Metin Türkay and Erdal Aydın (Eds.)
Proceedings of the 31st European Symposium on Computer Aided Process Engineering
June 6th to 9th, 2021, Istanbul, Turkey.

Novel Tool for Selecting Surrogate Modeling Techniques for Surface Approximation

Bianca Williams and Selen Cremaschi*

Department of Chemical Engineering, Auburn University, Auburn, AL, United States

selen-cremaschi@auburn.edu

Abstract

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 for several applications, including surface approximation and surrogate-based optimization. Many techniques have been developed for surrogate modeling; however, a systematic method for selecting suitable techniques for an application remains an open challenge. This work compares the performance of eight surrogate modeling techniques for approximating a surface over a set of simulated data. Using the comparison results, we constructed a Random Forest based tool to recommend the appropriate surrogate modeling technique for a given dataset using attributes calculated only from the available input and output values. The tool identifies the appropriate surrogate modeling techniques for surface approximation with an accuracy of 87% and a precision of 86%. Using the tool for surrogate model form selection enables computational time savings by avoiding expensive trial-and-error selection methods.

Keywords: surrogate model, process design/optimization, surface approximation

1. Introduction

Surrogate models are simplified approximations of more complex, higher-order models. They are used to map input data to outputs when the actual relationship between the two is unknown or computationally expensive to evaluate. Surrogate models are of particular interest where expensive simulations are used or when the fundamental relationship between the design variables and output variables is not well understood, such as in the design of cell manufacturing processes (Williams et al., 2020). Surrogate models can also be constructed for surrogate-based optimization when a closed analytical form of the relationship between input data and output data is not available or is not conducive for use in conventional gradient-based optimization methods. Several techniques have been developed for surrogate modeling, requiring a systematic approach for selecting which technique may be appropriate for an application.

Current standard practices for selecting which surrogate model form is appropriate rely on process-specific expertise. Numerous studies have been conducted to compare surrogate modeling techniques (Davis et al., 2017). However, most of these only evaluate a few models on a limited number of functions or applications. Recently, progress has been made in generalizing the process for selecting a surrogate model to approximate a surface by using meta-learning approaches to build selection frameworks (Cui et al., 2016; Garud et al., 2018). These frameworks provide “best” recommendations for surrogate modeling techniques based on the attributes calculated from the data being modeled and avoiding expensive trial-and-error methods. Few of

the developed meta-learning tools take model complexity into account, which can lead to overfitting, or consider that multiple models might perform similarly to the one identified as best.

This work aims to comprehensively investigate the performance of several different surrogate modeling techniques for approximating smooth, continuous functional relationships and to link that performance to the characteristics of the data being modeled. The performance metric used for evaluating how well the surrogate modeling techniques approximate surfaces is the adjusted R^2, which considers both model accuracy and complexity. Simulated data was generated using a suite of optimization test functions. Data attributes were calculated based only on input and output values for each dataset to represent its overall behavior. Attributes that have the most influential relationships for predicting the adjusted R^2 were selected using feature reduction. These attributes were used as inputs to construct a Random Forest based tool to make predictions on the surrogate models' performance and provide recommendations for which surrogate modeling technique(s) may be most accurate for the dataset.

2. Computational Experiments

2.1 Test Functions

The test functions used to simulate data for constructing the surrogate models and the recommendation tool are from the Virtual Library of Simulation Experiments optimization test suite (Surjanovic & Bingham, 2013). The functions are divided by their shapes, which include the categories: multi-local minima (29 functions), bowl-shaped (31 functions), plate-shaped (9 functions), valley-shaped (12 functions), and other-shaped (18 functions) that do not fit into the other four categories. Functions with two (XX functions), four (XX functions), six (XX functions), eight (XX functions), and ten (XX functions) inputs were used.

2.2 Surrogate Model Performance Comparison

Input-output pairs were generated from each test function using three different space-filling sampling methods: Halton Sequence Sampling, Sobol Sequence Sampling, and Latin Hypercube Sampling (LHS). Data was generated at seven different sample sizes sizes (50, 100, 400, 800, 1200, and 1600 samples), producing 693 total datasets. Eight surrogate modeling techniques were used for comparison: multivariate adaptive regression splines (MARS);(Friedman, 1991), random forests (RF);(Breiman, 2001) single hidden layer feed-forward artificial neural networks (ANN);(Haykin, 2009), extreme learning machines (ELM);(Haykin, 2009), Gaussian process regression (GP);(Rasmussen & Williams, 2005), support vector machines (SVM);(Drucker et al., 2002), Automated Learning of Algebraic Models using Optimization (ALAMO);(Cozad et al., 2014) and radial basis function networks (RBFN);(Gomm & Yu, 2000). Surrogate models were trained using the input-output pairs with each of the surrogate modeling techniques for the test functions. This process yielded 16,632 trained models. When necessary, the hyperparameters of each surrogate modeling technique (such as the number of hidden neurons for the neural network-based models and the number of trees in RF models) were optimized before training the models using ten-fold cross-validation. After the surrogate models were trained, the adjusted-R^2 values were calculated for each modeling technique-dataset pair.

2.3 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, including the first four statistical moments of the determinants of the estimated Hessian matrices of the datasets, and as the number of data points in the dataset, to use as potential inputs for predicting the model performance with the recommendation tool, resulting in a total of 40 attributes. The attributes aim to capture the overall behavior of the underlying model that generated the dataset. They were calculated for the datasets generated from the 99 test functions and used to construct the surrogate model recommendation tool.

A RF model was trained for each surrogate modeling technique to predict its adjusted-R^2 value using the identified attributes as inputs. Random forests are decision tree-based machine learning models, where the final output of the model is the average of the value predicted by every decision tree in the forest. Feature reduction was performed to determine which attributes had the most influence on the predicted output value for each modeling technique. Feature reduction techniques included linear and rank correlations (Zou et al., 2003) between the adjusted-R^2 value and the attributes, and the built-in feature selection method in RF models. In RFs, features are selected based on how well they improve the data separation at each decision node in each decision tree in the RF (Brieman, 2001). For each dataset, based on the adjusted-R^2 values, each of the surrogate modeling techniques was classified as either being recommended or not recommended for both the predicted and actual metric values. These classifications were compared and used to evaluate the quality of the selection recommendations.

3. Performance Metrics

The adjusted-R^2 value is used to assess the surrogate models' performance for surface approximation. The formula for calculating adjusted-R^2 ($\hat{R}^2$) is shown in Eq. (1).

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

In Eq. (1), R^2 is the R-squared regression coefficient, n is the number of data points in the training set, and k is the number of model parameters (or hyperparameters). The adjusted-R^2 takes into account both the surrogate model accuracy and complexity (Miles, 2005). Taking complexity into account is essential in ensuring that the model is not overfit as overfit models do not generalize well to new data. $\hat{R}^2$ values typically fall between zero and one, with an $\hat{R}^2$ of one indicating a perfect fit. However, with the adjustment for model size, adjusted-R^2 values can become negative.

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 (Sokolova & Lapalme, 2009), and the hit ratio (Cui et al., 2016). The accuracy is the percentage of recommendations that are correct. Precision is the probability that a model classified as recommended should actually be recommended. The hit ratio is the percentage of the time the model with the highest calculated adjusted-R^2 is included in the set of recommended models. All three performance metrics range from 0 to 100%. Monte Carlo cross-validation was used to evaluate the performance of the

recommendation tool with 100 Monte Carlo trials. Each trial had a test set size of 75, which was about 11% of the total simulated data.

4. Results and Discussion

4.1 Surrogate Model Performance

The surrogate modeling technique that yielded the model with the highest adjusted-R^2 value was selected as the "best" one. For each shape category, the number of times a technique was selected as best was tabulated. These tabulated values were divided by the total number of datasets in the category to calculate the fraction of datasets for which each surrogate modeling technique was selected as the best performing (Fig. 1). There was no significant difference in the adjusted-R^2 values among the three sampling methods. Therefore, only results for Sobol sequence sampling are shown here. For valley, bowl, and other-shaped functions, GP models provide the highest adjusted-R^2. However, ALAMO and MARS models produce the highest adjusted-R^2 most frequently for bowl and multi-local minima-shaped functions, respectively. These results indicate that the underlying function shape has an effect on which surrogate modeling technique may be most appropriate for approximating a dataset. While in general, GP models may provide the most accurate approximation, specific shape characteristics may lead to another technique's being more appropriate. It should be noted that although these results only reflect a single technique being selected with the highest adjusted-R^2, in many cases, there were multiple techniques with values that were not significantly different than that of the highest adjusted R^2 value.

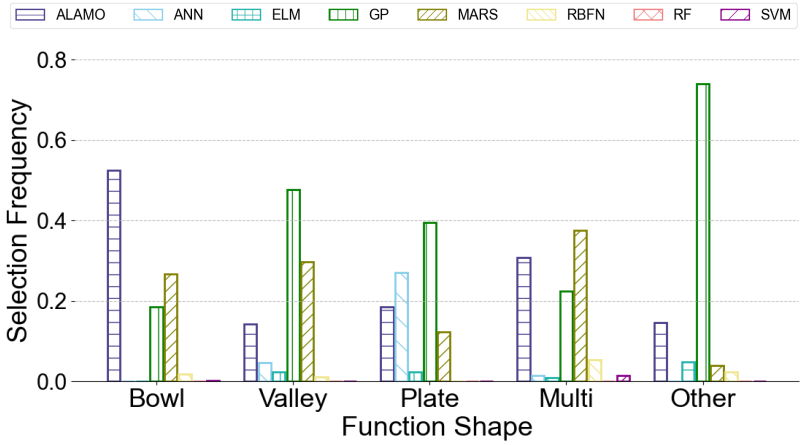

Figure 1- Percentage of datasets grouped by function shape for which each surrogate modeling technique had the highest adjusted-R^2

4.2 Attribute Selection for Adjusted-R^2 Prediction

From the comparison results, we can conclude that there is a relationship between the underlying shape of the surface being modeled and the performance of each of the surrogate modeling techniques. The minimum Mahalanobis distance (De Maesschalck et al., 2000) between any two points in the simulated dataset was moderately correlated to the calculated adjusted-R^2 of RF models, with a linear correlation coefficient of -0.58 and a rank correlation coefficient of -0.71. The position of the data points in the dataset

and how close they are to each other may be correlated to the approximation of RF models due to the need for the models to partition the design space of the surface when determining the decision nodes in each tree in the RF model. Data points that are closer together lead to smaller partitions and more accurate predictions.

For the feature selection by RF models, each technique had a different set of selected attributes for prediction. For ALAMO, ANN, RBFN, and SVM models, 18 different attributes were selected as important. Random forest models selected 19, 11, 20, and 17 attributes as important for ELM, GP, MARS and RF models, respectively. The attribute most commonly selected as being important for predicting the adjusted-R^2 was the minimum Mahalanobis distance between training points. Other commonly selected features include those related to the distributions of output values, specifically the relative size of the output distribution tails and the output distribution skewness, and the ratios of the average estimated gradient to the minimum and maximum estimated gradients for all of the data points in the dataset. These results suggest that the distribution and location of the sample points and the relative steepness and smoothness of the surface have a high level of influence on how well each of the surrogate models is able to approximate that surface.

For all of the neural network-based models (ANN, ELM, and RBFN) and RF models, the attribute selected with the highest importance was the percentage of the simulated data points that were located in the upper tail of the output distribution. The closely related attributes of the ratio of the upper and lower tail sizes and the skewness of the output value distribution were selected as most important for GP and MARS models, respectively. These attributes may have an effect on the accuracies of all these techniques as having data unevenly concentrated (or sparse) at the extreme values may skew models to predict more accurately in areas of data concentration and less so for other areas of the design space. For example, in the case of RF models, uneven tails could cause decision nodes in the model trees to split more frequently at the extremes of the output values while more finely split partitions are really needed elsewhere, such as where the gradients are steeper. For the neural network-based models, the on-off nature of the hidden layer nodes may make them more suitable for making accurate predictions for surfaces where large areas of the design space have similar output values, creating flat or nearly flat areas. The coefficient of variation (COV) was selected as the most important feature for the prediction of the performance of SVM models. The COV is inversely related to the signal-to-noise ratio of a surface (Wang et al., 2013). This attribute may be important for SVM model performance as the support vectors fitted in the model construction can easily become sensitive to noise as they are only dependent on a small set of the data used to train the model (Sabzekar et al., 2011). For ALAMO models, all of the selected attributes had roughly equal amounts of importance.

4.3 Recommendation Tool Performance

The selected attributes were used as inputs to train a RF model for the eight techniques to predict the adjusted-R^2 for a given dataset. Based on the predicted adjusted-R^2 value, the recommendation tool then classifies each of the surrogate models as being recommended or not for that dataset. This recommendation scheme allows for multiple similarly performing surrogate modeling techniques to be suggested for use in surface approximation. The selection tool identified which techniques should be recommended for the simulated datasets with an accuracy of 87%. The precision, or the probability that a recommended technique should actually be recommended, was 86%. The hit

ratio, the percentage of time techniques that had the highest adjusted-R^2 for a dataset were included in its set of recommended models, was 80%.

5. Conclusions

Selecting an appropriate surrogate modeling technique depends on the characteristics of the dataset being modeled. We identified attributes of datasets that are appropriate for use in predicting the adjusted-R^2 value. Using these attributes, we have constructed a tool that can recommend surrogate modeling techniques for approximating a dataset with 87% accuracy and 86% precision. Future work on the tool will include expanding it to surrogate-based optimization recommendations and investigation of additional attributes and machine learning techniques to improve recommendation quality.

References

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Cozad, A., Sahinidis, N. V., & Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. Aiche Journal, 60, 2211-2227.

Cui, C., Hu, M. Q., Weir, J. D., & Wu, T. (2016). A recommendation system for meta-modeling: A meta-learning based approach. Expert Systems with Applications, 46, 33-44.

Davis, S., Cremaschi, S., & Eden, M. (2017). Efficient Surrogate Model Development: Optimum Model Form Based on Input Function Characteristics. In A. Espuna, M. Graells & L. Puigjaner (Eds.), 27th European Symposium on Computer Aided Process Engineering (ESCAPE 27) (Vol. 40, pp. 457-462). Barcelona, Spain: Elsevier.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems, 50, 1-18.

Drucker, H., Shahrary, B., & Gibbon, D. C. (2002). Support vector machines: relevance feedback and information retrieval. Information Processing & Management, 38, 305-323.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines - Rejoinder. Annals of Statistics, 19, 123-141.

Garud, S. S., Karimi, I. A., & Kraft, M. (2018). LEAPS2: Learning based Evolutionary Assistive Paradigm for Surrogate Selection. Computers & Chemical Engineering, 119, 352-370.

Gomm, J. B., & Yu, D. L. (2000). Selecting radial basis function network centers with recursive orthogonal least squares training. Ieee Transactions on Neural Networks, 11, 306-314.

Haykin, S. (2009). Neural Networks and Learning Machines (3rd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Miles, J. (2005). R Squared, Adjusted R Squared. In Encyclopedia of Statistics in Behavioral Science: John Wiley & Sons Ltd.

Rasmussen, C. E., & Williams, C. K. I. (2005). Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, 1-247.

Sabzekar, M., Yazdi, H. S., & Naghibzadeh, M. (2011). Relaxed constraints support vector machines for noisy data. Neural Computing & Applications, 20, 671-685.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45, 427-437.

Surjanovic, S., & Bingham, D. (2013). Virtual Library of Simulation Experiments. In (Vol. 2018). Simon Fraser University.

Wang, B., Goodpaster, A. M., & Kennedy, M. A. (2013). Coefficient of variation, signal-to-noise ratio, and effects of normalization in validation of biomarkers from NMR-based metabonomics studies. Chemometrics and Intelligent Laboratory Systems, 128, 9-16.

Williams, B., Lobel, W., Finklea, F., Halloin, C., Ritzenhoff, K., Manstein, F., Mohammadi, S., Hashemi, M., Zweigerdt, R., Lipke, E., & Cremaschi, S. (2020). Prediction of Human Induced Pluripotent Stem Cell Cardiac Differentiation Outcome by Multifactorial Process Modeling. Front Bioeng Biotechnol, 8, 851.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. Radiology, 227, 617-622.