instances on an usually independent set of test instances. Most learning algorithms are highly influenced by the hyperparameter settings.[1] Thus, finding good hyperparameters is essential for training a precise classifier. For solving this problem, we can define and perform a single-criteria optimization (SCO), where the classification error or any other quality indicator is being minimized. Due to the large runtimes of ML algorithms, it has been shown earlier [1–3] that methods like Efficient Global Optimization (EGO) [4] can help to solve the problem, especially when budgets are limited. In EGO the optimization is internally performed on a surrogate model. The expected improvement (EI) [4] deals as *infill criterion* in an sequential optimization process, recommending which point to explore/to evaluate in the ML task.

However, in most cases the user is not only interested in high-quality prediction models, but also wants that model training and parameter tuning is performed in a reasonable amount of time. These objectives are usually conflicting, demanding for the concept of multi-criteria optimization (MCO). Nowadays, EGO-like algorithms are also available for multi-criteria optimization problems [5]. Up to now, this paradigm has received little attention in the ML community, and only little work has been spent to optimize the mentioned objectives at the same time.

One reason for this lies in the nature of ML parameter tuning, where the evaluation of the learning process is usually subject to noise. Besides the usual noise of labeling the ML data, it is mainly caused by the randomness in the selection of training, validation and test data which lead to different results, even for deterministic ML models. A surrogate model has to cope with such noisy responses. The noise in the quality response will increase if we look for solutions with low runtime, which is usually connected with smaller training set sizes.

Summarizing, the typical challenges of noisy MCO can be formulated as follows: (a) finding a good approximation of the Pareto front, (b) coping with the noise and (c) finding a good solution set with very restricted budgets of function evaluations.

1.1. Research questions

We aim at applying multi-criteria EGO variants to solve two criteria ML tasks. Therefore we can define the following research questions:

Q1 Is multi-criteria optimization with surrogate modeling (Kriging) possible for two criteria in the presence of strong noise?

* Corresponding author. Tel.: +49 226181966216.
E-mail addresses: patrick.koch@fh-koeln.de (P. Koch), wagner@isf.maschinenbau.uni-dortmund.de (T. Wagner), emmerich@liacs.nl (M.T.M. Emmerich), baeck@liacs.nl (T. Bäck), wolfgang.konen@fh-koeln.de (W. Konen).

[1] A hyperparameter is a parameter of the learning algorithm which is set to a specific value prior to the learning phase and controls the ability of the learning algorithm to adapt to new data (generalization).

http://dx.doi.org/10.1016/j.asoc.2015.01.005
1568-4946/© 2015 Elsevier B.V. All rights reserved.

Q2 Is it also possible when the budget for function evaluations (i.e. ML training runs) is very restricted?
Q3 Is it necessary to dampen the noise by averaging over repeated function evaluations, at the price of fewer allowed infills under the given budget?
Q4 Are the multi-criteria EGO approaches better in finding good approximation sets than traditional design of experiments (DoE) techniques, e.g., LHS (see Section 2.3)? Are there significant differences between the different EGO approaches?

The paper is structured as follows: in Section 1.2 we highlight related approaches. The basic idea of surrogate-based optimization is described for single-criteria problems in Section 2.1, and followed by a general introduction of MCO in Section 2.2. In Section 3 we describe the setup of the study for efficient two criteria ML experiments. The experimental results are discussed in Section 4 and we give concluding remarks in Section 5.

1.2. Related work

Because most supervised ML models like Support Vector Machines (SVMs) [6,7] are sensitive to their hyperparameter settings, an optimization is required until an optimal behaviour of the models can be guaranteed. This problem became popular as *model selection* [8,9]. Often hyperparameters of models like SVMs are set by grid search or local search heuristics. E.g., Chapelle et al. [10] proposed an approach based on gradient descent, while Keerthi et al. [11] tuned parameters using a BFGS optimizer. Later, Keerthi et al. [12] also showed that hyperparameter optimization can be efficiently done with BFGS even for large-scale problems. Instead, global optimization heuristics were firstly proposed by Cohen et al. [13], who used Genetic Algorithms (GA) for selecting the best SVM model. Friedrichs and Igel [14] later optimized SVM hyperparameters with the CMA-ES [15,16] and demonstrated a superior performance compared with grid search. Also, Glasmachers and Igel [17,18] presented an improved approach for general Gaussian kernels and handling uncertainty.

Although global optimization methods like Evolutionary Algorithms are suitable for many different problems, they often suffer from requiring too many objective function calls. Alternative approaches are known as response surface methodology (RSM) [19] or model-assisted optimization. In model-assisted optimization a surrogate model of the objective function is learned during the optimization process, which can be used to replace evaluations on the real expensive function. Model-assisted optimization has received a lot of interest with the integration of Kriging surrogate models in the context of design and analysis of computer experiments (DACE) [20]. An overview about surrogate models in Evolutionary Computation has been given by Jin [21]. Kriging surrogate models for reducing the number of function evaluations in Evolutionary Computation have been proposed by Ratle [22], Emmerich et al. [23] and Zhou et al. [24]. Lim et al. [25] describe a generalized evolutionary framework using ensembles and smoothing of surrogate models to generate reliable fitness approximations. In their approach they compare Kriging, polynomial regression and radial basis functions.

In case of noisy evaluations, there is a large body of work on Noisy Kriging-based Optimization (NKO) for single-criteria optimization tasks. Forrester et al. [26] extended the deterministic DACE [20] method to noisy experiments using a method named re-interpolation, which is also used in this work. In the same year, Huang et al. [27] proposed an approach based on the so-called Augmented Expected Improvement (AEI) criterion for noisy evaluations. Picheny et al. [28] introduced the quantile-expected improvement and an online computation time allocation method. A comprehensive overview about these approaches is given in [29]. In this article a benchmark of several NKO-approaches on a variety of well-known hard optimization problems with a steerable amount of additive noise has been performed. In (noisy) ML parameter tuning Konen et al. [3] showed that a model-assisted tuning using Kriging performs better on a set of benchmark problems than other state-of-the-art optimization heuristics. On basis of these results Koch and Konen [2] discovered that tuning with small fractions of the available training data can lead to good parameter settings. But any a-priori setting of the training set size without special problem knowledge remained virtually impossible.

A solution to this issue can be to explore a set of solutions, representing alternatives between small and large training set sizes, so that a suitable size is finally delivered to the user. As a necessity, this approach requires the definition of multiple objectives. In earlier ML research, MCO was firstly proposed by Liu and Kadirkamanathan [30]. They optimized a radial basis function network, where two objectives functions were considered to optimize the differences between the real non-linear system and the non-linear model, and another function to emphasize on simpler models. Freitas [31] and Jin and Sendhoff [32] give comprehensive reviews about the employment of multi-criteria algorithms in ML. Jin [33] advocates to use Pareto-based approaches, covering both supervised and unsupervised learning. For MCO often set-based approaches based on evolutionary multi-objective algorithms (EMOA) are proposed. As a drawback, the required number of real function evaluations is considerably higher for these algorithms. This can be problematic, because the computation time is usually very limited. Instead Knowles and Nakayama [34] discuss the use of surrogate-modeling techniques also for multi-criteria optimization problems. The first EMOA using surrogate models was proposed by Giannakoglou et al. [35], whereas Emmerich and Naujoks [36] introduced Kriging models that make use of error prediction to EMOA. Ascia et al. [37] performed an extensive comparison of state-of-the-art EMOA with an approach based on fuzzy approximation to speed up evaluations. A popular variant of EGO for multi-criteria optimization was given by Knowles [38], the Pareto-EGO (Par-EGO). Later, EGO approaches using the hypervolume as infill criterion were introduced, e.g., the SMS-EGO by Ponweiser et al. [39], or the hypervolume-based EI criterion by Emmerich et al. [40], also referred to as $\mathcal{S}$-metric based Expected Improvement (SExI) [5]. Another approach based on decomposition is the MOEA/D by Zhang et al. [41]. Recently, Zaefferer et al. [42] compared SMS-EMOA, a well-known solver without surrogate modeling, and four cutting-edge multi-criteria solvers with surrogate modeling (SMS-EGO, SExI-EGO, MSPOT and MEI-SPOT) on an optimization task without noise. An overview of the properties of multi-criteria EGO variants was given by Wagner et al. [5]. As an alternative to Kriging-based methods some authors propose topology-based methods using Delaunay regression or SOM to capture the topology of the underlying data [43].

Only few authors, see Knowles et al. [44], address the topic of NKO for multi-criteria optimization tasks. The new contribution of the present paper is that we investigate for the first time (i) hypervolume-based expected improvement in surrogate models for MCO, and (ii) MCO for tuning noisy ML problems. We will show that the proposed NKO methods can achieve good solutions with relatively few function evaluations and that they can cope with the specific noise which arises in the field of ML mainly from the subsampling of the training data.

2. Methods

2.1. Efficient single-criteria optimization

In single-criteria optimization we seek an optimal vector $\vec{x}^* \in S \subseteq \mathbb{R}^n$ of a function $f(\vec{x}) \to \mathbb{R}$ which gives the minimal[2] function value f^* for that function:

$$\min_{\vec{x} \in S} f(\vec{x}) = f^* \tag{2.1}$$

Because in our case the evaluation of f includes a complete training and evaluation of a ML model, the direct optimization of f can be expensive. Instead of performing the optimization on f, we fit a surrogate function $\hat{f}$ approximating the real function f. The function is first evaluated at design points $\vec{x}^{(1)}, \ldots, \vec{x}^{(k)}$, in order to train the approximation function $\hat{f}$. We will denote the real evaluations – also called observations – by $y^{(1)}, \ldots, y^{(k)}$. These k observations are used to build a good start point for the following optimization on the model:

$$init_{des} := (\vec{x}^{(1)}, y^{(1)}), (\vec{x}^{(2)}, y^{(2)}), \ldots, (\vec{x}^{(k)}, y^{(k)}) \tag{2.2}$$

This initial design consisting of k points is usually selected from a point set, which is uniformly distributed over the search space, e.g., by calculating an optimized LHS. Then, a regression function can be fitted based on the initial design. In general, any regression technique can be used, however we prefer to use Kriging surrogate models, which were proposed in the context of Design and Analysis of Computer Experiments (DACE) [20], because Kriging performed best in earlier studies. As another advantage, it can handle more complex fitness landscapes and augments predictions with an estimate of the corresponding uncertainty.

Kriging. Kriging is a regression technique named after the geostatistician Krige [45]; the theory of Kriging was mathematically formalized by Matheron [46]. In Kriging the Best Linear Unbiased Predictor (BLUP) and Kriging variance or uncertainty are used for (error) prediction. This coincides with the conditional mean and variance, whenever the random process is Gaussian [47]. We will proceed here with this Bayesian interpretation also known as Ordinary Kriging (OK) [29]. In a first step the results $\vec{y} = (y_1, \ldots, y_k)^T = f(\vec{x}^{(1)}), \ldots, f(\vec{x}^{(k)})$ of an arbitrary initial design are taken as input to the OK model. For any design point $\vec{x}^{(i)}$ the approximation function

$$Y(\vec{x}^{(i)}) = \mu + Z(\vec{x}^{(i)}) \tag{2.3}$$

[2] Note that for simplicity we always refer to minimization problems in this paper, but maximization problems can easily be transformed to minimization problems. Moreover, we assume the existence of an optimum.