10th World Congress on Structural and Multidisciplinary Optimization
May 19 - 24, 2013, Orlando, FL

MODEL SELECTION BASED ON GENERALIZED-REGIONAL ERROR ESTIMATION FOR SURROGATE

Ali Mehmani[1], **Souma Chowdhury**[2], **Jie Zhang**[3], **Achille Messac**[4]

[1] Syracuse University, Syracuse, NY, USA, amehmani@syr.edu
[2] Syracuse University, Syracuse, NY, USA, sochowdh@syr.edu
[3] National Renewable Energy Laboratory, Golden, CO, USA, jie.zhang@nrel.gov
[4] Syracuse University, Syracuse, NY, USA, messac@syr.edu, Corresponding Author

Abstract

The analysis of complex system behavior often demands expensive experiments or computational simulations. Surrogates modeling techniques are often used to provide a tractable and inexpensive approximation of such complex system behavior. Owing to the lack of knowledge regarding the suitability of particular surrogate modeling techniques, model selection approach can be helpful to choose the best surrogate technique. Popular model selection approaches include: (i) split sample, (ii) cross-validation, (iii) bootstrapping, and (iv) Akaike's information criterion (AIC) (Queipo et al. 2005; Bozdogan et al. 2000). However, the effectiveness of these model selection methods is limited by the lack of accurate measures of local and global errors in surrogates.

This paper develops a novel and model-independent concept to quantify the local/global reliability of surrogates, to assist in model selection (in surrogate applications). This method is called the Generalized-Regional Error Estimation of Surrogate (G-REES). In this method, intermediate surrogates are iteratively constructed over heuristic subsets of the available sample points (i.e., intermediate training points), and tested over the remaining available sample points (i.e., intermediate test points). The fraction of sample points used as intermediate training points is fixed at each iteration, with the total number of iterations being pre-specified. The estimated median and maximum relative errors for the heuristic subsets at each iteration are used to fit a distribution of the median and maximum error, respectively. The statistical mode of the median and the maximum error distributions are then determined. These mode values are then represented as functions of the density of training points (at the corresponding iteration). Regression methods, called *Variation of Error with Sample Density (VESD)*, are used for this purpose. The VESD models are then used to predict the expected median and maximum errors, when all the sample points are used as training points.

The effectiveness of the proposed model selection criterion is explored to find the best surrogate between candidates including: (i) Kriging, (ii) Radial Basis Functions (RBF), (iii) Extended Radial Basis Functions (ERBF), and (iv) Quadratic Response Surface (QRS), for standard test functions and a wind farm capacity factor function. The results will be compared with the relative accuracy of the surrogates evaluated on additional test points, and also with the prediction sum of square (PRESS) error given by leave-one-out cross-validation.

The application of G-REES to a standard test problem with two design variables (Branin-hoo function) show that the proposed method predicts the median and the maximum value of the global error with a higher level of confidence compared to PRESS. It also shows that model selection based on G-REES method is significantly more reliable than that currently performed using error measures such as PRESS. The application of G-REES method to high dimensional problems (results to be included in the full manuscript) is expected to further establish the unique benefits of this new model selection method.