Transfer of Knowledge for Surrogate Model Selection in Cost-Aware Optimization

Zbyněk Pitra[1,2], Jakub Repický[1,3], and Martin Holeňa[1]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic
{pitra, repicky, holena}@cs.cas.cz
[2] Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
[3] Faculty of Mathematics and Physics, Charles University in Prague

1 Introduction

Surrogate model selection is an active-learning approach to cost-aware continuous black-box optimization in domains where the evaluation of the black-box objective function is expensive, e. g., obtained experimentally or resulting from comprehensive simulations. Active reusing of knowledge represented by landscape properties of the objective function accross different tasks can provide additional information for more reliable decisions in terms of a suitable surrogate model and a suitable setting of its hyperparameters. However, research into using metalearing [13] and especially *Exploratory Landscape Analysis* (ELA) [14] in this context is only starting [20]. Our goal is to develop a learning system capable to recommend a surrogate model on the basis of the knowledge obtained in previous black-box optimization tasks.

In this paper, we provide a first step necessary to construct a learning system applying knowledge from previous tasks to a new one: a study of the applicability of ELA to two important kinds of surrogate models – Gaussian processes (GP) [18] and ensembles of regression trees (random forests, RF) [3,4,6]. Results using the noiseless benchmarks of the Comparing-Continuous-Optimisers (COCO) platform [9] in the expensive scenario, where at most 50D evaluations are available, are analysed for statistical dependences between model performance and a broad variety of landscape features.

2 Exploratory Analysis of Fitness Landscapes

In order to achive our goal, the relationships between data properties and surrogate model performances has to be analysed in detail first. Second, the investigated relationships will be used to design a system capable to transfer knowledge about relationships from processed tasks to new ones.

The surrogate model selection problem is analogous to the algorithm selection problem (ASP) formulated in [19] and it aims at selecting the most suitable surrogate model for a specific optimization task. Considering ASP, ELA [14] aims at characterizing the landscape of an investigated function and deriving rules how those characteristics influence the performance of the optimization algorithm.

We analyze relationships between the mean-squared error (MSE) of 29 different settings of GP or RF described in Appendix A.1 and 79 out of 91 ELA features (see also Appendix A.1) which didn't yield constant over 24 noiseless benchmark functions from the COCO framework [9] in dimensions $D = \{2, 5, 10\}$ and their 15 instances[4] for any of the tested GP or RF settings. The datasets consisting of $50D$ points for each instance per function were generated by a random improved Latin Hypercube design [1] covering the input space $[-5, 5]^D$. The overall predictive performance of the surrogate models was tested through 5-fold cross-validation on the generated datasets.

As a first step, we performed a simple correlation analysis using the Spearman correlation coefficient between the MSE of the considered models and the investigated ELA features. However, no single ELA feature was found to be discriminative for surrogate model performance, although a few features were positively (or negatively) correlated with all considered models, which indicates the landscape to be difficult (or easy) for fitting any of them.

As a second step, a classification tree representing a multivariate statistical analysis was built using the obtained results. The resulting tree is depicted in Figure 1. 79 ELA features were classified into 29 classes according to which of the 29 considered settings of GP and RF achieved the lowest MSE for the respective combination of dimension and function among all evaluated settings. Features describing the global structure of the objective function landscape were detected as most distinctive (f12, f15, f61, f62, f64, f70, and f71). Global structure of the landscape can possibly influence the performance of a particular model. Very interesting is the discovered importance of basic features such as dimension (f1) or extreme values of the objective function (f3). In addition, skewness (f35) and the kurtosis (f36) also had influence on surrogate model selection. The last mentioned observations may suggest that even a set of simple features can provide valuable information about the model suitability.

3 Discussion

The results suggest that clear relationships between the performance of the 29 compared settings of GP and RF models and the considered features are not easy to derive. Features describing global properties of the landscape are very useful in case of selection of the surrogate model and its settings. On the other hand, simple features can also provide important knowledge useful for future decisions.

The intended direction for our future research is to apply the obtained knowledge to select a suitable surrogate model for previously unseen data in designing a metalearning system. Another important research direction is to investigate the impact of the sampling strategy in the input space to the resulting landscape features and their relationship with the perfomance of the considered models and their various settings.

[4] Function instances are defined by transformations (translations, rotations, and shifts) of both the search space and function values.

Acknowledgements The reported research was supported by the Czech Science Foundation grant No. 17-01251, by the Grant Agency of the Czech Technical University in Prague with its grant No. SGS17/193/OHK4/3T/14, and by Specific College Research project number 260 453. Computational resources were provided by the CESNET LM2015042 under the programme "Projects of Large Research, Development, and Innovations Infrastructures".

References

1. Beachkofski, B., Grandhi, R.: Improved distributed hypercube sampling. In: Proceedings of the 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. p. 1274. American Institute of Aeronautics and Astronautics (2002)
2. Beirlant, J., Dudewicz, E.J., Györfi, L., Van der Meulen, E.C.: Nonparametric entropy estimation : an overview. International Journal of Mathematical and Statistical Sciences **6**(1), 17–39 (1997)
3. Breiman, L.: Classification and regression trees. Chapman & Hall/CRC (1984)
4. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
5. Chaudhuri, P., Huang, M.C., Loh, W.Y., Yao, R.: Piecewise-polynomial regression trees. Statistica Sinica **4**(1), 143–167 (1994)
6. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. pp. 785–794. KDD '16, ACM (2016)
7. Dobra, A., Gehrke, J.: SECRET: A scalable linear regression tree algorithm. pp. 481–487. KDD '02, ACM (2002)
8. Duvenaud, D.K., Nickisch, H., Rasmussen, C.E.: Additive gaussian processes. In: Advances in Neural Information Processing Systems 24, pp. 226–234. Curran Associates, Inc. (2011)
9. Hansen, N., Finck, S., Ros, R., Auger, A.: Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Tech. Rep. RR-6829, INRIA (2009), updated February 2010
10. Hinton, G.E., Revow, M.: Using pairs of data-points to define splits for decision trees. In: Advances in Neural Information Processing Systems. vol. 8, pp. 507–513. MIT Press (1996)
11. Kerschke, P.: Comprehensive feature-based landscape analysis of continuous and constrained optimization problems using the R-package flacco. ArXiv e-prints (2017)
12. Kerschke, P., Dagefoerde, J.: flacco: Feature-Based Landscape Analysis of Continuous and Constraint Optimization Problems (2017), `https://cran.r-project.org/package=flacco`, R-package v. 1.7
13. Lemke, C., Budka, M., Gabrys, B.: Metalearning: a survey of trends and technologies. Artificial Intelligence Review **44**(1), 117–130 (Jun 2015)
14. Mersmann, O., Bischl, B., Trautmann, H., Preuss, M., Weihs, C., Rudolph, G.: Exploratory landscape analysis. pp. 829–836. GECCO '11, ACM (2011)
15. Murthy, S.K., Kasif, S., Salzberg, S.: A system for induction of oblique decision trees. J. Artif. Int. Res. **2**(1), 1–32 (1994)
16. Neal, R.M.: Bayesian Learning for Neural Networks. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1996)
17. Pitra, Z., Repický, J., Holeňa, M.: Boosted regression forest for the doubly trained surrogate covariance matrix adaptation evolution strategy. ITAT 2018, CreateSpace Independent Publishing Platform, North Charleston, USA (2018)

18. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptative computation and machine learning series, MIT Press (2006)
19. Rice, J.R.: The algorithm selection problem. Advances in Computers, vol. 15, pp. 65 – 118. Elsevier (1976)
20. Yu, H., Tan, Y., Sun, C., Zeng, J., Jin, Y.: An adaptive model selection strategy for surrogate-assisted particle swarm optimization algorithm. pp. 1–8. SSCI '16 (2016)

A Appendix

A.1 Experimental setup

The GP regression model in `gpml` implementation[5] was employed using 9 different covariance functions, listed in Table 1, and constant mean $\mu(\mathbf{x}) = \text{mean}(\mathbf{y})$, where $\mathbf{y}$ are the outputs of the training set. The hyperparameters were optimized with MATLAB's `fmincon` using 5 optimization trials, except for the additive covariance function k_{ADD}, which was optimized with only 3 trials due to its relatively high complexity. The rest of initial values for hyperparameters, together with their bounds are reported Table 2. The initial values for repeated optimization trials were sampled.

The RF models were tested using the full-factorial desing on the ensemble method, splitting method, and error gain function. In addition, the number of trees n_{tree}, the number of points N_t, and the number of dimensions used for training the individual tree n_D were sampled from the values in Table 3. Thus, the RF experimental part sampled RF models from 1600 different settings. MSE (err_{MSE}), variance of predicted y-values (err_{var}), and nearest-neighbor entropy estimator [2] (err_{NN}) were employed as error gain functions (err). In bagged RF, cross-validation pruning [3] was utilized to optimize the tree structure. In addition, the following five regression models were used in leaves: constant, linear, linear with interactions, quadratic without interactions, and full quadratic. The model providing the best fit according to the MSE loss function was always selected for the relevant leaf and appropriate data. In boosted RF, the maximum tree depth was set to 8, in accordance with [6].

Considering decision tree settings regardless the ensemble method, the five splitting methods from the following algoritms were employed: CART [3], SECRET [7], OC1 [15], SUPPORT [5], and a method from [10] (PAIR). The remaining decision tree parameters have been taken identical to settings from [17].

The 91 calculated landscape features were from the following 11 ELA feature sets [11,12]: *y-Distribution*, *Levelset*, *Meta-Model*, *CM-Angle*, *CM-Gradient Homogeneity*, *CM-Convexity*, *NBC*, *Dispersion*, *Information Content*, *Basic*, and *PCA*. Feature sets requiring additional evaluations of the objective function (*Convexity*, *Local Search*, and *Curvature*) and cell-mapping feature sets with high computational or memory requirements in higher dimension (*GCM*, *Barrier Trees*, and *Linear Model*) were omitted. All landscape features were calculated using default settings from [12].

[5] `http://www.gaussianprocess.org/gpml/code/matlab/doc/`

Table 1. Experimental settings of GP covariances: d – metric $d(\mathbf{x}_p, \mathbf{x}_q)$, P – isotropic distance measure $P = l^{-2} I_D$, $\tilde{\mathbf{x}}_p, \tilde{\mathbf{x}}_q$ – inputs augmented by a bias unit, $k^{(i)}\left(\mathbf{x}_p^{(i)}, \mathbf{x}_q^{(i)}\right)$ – one-dimensional k_{SE}, $R \subseteq \{1, \ldots, D\}$ – set of selected degrees of interactions. k_{SE} and k_{RQ} were used in both isotropic and automatic relevance determination (ARD) versions ($k_{\mathrm{SE}}^{\mathrm{ARD}}$, $k_{\mathrm{RQ}}^{\mathrm{ARD}}$).

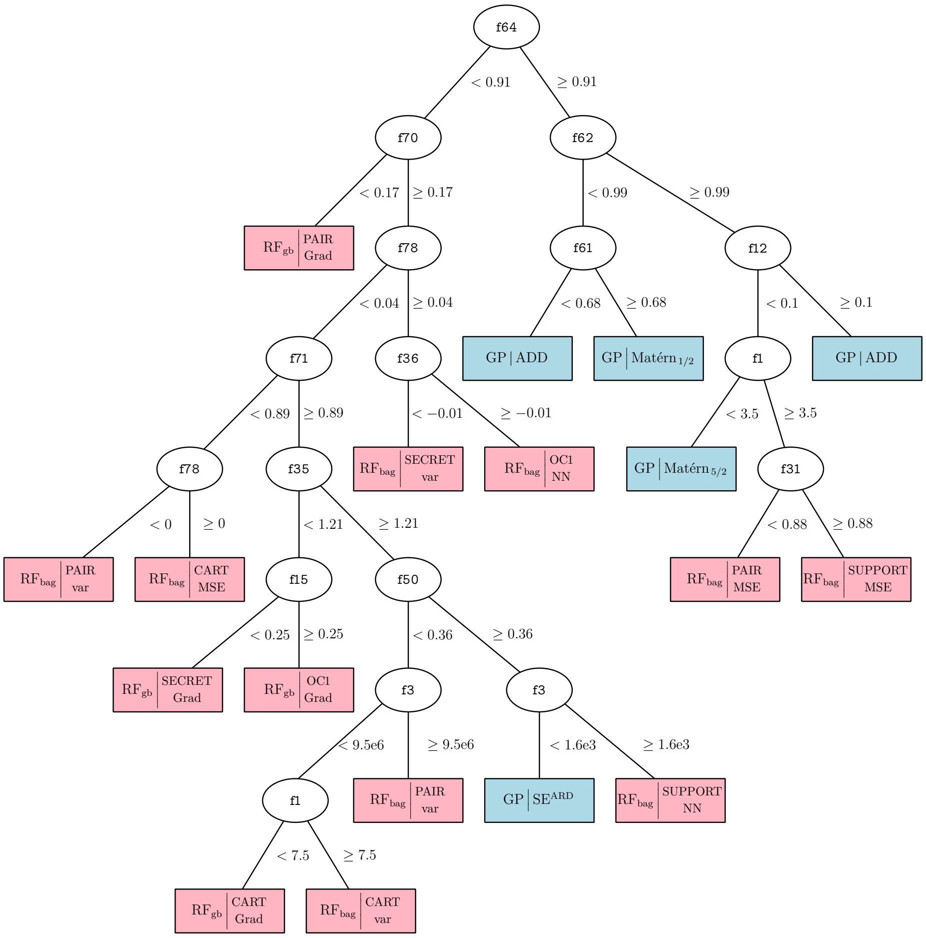
name	kernel
squared-exponential	$k_{\mathrm{SE}}(d; \sigma_f, l) = \sigma_f^2 \exp\left(-\frac{d^2}{2l^2}\right)$
Matérn family	$k_{\text{Matérn}}^{\nu=\frac{1}{2}}(d; \sigma_f, l) = \sigma_f^2 \exp\left(-\frac{d}{l}\right)$
	$k_{\text{Matérn}}^{\nu=\frac{3}{2}}(d; \sigma_f, l) = \sigma_f^2 \left(1 + \frac{\sqrt{3}d}{l}\right) \exp\left(-\frac{\sqrt{3}d}{l}\right)$
	$k_{\text{Matérn}}^{\nu=\frac{5}{2}}(d; \sigma_f, l) = \sigma_f^2 \left(1 + \frac{\sqrt{5}d}{l} + \frac{5d^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}d}{l}\right)$
rational quadratic	$k_{\mathrm{RQ}}(d; \sigma_f, l) = \sigma_f^2 \left(1 + \frac{d^2}{2l^2\alpha}\right)^{-\alpha}$
neural network [16]	$k_{\mathrm{NN}}(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \arcsin\left(\frac{2\tilde{\mathbf{x}}_p^T P \tilde{\mathbf{x}}_q}{\sqrt{(1+2\tilde{\mathbf{x}}_p^T P \tilde{\mathbf{x}}_p)(1+2\tilde{\mathbf{x}}_q^T P \tilde{\mathbf{x}}_q)}}\right)$
additive [8]	$k_{\mathrm{ADD}}(\mathbf{x}_p, \mathbf{x}_q, R) =$ $\sum_{r \in R} \sigma_f^{(r)} \sum_{1 \leq i_1 < i_2 < \cdots < i_r \leq D} \left(\prod_{d=1}^{r} k^{(i_d)}\left(\mathbf{x}_p^{(i_d)}, \mathbf{x}_q^{(i_d)}\right)\right)$

Table 2. Experimental settings of GP. σ_n and l apply to all covariances. σ_f applies to all except the k_{ADD}, in which $\sigma_f^{(r)}$ scales each degree $r \in R$, $R = \{1, 2, 3, 5, 7, 10\} \cap \{1, \ldots, D\}$ of interaction separately. $\sigma_f^{(r)}$ and its upper bound were initialized proportionally to $\binom{D}{r}$.

GP hyperparam	initial value	constrains
σ_n	1e−2	[1e−3, 1e1]
l	std(X)	[1e−2, 1e2]
σ_f	$\frac{\mathrm{std}(\mathbf{y})}{\sqrt{2}}$	[1e−2, 1e6]

Table 3. Experimental settings of RF: n_{tree} – number of trees in RF, N_t, n_D – number of tree points and dimensions, N – number of RF points, D – input space dimension, $\mathrm{err}_{\mathrm{Grad}}$ – gradient error gain. Split methods and error gain functions err are tested using full-factorial design, n_{tree}, N_t, and n_D are sampled.

RF param	bagging	boosting
err	$\{\mathrm{err}_{\mathrm{MSE}}, \mathrm{err}_{\mathrm{var}}, \mathrm{err}_{\mathrm{NN}}\}$	$\mathrm{err}_{\mathrm{Grad}}$
split	{CART, SECRET, OC1, SUPPORT, PAIR}	
n_{tree}	$\{64, 128, 256, 512, 1024\}$	
N_t	$\lceil\{0.25, 0.5, 0.75, 1\} \cdot N\rceil$	
n_D	$\lceil\{0.25, 0.5, 0.75, 1\} \cdot D\rceil$	



Legend	
f1	basic.dim
f3	basic.objective_min
f12	cm_angle.y_best2worst_std
f15	cm_convexity.convex_soft
f31	dispersion.ratio_mean_25
f35	ela_distribution.skewness
f36	ela_distribution.kurtosis
f50	ela_levelset.mmce_lda_50
f61	ela_metamodel.lin_w_interact_adj_r2
f62	ela_metamodel.quad_simple_adj_r2
f64	ela_metamodel.quad_w_interact_adj_r2
f70	nearest_better.nb_std_ratio
f71	nearest_better.nb_mean_ratio
f78	pca.pca_pc1_cov_init

Figure 1. Classification tree demonstrating the influence of ELA features on model suitability. Light blue: Gaussian processes. Light pink: Random forests. The RF models use the notation $\mathrm{RF}_{\text{ensemble method}}\Big|^{\text{split method}}_{\text{error gain}}$, where bag = bagging, gb = gradient boosting, and Grad = gradient error gain. The GP models use the notation $\mathrm{GP}\,\Big|\,\text{covariance function}^{\text{ARD}}$, where $^{\text{ARD}}$ = automatic relevance determination (scaling in each dimension separetely).