Determining the Feasible Region of Non-Linear Dynamic Process Models for Optimization Through Data-Driven Regression with Classification

Torben Talis,a Gerardo Brand Rihm,a Raoul Heese,b Erik Esche,a Michael Bortz,b Jens-Uwe Repkea

aTechnische Universität Berlin, Process Dynamics and Operation Group, Str. des 17. Juni 135, Berlin 10623, Germany

bFraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1, Kaiserslautern 67663, Germany

t.talis@tu-berlin.de

Abstract

Surrogate models from machine learning have no guarantees regarding extrapolation. However, they will also predict results in areas where the original model is known to be infeasible. Classifiers can be used to model the boundary of the feasible region. Here, classifiers are analysed and extended towards dynamic systems. A rigorous, dynamic model is sampled in different ways and the feasible region is identified by data-driven classifiers. The suitability of different architectures is evaluated for the case study of a flash separation unit. The most suitable classifier will be used to improve real time optimization using surrogate models by preventing faulty extrapolations and forcing all points of the solution to be in the feasible region.

**Keywords**: imbalanced classification, data-driven models, feasible region, dynamic systems.

* 1. Introduction

Chemical processes can be described by complex nonlinear dynamic models. These models have a limited feasible region, e.g., a common flash unit model is only valid in two-phase region or a distillation column model is only defined in the proper fluid dynamic operation area. Hard and hidden constraints like these are included in the formulation of first principles models. However, the computational cost of these models often prohibits their application in real time optimization. In contrast, data-driven surrogate models are fast to evaluate, but there is typically no information on the region of validity included in the model. Hence, these models return results even for points outside the feasible region, which are obviously bad extrapolations. Real plant operation is frequently performed close to these bounds, so information on the feasible region should be added to the surrogate models. Similar has been done before for steady state models by e.g., (Penteado et al., 2020). (Schweidtmann et al., 2022) use a classifier to identify the boundaries of a training dataset und use this as the region of validity for their surrogate model. The actual feasible region of the process is not investigated. However, the method is applied for optimization on a dynamic model. We focus on exploring the feasible region of the rigorous model and matching the surrogate model’s region of validity to it. This enables safe operation close to the bounds. For this, it is mandatory to incorporate data of non-converging trajectories into the training dataset.

* 1. Methods

In this study we use a rigorous model of a flash unit with non-ideal thermodynamics, depicted in Fig. 1, to first generate data and then investigate the suitability of the following classifier architectures: random forest, k-nearest neighbors, and balanced bagging classifier (Guillaume Lemaître et al., 2017). The rigorous model is only valid in the two-phase region and is described in more detail in (Brandner et al., 2023). Different cases are investigated, where scenarios for heat flux and feed composition are generated by a step or a pseudorandomized sequence. These signals have a wide spread, so the simulation diverges at the end of almost all simulation scenarios, whenever the system leaves the two-phase region. However, this results in highly imbalanced datasets with a lot of valid points (along the time series) and only very few points describing the boundary of convergence and beyond (typically the last point of such a time series). 28 datasets with 3 different sampling rates (10 / 100 / 1000 s-1) are investigated. The share of non-converging points amounts to 0.8, 0.1, and 0.01 %, respectively for the three sampling rates. For the classification, four different sets of features are selected: (i) the states *x (T, p, xB, yD)*, (ii) *x* and the controls *u (Q, xF)*, (iii) *x* and gradient information *dx*, (iv) *x*, *u*, and *dx*. For ease of implementation, the gradient of the systems internal energy and the component hold-ups were used instead of the before mentioned states. Although, it might be less intuitive, the system is fully described by this information. All datasets are then split into 80% training and 20% test data, while keeping the same ratios of the two classes in both sets and are used to train different classifiers. Tab. 1 shows the truth table for the models. The geometric mean score (g-score), which is the geometric mean of sensitivity and specificity (Eq. (1)), was used to score their performance.

Figure 1: Flowsheet of the flash unit with controls *Q* and *xF*



|  |  |
| --- | --- |
| $$g= \sqrt{\frac{TP}{TP+FN}∙\frac{TN}{TN+FP}}$$ | (1) |

Table 1: Truth table for classifier

|  |  |  |
| --- | --- | --- |
|  | Classifier predictsnon-convergence | Classifier predicts convergence |
| Rigorous modeldoes not converge | TP | FN |
| Rigorous modelconverges | FP | TN |

* 1. Results and Discussions

There are eight datasets with a timestep of 0.001, and each 10 with a timestep of 0.01 and 0.1. The mean, as well as the min and max value of the g-score of all datasets are shown in Fig. 2 for four different architectures (3-nearest-neighbors (3-NNC), 5-nearest-neighbors (5-NNC), random-forest (RFC), and balanced-bagging classifier (BBC)) against the four different feature sets. The results show that in general a higher sampling rate renders the classification problem more difficult and thus leads to a worse classification accuracy. However, while the classification metric is better for small sampling rates, the actual bounds of the feasible region cannot be described accurately in these cases. In contrast to the classic k-NN- and RFC classifiers, the BBC is an ensemble method, that makes use of the undersampling technique and was specifically developed for imbalanced datasets. The achievable g-scores are much better than with the other architectures and the metric seems to be surprisingly independent of both the timestep and the selected features.



Figure 2: Comparison of different classifier architectures. Mean, min, and max value of all datasets for different feature sets.

* 1. Conclusions

In a case study, we evaluated the suitability of different data-driven classifiers for describing the feasible region of a dynamic model of a flash separation unit. Different sampling rates and classification techniques have been analyzed and compared. Of the investigated model architectures, the balanced bagging classifier is best in predicting, whether the rigorous model is solvable at a given state. To the best of our knowledge this is the first instance of a classifier used for describing the feasible region, instead of the valid range of training data. We consider the present work a proof of concept for dynamic models in general. This classifier will subsequently be used in combination with a data-driven regression model, to guarantee feasibility of the solutions of optimization problems.

* 1. Acknowledgement

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466380688 – within the Priority Programme “SPP 2331: Machine Learning in Chemical Engineering. Knowledge Meets Data: Interpretability, Extrapolation, Reliability, Trust”.

**References**

Brandner D., Talis T., Esche E., Repke J.-U., Lucia S., 2023, Reinforcement learning combined with model predictive control to optimally operate a flash separation unit In: M.C.G.E.P. Antonios C. Kokossis (Ed), 33 European Symposium on Computer Aided Process Engineering, Vol 52, Elsevier, 595–600.

Guillaume Lemaître, Fernando Nogueira, Christos K. Aridas, 2017, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, 18, 1–5.

Penteado A.T., Esche E., Weigert J., Repke J.-U., 2020, A Framework for Stochastic and Surrogate-Assisted Optimization using Sequential Modular Process Simulators In: 30th European Symposium on Computer Aided Process Engineering, Vol 48, Elsevier, 1903–1908.

Schweidtmann A.M., Weber J.M., Wende C., Netze L., Mitsos A., 2022, Obey validity limits of data-driven models through topological data analysis and one-class classification, Optimization and Engineering, 23, 855–876.