Process-based Modeling and Design of Dynamical Systems

Jovan Tanevski[1], Nikola Simidjievski[1], Ljupčo Todorovski[2,1], and Sašo Džeroski[1]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] University of Ljubljana, Slovenia
{name.surname}@ijs.si; ljupco.todorovski@fu.uni-lj.si

Abstract. Process-based modeling is an approach to constructing explanatory models of dynamical systems from knowledge and data. The knowledge encodes information about potential processes that explain the relationships between the observed system entities. The resulting process-based models provide both an explanatory overview of the system components and closed-form equations that allow for simulating the system behavior. In this paper, we present three recent improvements of the process-based approach: (i) improving predictive performance of process-based models using ensembles, (ii) extending the scope of process-based models towards handling uncertainty and (iii) addressing the task of automated process-based design.

1 Introduction

Process-based modeling (PBM) supports knowledge discovery by learning understandable and communicable models of dynamical systems. PBM uses domain-specific knowledge as declarative bias in combination with observed time-series data to address the task of modeling real-world systems. It performs both structure identification and parameter estimation, resulting in a process-based model which specifies a set of differential equations. In turn, such models accurately capture the complex and nonlinear behavior of a dynamical system through time.

Learning models of dynamical systems is a supervised machine learning task: the predictive variables correspond to observed system variables, while the targets correspond to their time derivatives. However, the task bears two specific properties that limit the use of traditional machine learning approaches. First, the resulting models take the form of a set of entities, processes and differential equations, i.e., artifacts used by scientists and engineers to construct explanatory models. On the other hand, machine learning methods operate on classes of predictive models that generalize well over arbitrary data, while keeping the complexity of training and evaluation procedures low. Second, the observed variables are measured at consecutive time points, so the data instances breach the common assumption of their mutual independence.

The PBM approach relies on the paradigm of computational scientific discovery [4] and more specifically, on approaches to inductive process modeling.

On one hand, research in this area has a long tradition and has been applied to a variety of domains [1, 2, 11, 3]. However, while successful, it has been at the margins of mainstream machine learning. On the other hand, the PBM approach has so far focused primarily on applications within a narrow class of problems that emphasize descriptive and deterministic models at output, given a single data type at input. In terms of output, such models are typically simulated and analyzed using the learning data. Therefore, they have a tendency to overfit – rendering them incapable at accurately predicting future system's behavior. Also, these models do not capture the intrinsic uncertainty of the interactions in the system. They always predict exactly the same behavior of the system at output in a deterministic manner: determined only by initial conditions and ignoring the uncertainty in real-world systems. In terms of input, an assumption of the PBM is that time-series of observations are always available and sufficient. This, however, does not hold for problems with limited observability or tasks, such as design, where different types of input are required.

In response, our recent developments of the PBM approach have aimed at bridging the gap between machine learning and domains of application within physical and life sciences. We address the limitations of the PBM approach by broadening the classes of tasks it can address. We build on the tradition of constant performance improvement, but also extend the scope of potential applications. In particular, to improve the performance on the task of predictive modeling, we support the learning of different types of ensembles of process-based models [5, 7, 6]. Next, we extended the output to include process-based models that describe stochastic interactions [8]. Finally, in order to address tasks of modeling dynamical systems under limited observability and tasks of design of dynamical systems, we consider different types of input data. Namely, in addition to time-series of observations of system variables we allow for the definition of expected properties of the behavior of the dynamical system [9, 10].

2 Methods

The PBM learning task takes domain-specific knowledge and time-series data at input (Figure 1). The resulting model comprises system variables represented as *entities* and their interactions that define the underlying model structure represented as *processes*. This representation allows for straightforward mapping of process-based models into a set of differential equations. The model parameters are fitted to the data using evolutionary optimization methods with the sum-of-squares loss function as the objective. The PBM approach, however, adds an extra layer to the model equations. In particular, the models are constructed using components from a library of domain-knowledge, represented by *template entities and processes*. These templates encode taxonomies of variable and constant properties of the constituents in the dynamical systems as well as the taxonomies of processes (interactions) among them. The (partial) instantiations of such templates, taken from arbitrary levels of the respective taxonomies, define and constrain the model structure search space for a specific modeling task.

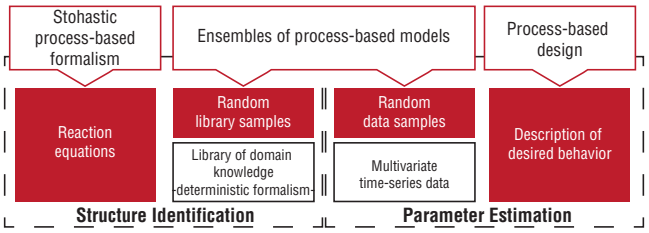

Fig. 1. General overview of the three extensions of PBM presented in this paper.

PBM has four distinguishing features. First, it produces *understandable* models, which give clear insight into the structure of a dynamical system building on the traditional mathematical description. The processes relate specific parts of the set of differential equations to understandable real world causal relations between the system's components. Second, process-based models retain the *utility* of traditional mathematical models. They can be readily simulated and analyzed using well established numerical approaches. Third, PBM is *generally applicable* to domains that require models described in terms of equations. Finally, the PBM approach is *modular*. The domain-knowledge library can be instantiated into a number of different modeling components specific to a particular modeling task. It captures the basic modeling principles in a given domain and can be reused for different modeling applications within the same domain.

We report on three extensions of PBM (Figure 1). To improve the capability to predict future system's behavior, we consider learning of ensembles of process-based models. The constituent base process-based models are learned either from different samples of the measured data [5], random samples of the library of domain knowledge [7] or both [6]. Such sampling approaches have a direct effect on the generalization ability of the ensembles, leading to improved predictive performance. Second, the ensembles of process-based models can provide long-term predictions, relying only on the initial values of the state variables as opposed to traditional ML ensembles (in the context of time-series) that are typically used for short-term prediction.

To capture the intrinsic uncertainty of interactions within real world dynamical systems, we propose an improved finer grained formalism for representing domain knowledge [8]. It encodes the interactions between entities, i.e., processes in the form of reaction equations allowing for both deterministic and stochastic interpretation of process-based models and knowledge.

We extended the input to the PBM approach to different types of data, which allows handling a broader set of tasks ranging from completely data-driven to completely knowledge-driven modeling. In this context, we first strengthen the evaluation bias of modeling tasks with limited observability [10]. We use domain-specific criteria for model selection as part of a general regularized objective function for parameter optimization and model selection. Second, we formulate the novel task of process-based design of dynamical systems [9]. This approach does not take measured data at input, but is completely based on the description of desired properties of the behavior of a dynamical system. We further generalize the task by taking advantage of methods for simultaneous optimization of

multiple conflicting objectives (desired properties of the behavior). We use the complete information from the Pareto front of optimal solutions (obtained for every candidate design) to rank the designs and make a well informed selection.

3 Significance and Challenges

The methodology for learning ensembles of PBMs extends the scope of the traditional ensemble paradigm in machine learning towards modeling dynamical systems. It improves the generalization power of PBMs, providing more accurate simulation of the future behavior of the modeled systems. The proposed methodology employs four different methods for constructing ensembles of process-based models. Each of these significantly improves the predictive performance (on average up to 60% of relative improvement) over individual models on tasks of modeling population dynamics in three lake ecosystems [5, 7, 6].

The extension of the PBM approach towards stochastic process-based models has allowed us to model dynamical systems that are out of the scope of deterministic models. We have demonstrated that the stochastic PBM is capable of reconstructing known, manually constructed models from synthetic and real-world data in the domains of systems biology and epidemiology [8].

The capability of PBM to handle different inputs and multiple modeling objectives has led to important contributions in the domains of systems and synthetic biology. In particular, PBM can address the problem of high structural uncertainty (many candidate model structures) and incomplete data (i.e., limited observability of the system variables). In system biology, our approach can alleviate the model selection problem by strengthening the evaluation bias with introducing domain-specific model selection criteria [10]. In synthetic biology, we can now use PBM to solve the task of automated design. Our results show that PBM is capable of reconstructing known/good designs, as well as proposing novel alternative designs of a synthetic stochastic switch and a synthetic oscillator [9].

Note, finally, that all three extensions of the PBM approach are designed and implemented as independent modular components. Therefore, they are interoperable. They can be, in principle, arbitrarily combined and applied to novel tasks, such as learning ensembles of stochastic process-based models.

Several challenges, that we are aware of and currently working on, remain in PBM. The exhaustive combinatorial search currently in use is computationally inefficient and does not scale well with the number of candidate model structures. It is therefore necessary to integrate methods for heuristic search in our current implementation. An alternative approach to reducing search complexity is to use higher-level constraints on model structures that are more expressive than the current constraints. They can be based on the topological properties of the candidate model structures, or can define a probability distribution over the model structures. Finally, both process-based modeling and design require further evaluation on other related domains, such as neurobiology, systems pharmacology

and systems medicine, or on completely new domains. The new applications will most certainly open up new directions for improvement of the PBM approach.

Acknowledgements. The authors acknowledge the financial support of the Slovenian Research Agency (research core funding No. P2-0103, No. P5-0093 and project No. N2-0056 Machine Learning for Systems Sciences) and the Ministry of Education, Science and Sport of Slovenia (agreement No. C3330-17-529021).

References

1. Bridewell, W., Langley, P., Todorovski, L., Džeroski, S.: Inductive Process Modelling. Machine Learning 71, 109–130 (2008)
2. Džeroski, S., Todorovski, L. (eds.): Computational Discovery of Scientific Knowledge. Springer (2007)
3. Čerepnalkoski, D., Taškova, K., Todorovski, L., Atanasova, N., Džeroski, S.: The influence of parameter fitting methods on model structure selection in automated modeling of aquatic ecosystems. Ecological Modelling 245, 136–165 (2012)
4. Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M.: Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press (1992)
5. Simidjievski, N., Todorovski, L., Džeroski, S.: Predicting long-term population dynamics with bagging and boosting of process-based models. Expert Systems with Applications 42(22), 8484–8496 (2015)
6. Simidjievski, N., Todorovski, L., Džeroski, S.: Learning ensembles of process-based models by bagging of random library samples. In: Proceedings of the Nineteenth International Conference on Discovery Science. pp. 245–260. Springer (2016)
7. Simidjievski, N., Todorovski, L., Džeroski, S.: Modeling Dynamic Systems with Efficient Ensembles of Process-Based Models. PLoS One 11(4), 1–27 (2016)
8. Tanevski, J., Todorovski, L., Džeroski, S.: Learning stochastic process-based models of dynamical systems from knowledge and data. BMC Systems Biology 10(1), 1–30 (2016)
9. Tanevski, J., Todorovski, L., Džeroski, S.: Process-based design of dynamical biological systems. Scientific Reports 6(1), 1–13 (2016)
10. Tanevski, J., Todorovski, L., Kalaidzidis, Y., Džeroski, S.: Domain-specific model selection for structural identification of the Rab5-Rab7 dynamics in endocytosis. BMC Systems Biology 9(1), 1–31 (2015)
11. Todorovski, L., Bridewell, W., Shiran, O., Langley, P.: Inducing hierarchical process models in dynamic domains. In: Proceedings of the Twentieth National Conference on Artificial Intelligence. pp. 892–897. AAAI Press (2005)