Automatic Regression Modeling with Active Learning

Dirk Gorissen DIRK.GORISSEN@UGENT.BE
Tom Dhaene TOM.DHAENE@UGENT.BE
Eric Laermans ERIC.LAERMANS@UGENT.BE
Ghent University - IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8, Bus 201, 9050 Ghent, Belgium

Abstract

Many complex, real world phenomena are difficult to study directly using controlled experiments. Instead, the use of computer simulations has become commonplace as a feasible alternative. However, due to the computational cost of these high fidelity simulations, the use of neural networks, kernel methods, and other surrogate modeling techniques have become indispensable. Surrogate models are compact and cheap to evaluate, and have proven very useful for tasks such as optimization, design space exploration, visualization, prototyping, and sensitivity analysis. Consequently, there is great interest in techniques that facilitate the construction of such regression models, while minimizing the computational cost and maximizing model accuracy. We present a fully automated machine learning toolkit for regression modeling and active learning to tackle these issues. We place a strong focus on adaptivity, self-tuning and robustness in order to maximize efficiency and make our algorithms and tools easily accessible to other scientists in computational science and engineering.

1. Introduction

For many problems from science and engineering it is impractical to perform experiments on the physical world directly (e.g., airfoil design, earthquake propagation). Instead, complex, physics-based simulation codes are used to run experiments on computer hardware. While allowing scientists more flexibility to study phenomena under controlled conditions, computer experiments require a substantial investment of computation time (one simulation may take many minutes, hours, days or even weeks) (Wang & Shan, 2007). As a result, the use of various approximation methods that mimic the behavior of the simulation model as closely as possible has become standard practice. This work concentrates on the use of data-driven, **global** approximations using compact surrogate models (also known as metamodels, or response surface models (RSM)). Popular metamodel types include: neural networks, Kriging models, and Support Vector Machines (SVM).

Global surrogate models provide a fast and efficient way for the engineer to explore the relationship between parameters (design space exploration), study the influence of various boundary conditions on different optimization runs, or enable the simulation of large scale systems where this would normally be too cumbersome. For the last case a classic example is the full-wave simulation of an electronic circuit board. Electromagnetic modeling of the whole board in one run is almost intractable. Instead the board is modeled as a collection of small, compact, accurate replacement surrogate models that represent the different functional components (capacitors, resistors, ...) on the board.

2. Motivation

However, in order to come to an acceptable approximation, numerous problems and design choices need to be overcome: what data collection strategy to use (active learning), what model type is most applicable (model selection), how should model parameters be tuned (hyperparameter optimization), how to optimize the accuracy vs. computational cost trade-off, etc. Particularly important is the data collection strategy. Since data is computationally expensive to obtain, it is impossible to use traditional, one-shot, space filling experimental designs. Data points must be selected iteratively, there where the information gain will be the greatest. An intelligent sampling function is needed that minimizes the number of sample points selected in each iteration, yet maximizes the informa-

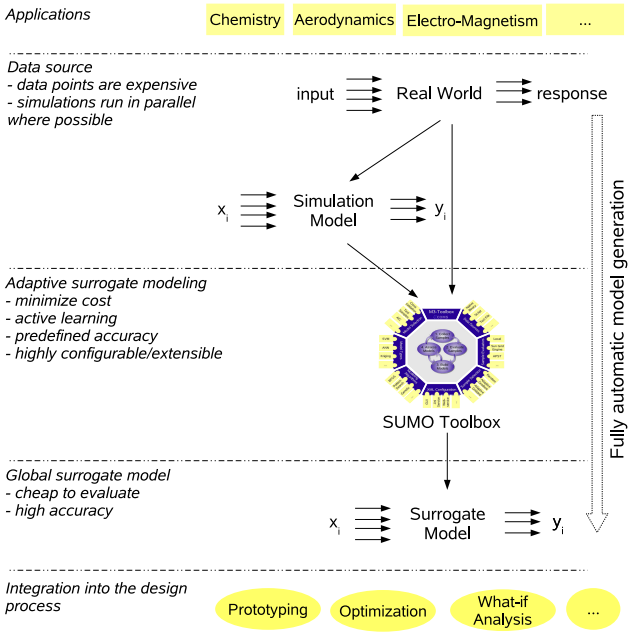

Figure 1. Automatic Adaptive Surrogate Modeling

tion gain of each iteration step. This is the process of active learning (Sugiyama & Ogawa, 2002), but it is also known as adaptive sampling or sequential design.

Together this makes that there are an overwhelming number of options available to the designer: different model types, different experimental designs, different model selection criteria, etc. However, in practice it turns out that the designer rarely tries out more than one subset of options. All too often, surrogate model construction is done in a one-shot manner. Iterative and adaptive methods, on the other hand, have the potential of producing a much more accurate surrogate at a considerably lower cost (less data points) (Busby et al., 2007). We present a state-of-the-art machine learning platform that provides an automatic, flexible and rigorous means to tackle such problems and that can easily be integrated in the engineering design process: the **SU**rrogate **MO**deling (SUMO) Toolbox.

3. SUMO Toolbox

The SUMO Toolbox (Gorissen et al., 2006) is an adaptive tool that integrates different modeling approaches and implements a fully automated, adaptive global surrogate model construction algorithm. Given a simulation engine, the toolbox automatically generates a surrogate model within the predefined accuracy and time limits set by the user (see figure 1). However, at the same time keeping in mind that there is no such thing as a 'one-size-fits-all', different problems need to be modeled differently. Therefore the toolbox was designed to be modular and extensible but not be too cumbersome to use or configure. Different plugins are supported: model types (neural networks, SVMs, splines, ...), hyperparameter optimization algorithms (Pattern Search, Genetic Algorithm (GA), Particle Swarm Optimization (PSO), ...), active learning (density based, error based, gradient based, ...), and sample evaluation methods (local, on a cluster or grid). The behavior of each component is configurable through a central XML configuration file and components can easily be added, removed or replaced by custom, problem-specific, implementations.

The difference with existing machine learning toolkits such as Rapidminer (formerly Yale), Spider, Shogun, Weka, and Plearn is that they are heavily biased towards classification and data mining (vs. regression), they assume data is freely available and cheap (no active learning), and they lack advanced algorithms for the automatic selection of the model type and model complexity.

Our approach has been successfully applied to a very wide range of fields ranging from combustion modeling in chemistry and metallurgy, semi-conductor modeling (electromagnetism), aerodynamic modeling (aerospace), to structural mechanics modeling in the car industry. Its success is primarily due to its flexibility, self tuning implementation, and its ease of integration into the larger computational science and engineering pipeline.

References

Busby, D., Farmer, C. L., & Iske, A. (2007). Hierarchical nonlinear approximation for experimental design and statistical data fitting. *SIAM Journal on Scientific Computing*, *29*, 49–69.

Gorissen, D., Hendrickx, W., Crombecq, K., & Dhaene, T. (2006). Integrating gridcomputing and metamodelling. *Proceedings of 6th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2006)* (pp. 185–192). Singapore.

Sugiyama, M., & Ogawa, H. (2002). Release from active learning/model selection dilemma: optimizing sample points and models at the same time. *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on* (pp. 2917–2922).

Wang, G. G., & Shan, S. (2007). Review of metamodeling techniques in support of engineering design optimization. *Journal of Mechanical Design*, *129*, 370–380.