Avoiding overfitting in surrogate modeling: an alternative approach

Keywords: Surrogate modeling, LRM, cross-validation, overfitting

Huu Minh Nguyen HUUMINH.NGUYEN@UGENT.BE
Ivo Couckuyt IVO.COUCKUYT@UGENT.BE
Yvan Saeys YVAN.SAEYS@UGENT.BE
Luc Knockaert LUC.KNOCKAERT@UGENT.BE
Tom Dhaene TOM.DHAENE@UGENT.BE
Ghent University - IBBT, Sint-Pietersnieuwstraat 25, 9000, Gent, Belgium

Dirk Gorissen DIRK.GORISSEN@SOTON.AC.UK
University of Southampton, Room 2041, Building 25, Highfield Campus, School of Engineering Sciences, University of Southampton, SO17 1BJ, UK

1. Introduction

In many simulation applications , performing routine design tasks such as visualization, design space exploration or sensitivity analysis quickly becomes impractical due to the (relatively) high cost of computing a single design (Forrester et al., 2008). Therefore, in a first design step, surrogate models are often used as replacements for the real simulator to speed up the design process (Queipo et al., 2005). Surrogate models are mathematical models which try to generalize the complex behavior of the system of interest, from a limited set of data samples to unseen data and this as accurately as possible. Examples of surrogate models are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Kriging models and Radial Basis Function (RBF) models. Surrogate models are used in many types of applications, however in this work we concentrate on noiseless simulation data, as opposed to measurement data or data coming from stochastic simulators.

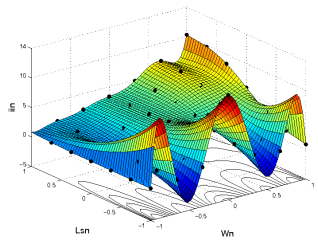
An important consideration when constructing the surrogate model is the selection of suitable hyperparameters, as they determine the behavior of the model. Finding a good set of hyperparameters is however non-trivial, as it requires estimating the generalization ability of the model. When dealing with sparse data, the search for good hyperparameters becomes even harder. Bad hyperparameters will lead to models with high training accuracy (as shown in Fig. 1(a)) but which fail to capture the true behavior (Fig. 1(c)) and instead exhibiting artificial ripples and bumps.

A common strategy for optimizing model hyperparameters when only a limited amount of training data is available, is cross-validation. However, because many models have to be trained, performing cross-validation can be quite time and resource consuming, especially if the cost of model building is high. Moreover, cross-validation is not always efficient at preventing artificial model behavior (Gorissen et al., 2009).

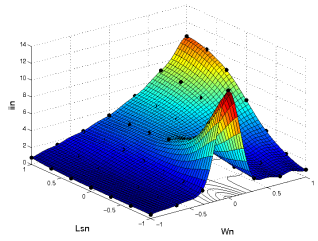
We present in this work a new generic auxiliary model selection measure, called the Linear Reference Model (LRM), which is designed to be fast to compute and which reduces the chance of overfitting. LRM identifies regions where the model exhibits complex behavior (such as oscillations) but lacks the data to support this and penalizes the model accordingly. Overfitted regions are identified by comparing the predicted output values of the surrogate model to that of a local linear interpolation. Large deviations are an indication of overfitting, and models are penalized more heavily if they diverge more from the local linear interpolation. Note that the risk of underfitting is usually neglible as the high accuracy typically required in surrogate modeling can only achieved by using high complexity models, in which case LRM will only reduce their tendency to overfit but never to the extent that the models will underfit. Fig. 1(b) shows the effect of applying the LRM measure. Although the training accuracy of the model is now worse, the intermediate surrogate model is better at capturing the true behavior of the system.

The LRM score is calculated as follows. First, a Delau-

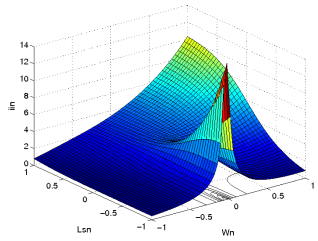
Appearing in *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands.*



(a) Overfitted model exhibiting artificial behavior



(b) Minimum LRM score



(c) True function

Figure 1. surrogate models of the input noise current ($\sqrt{\overline{i_{in}^2}}$) of a Low Noise Amplifier (Gorissen et al., 2009) generated with different model selection criteria. The dots represent a sparse intermediate training during model construction (7×7 samples).

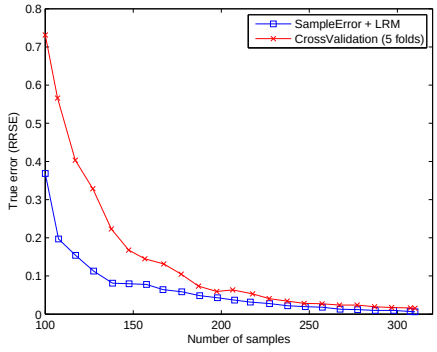


Figure 2. Academic example error on the independent test set as function of the number of data samples for "cross-validation" and "LRM in conjuction with the in-sample error".

nay tesselation of the input space is contructed after which, for each simplex given by the tesselation, a hyperplane through the corresponding samples is built. These hyperplanes form the local linear interpolation which will be used as reference for the surrogate model. Next, the surrogate model is compared with the local linear interpolation at every simplex, and the difference between the two is calculated. The LRM score for the surrogate model is then simply the average difference over all simplices.

We applied the LRM measure to both an analytic academic example and a real world application (Gorissen et al., 2009) using an adaptive model building scheme. In this scheme, a sequential sampling algorithm adds a small number of new samples to the training data at each iteration after which the surrogate model is rebuilt. Both cross-validation and LRM are then used to evaluate the updated models. Our experiments show that, in this context, the accuracies of the models selected by LRM converge faster and are better or comparable to accuracies of models selected using cross-validation. This is illustrated in Fig. 2 for the academic example, where the accuracy on an independent test set is plotted as a function of the number of selected samples. When the number of training samples is relatively small and the models are prone to overfitting, LRM (in combination with the in-sample error) achieves much lower errors than cross-validation. As the number of training samples increases, the difference in accuracy of the models selected by both approaches diminishes. However, LRM achieves this accuracy at much reduced computational cost and can thus provide an interesting alternative to cross-validation in simulation-based engineering design.

References

Forrester, A., Sobester, A., & Keane, A. (2008). *Engineering design via surrogate modelling: A practical guide*. Wiley.

Gorissen, D., De Tommasi, L., Crombecq, K., & Dhaene, T. (2009). Sequential modeling of a low noise amplifier with neural networks and active learning. *Neural Computing and Applications*, *18*, 485–494.

Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., & Tucker, P. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, *41*, 1–28.