NOVEL TOOL TO SELECT SURROGATE MODELING TECHNIQUE FOR DESIGN SPACE APPROXIMATION

B. Williams and S. Cremaschi*
Auburn University
Auburn, AL 36849

Abstract Overview

Surrogate models are used to map input data to output data when the actual relationship between the two is unknown or computationally expensive to evaluate. We have constructed a tool to recommend the appropriate surrogate modelling technique for a given dataset using attributes calculated from the input and output values. The tool identifies the appropriate surrogate modeling techniques with an accuracy of 98% and a precision of 91%.
process design/optimization, design space approximation, random forests, artificial neural networks

Introduction

Surrogate models, also known as response surfaces or black-box models, can be used to reduce computational cost by approximating more complex, higher order models (Wang et al., 2014). Surrogate modeling techniques are of particular interest where high-fidelity, thus expensive, simulations are used (Han and Zhang, 2012) or when the fundamental relationship between the design variables and output variables is not well understood, such as in design of cell or tissue manufacturing processes (Machin, M. et al., 2011). These techniques have been receiving increasing attention in a wide range of applications, for example, in optimization of process design, scheduling, and control (Burnak et al., 2019).

Several machine learning and regression techniques have been developed for constructing surrogate models. Current common practices for selecting which surrogate model form is appropriate rely on process specific expertise. Numerous studies have been comparted the performance of surrogate modeling techniques (Davis et al., 2017; Bhosekar and Ierapetritou, 2018). The majority of these only compare a few models on a limited number of functions or applications. Progress has been made in recent works in generalizing the process for selecting a surrogate model to approximate a design space by using meta-learning approaches to build selection frameworks (Garud et al., 2018; Cui et al., 2016). These frameworks provide "best" recommendations for surrogate modeling techniques, based on the attributes calculated from the data being modeled. In addition, the framework developed by Garud et al. (2018) gives a ranking of all the considered surrogate models based on the predicted accuracy of the model. However, neither framework takes model complexity into account, which can lead to overfitting, or considers that multiple models might perform similar to the one identified as best in terms their accuracies.

To address the knowledge gap, this work compares the performance of eight different surrogate modeling techniques on a collection of generated datasets. Using information extracted from those datasets and building upon previous meta-learning approaches, we construct a tool to provide recommendations for the appropriate modeling techniques for the datasets based only on the characteristics of the data being modeled. The performance metric used to evaluate the model performance is the adjusted-R^2 value (Miles, 2014), which balances the model accuracy with the size, or complexity, of the model. Data sets for training surrogate models were generated from a large set of test functions with different characteristics. The effect both the

* To whom all correspondence should be addressed

underlying shape of the function used to generate the dataset and the number of inputs on the performance of each technique is assessed to provide guidance on which surrogates provide the best predictions and give general "rules of thumb." Additional characteristics, i.e., attributes, were calculated for each dataset with the goal of representing its overall behavior. These attributes were used as inputs, with the actual adjusted-R^2 values as outputs, to train random forest models to provide predictions of adjusted-R^2 values for each technique. Based on the predicted adjusted-R^2 values, the tool identifies which surrogate modeling techniques are recommended for use given a set of data.

Computational Experiments

Surrogate Model Performance Comparison

To evaluate the performances of the surrogate modeling techniques, 1000 input-output pairs were generated from each test function using Sobol sequence sampling (Joe and Kuo, 2003). Eight surrogate modeling techniques are used for comparison: multivariate adaptive regression splines (MARS);(Friedman, 1991), random forests (RF);(Brieman, 2001), single hidden layer feed forward artificial neural networks (ANN);(Haykin, 2009), extreme learning machines (ELM);(Huang et al., 2006), Gaussian process regression (GP);(Rasmussen and Williams, 2006), support vector machines (SVM);(Drucker et al., 1997), Automated Learning of Algebraic Models using Optimization (ALAMO);(Cozad et al., 2014), and radial basis function networks (RBFN);(Jin et al, 2001). Models were trained using the input-output pairs with each of the surrogate modeling technique for each function. This resulted in 808 surrogate models.

When necessary, the hyperparameters of each surrogate modeling technique (such as the number of hidden neurons for ANNs) were optimized prior to training the models for each dataset. After the surrogate models were trained for each dataset, the adjusted-R^2 values were calculated for each modeling technique-dataset pair.

Recommendation Tool Construction

Cui et al. (2016) and Garud et al. (2018) extract information from the datasets for use in their recommendation frameworks in the form of attributes. The attributes include common statistical measures, such as mean and standard deviation, gradient based attributes, and attributes related to the extrema of the output values. We have defined additional attributes related to both the estimated gradients of the datasets and the extreme values of the outputs to use as potential inputs for predicting the model performance with our recommendation tool, resulting a total of 32 attributes.

A random forest model was trained for each surrogate modeling technique to predict its adjusted-R^2 value using the identified attributes as inputs. Feature reduction was performed on the attributes to determine which attributes had the most influence on the predicted output value for each modeling technique. Each technique had a different set of selected attributes for prediction, with the only common attribute among all the techniques being the average value of the gradient estimates. For each dataset, based on the adjusted-R^2 values, the surrogate modeling techniques were classified as either being recommended or not recommend for both the predicted and actual metric values. These classifications were compared and used to evaluate the quality of the selection recommendations.

Adjusted-R^2 for Surrogate Model Selection

The formula for calculating adjusted-R^2 ($\hat{R}^2$) is shown in Eq. (1).

$$\hat{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right] \quad (1)$$

In Eq. (1), R^2 is the R-squared, n is the number of data points in the training set, and k is the number of model parameters (or hyperparameters).

Classification Evaluation Metrics

The metrics used to evaluate the performance of the recommendation tool (i.e., the classification of surrogate modeling techniques given a dataset) are accuracy, precision, recall (Sokolova and Lapalme, 2009), and the Matthews correlation coefficient (MCC). The MCC (Matthews, 1975) is the correlation between actual and predicted classification. It has a value between -1 and 1, with one being a perfect correlation, -1 being a completely negative correlation, and 0 being no correlation or random assignment of classifications. Five-fold cross validation was used to evaluate the performance of the recommendation tool.

Results

Adjusted-R2 Performance

Adjusted-R^2 values were calculated for all the modeling techniques for each dataset. The percentage of the time each surrogate model had the highest adjusted-R^2 value was used to calculate the fraction of the available datasets for which a technique was identified as being the most accurate (Fig. 1). The number of datasets included in each category is included below the *x*-axis.

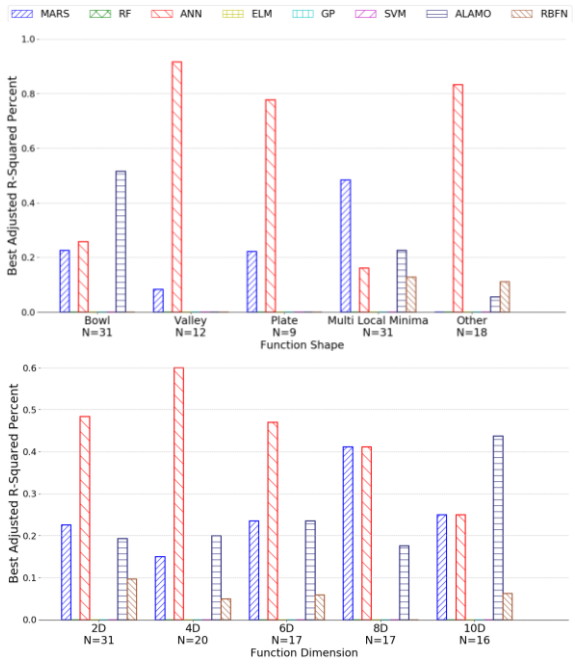

Figure 1. Percentage of datasets for which each model had the highest adjusted-R2 when datasets are grouped by (a) function shape and (b) input dimension

When the datasets are grouped by the function shape, ANN models have the highest adjusted-R^2 values. For bowl and plate shaped functions, ALAMO and MARS models, respectively, give the highest values for the largest percentage of the datasets. When the datasets are grouped by input dimension, ANN is the best performing model the highest percentage of the time at low input dimension. However, as the dimension increases, other models begin to perform as well or better than ANN models. This result indicates that there is some dependence of the surrogate model performance on the overall shape of the function the dataset was generated from and on the number of inputs.

Recommendation Tool Performance

The surrogate model selection tool identified which surrogate modeling techniques should be recommended for a dataset with an accuracy of 85%. The precision, or the probability that a recommended technique should actually be recommended, was 91%.

Conclusions

Selection of the appropriate surrogate modeling technique depends on the characteristics of the dataset being modeled. In general, MARS and ANN models give the most accurate predictions for approximating a design space. We have identified attributes of datasets that are appropriate for use in predicting the adjusted-R^2 value for a technique. Using these attributes, we have constructed a random forest model-based tool that can recommend appropriate surrogate modeling techniques for use with a dataset with a 98% accuracy.

Acknowledgments

This work was partially funded by Department of Education GAANN grant #P200A150075, NSF grant #1743445, and RAPID Manufacturing Institute, U.S.A. The authors would also like to acknowledge the Auburn HPC cluster for support on this work.

References

Bhosekar, A., Ierapertritou, M. (2018). Advances in surrogate based modeling feasibility analysis, and optimization: A review. *Comput. Chem. Eng.*, 108, 250.

Brieman, L. (2001). Random Forests. *Machine Learning*, 45, 5.

Burnak et al. (2019). Integrated process design, scheduling, and control using multiparametric programming. *Comput. Chem. Eng.*, 125, 164-184.

Cozad, A. et al. (2014). Learning Surrogate Models for Simulation Based Optimization. *AICHE Journal*, 60, 2211.

Cui, C. et al. (2016). A recommendation system for meta-modeling: A meta-learning based approach. *Expert Systems with Applications*, 46, 33.

Davis, S. et al. (2017). Efficient surrogate model development: Optimum model form based on input function characteristics. *Proceedings of the 27th European Symposium on Comp. Aided Process Engineering*, 27.

Drucker, et al. (1997). Support Vector Regression Machines. *Neural Information Processing*, MIT Press, 155-161.

Friedman, J. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics,* 19, 1.

Garud, S. et al. (2018). Learning based Evolutionary Assistive Paradigm for Surrogate Selection. *Comput. Chem. Eng.*, 119, 352.

Han, Z. and Zhang, K. (2012). Surrogate-Based Optimization. *Real-World Applications of Genetic Algorithms*. InTech Europe, Rijeka, Croatia.

Haykin, S. (2009). Neural Networks and Learning Machines. Prentice Hall PTR.

Huang, G. et al. (2006). Extreme Learning Machine Theory and Applications. *Neurocomputing*, 70, 489.

Jin, R. et al. (2001). Comparative Studies of Metamodelling Techniques under Multiple Modeling Criteria. *Structural and Multidisciplinary Optimization,* 23, 1.

Joe, F. and Kuo,F. (2008). Constructing Sobol' sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30, 2635.

Machin, M. et al. (2011). Implementation of modeling approaches in the QbD framework: examples from the Norvartis experience. *Eur Pharm Rev,* 16, 39-42.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBE*, 405(2), 442-451.

Rasmussen, C. and Williams, C. 2006 Gaussian Processes for Machine Learning, Cambridge, MA: MIT Press.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Comm. Com. Inf. Sc,* 45(4), 427-437.

Surjanovic, S., Bingham, D. (2013). "Virtual Library of Simulation Experiments: Test Functions and Datasets. http://www.sfu.ca/~ssurjano.

Wang, C. et al. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environ. Modell. Softw.,* 60, 167.