Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

CONSTRUCTING CLASSIFIERS OF EXPENSIVE SIMULATION-BASED DATA BY SEQUENTIAL EXPERIMENTAL DESIGN

Joachim van der Herten
Ivo Couckuyt
Dirk Deschrijver
Tom Dhaene

Internet Based Communication Networks and Services (IBCN)
Ghent University - iMinds
Gaston Crommenlaan 8 (Bus 201)
B-9050 Gent, Belgium

ABSTRACT

Sequential experimental design for computer experiments is frequently used to construct surrogate regression models of complex blackbox simulators when evaluations are expensive. The same methodology can be used to train classifiers of labeled data which is expensive to obtain. For certain problems classification can be a more appropriate method to obtain a solution with fewer samples.

1 INTRODUCTION

Most supervised learning research on classification focusses on large data sets. The goal is to obtain a separation of all data in two or more distinct groups based on characteristics of the training data (*features* or *variables*) as input and the corresponding group as output (*class label*), without prior knowledge about the relation. After training is completed the classifier can predict the label of unseen data points. Amongst other reasons, the accuracy of these predictions depend on the correlation between the features and the output label.

However, acquisition of the output labels can be *expensive*: they can be the result of costly or possibly dangerous real-life experiments, or computationally intensive evaluation of high-fidelity simulators (computer experiments). Acquiring the labels of a big dataset and train a classifier is infeasible in these situations as the computational costs will be high. Constructing a regression model of a computationally expensive deterministic simulator of a complex blackbox system is a problem known as surrogate modeling, metamodeling or response surface modeling (Goethals et al. 2012). This methodology is often applied during product design to avoid building many prototypes (virtual prototyping or CAD). The goal is to obtain an accurate regression model using a minimum amount of (often high-fidelity) evaluations of the simulator. The surrogate model is then used to answer design questions involving many evaluations (such as exploration, optimization or sensitivity analysis). In the context of surrogate modeling the simulator is usually considered to be deterministic.

The use of Sequential Experimental Design as used in surrogate modeling (Gorissen et al. 2010) can be used successfully to construct accurate classifiers when labels are to be obtained from expensive simulations, and the use of classifiers instead of regression based models has the potential to solve some typical surrogate modeling problems more efficiently.

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

2 SEQUENTIAL DESIGN

Traditionally, surrogate (regression) models were constructed by a *one-shot* design. Because traditional Design of Experiments (DoE) concepts such as blocking and replication loose their relevance in the context of computer experiments (as they are deterministic). These designs focus on space-fillingness, such as the popular (maximin) Latin Hypercube. This approach has two major disadvantages: first of all it is not known a priori how much samples are needed to obtain the desired level of accuracy. Secondly, the sampling distribution of a one-shot space-filling designs is usually suboptimal, as regions with non-linear response behaviour typically require a more dense sampling distribution, compared to less dynamic regions.

In sequential design, the starting point is a small space-filling design to seed the process. Each iteration, intermediate models are constructed to track the progress of the accuracy, new locations for additional data points are chosen intelligently based on all acquired knowledge (space-fillingness, responses, model errors) and the responses are calculated by evaluation of the simulator. This process continues until the predefined modeling goals have been met.

The methodology of sequential design can also be adopted for the construction of classifiers when computing the labels for data points is expensive. Instead of intermediate (regression) models, classifiers are constructed and scored according to appropriate performance metrics. New data points are chosen (for example near the discovered class boundaries to improve the accuracy of the boundary) until the classifier has obtained sufficient accuracy.

This approach offers advantages for several applications, such as identification of *feasible regions* (Couckuyt et al. 2013) as often required in product design, or surrogate modeling of high-dimensional problems by identifying key regions and only constructing a regression for these regions during a second phase: this approach can also be used to optimize high-dimensional objective functions (Singh et al. 2014). Another application are constraint satisfaction problems such as encountered in Surrogate-Based Optimization (SBO) with black-box constraints.

3 CONCLUSION

Applications indicate the use of classification models can greatly enhance answering problem statements in product design involving feasible regions and optimization with black-box constraints. These problems do not require all knowledge on the output variability and can therefore focus on the boundary of interest. It can also assist in the optimization process of high-dimensional objective functions. Our suggested approach greatly enhances the capabilities of engineers during the virtual prototyping phase of product design.

ACKNOWLEDGMENTS

This research has been funded by the Interuniversity Attraction Poles Programme BESTCOM initiated by the Belgian Science Policy Office. Ivo Couckuyt is a post-doctoral research fellow of FWO-Vlaanderen.

REFERENCES

Couckuyt, I., J. Aernouts, D. Deschrijver, F. D. Turck, and T. Dhaene. 2013. “Identification of quasi-optimal regions in the design space using surrogate modeling.”. *Eng. Comput. (Lond.)* 29 (2): 127–138.

Goethals, K., I. Couckuyt, T. Dhaene, and A. Janssens. 2012. “Sensitivity of night cooling performance to room/system design : surrogate models based on CFD”. *Building and Environment* 58:23–36.

Gorissen, D., K. Crombecq, I. Couckuyt, P. Demeester, and T. Dhaene. 2010. “A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design”. *Journal of Machine Learning Research* 11:2051–2055. Available at http://sumo.intec.ugent.be.

Singh, P., F. Ferranti, D. Deschrijver, I. Couckuyt, and T. Dhaene. 2014. “Classification Aided Domain Reduction for High Dimensional Optimization”. In *Proceedings of the 2014 Winter Simulation Conference*, WSC ’14, 3928–3939. Piscataway, NJ, USA: IEEE Press.