Ecological Modelling 306 (2015) 305–317

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Learning ensembles of population dynamics models and their application to modelling aquatic ecosystems

Nikola Simidjievski [a,b,*], Ljupčo Todorovski [c], Sašo Džeroski [a,b]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[c] Faculty of Administration, University of Ljubljana, Slovenia

ARTICLE INFO

Article history:
Available online 5 September 2014

Keywords:
Aquatic ecosystems
Population dynamics
Phytoplankton growth
Process-based modelling
Ensembles
Bagging

ABSTRACT

Ensemble methods are machine learning methods that construct a set of models and combine their outputs into a single prediction. The models within an ensemble can have different structure and parameters and make diverse predictions. Ensembles achieve high predictive performance, benefiting from the diversity of the individual models and outperforming them.

In this paper, we develop a novel method for learning ensembles of process-based models. We build upon existing approaches to learning process-based models of dynamic systems from observational data, which integrates the theoretical and empirical paradigms for modelling dynamic systems. In addition to observed data, process-based modelling takes into account domain-specific modelling knowledge.

We apply the newly developed method and evaluate its utility on a set of problems of modelling population dynamics in aquatic ecosystems. Data on three lake ecosystems are used, together with a library of process-based knowledge on modelling population dynamics. Based on the evaluation results, we identify the optimal settings of the method for learning ensembles of process-based models, i.e., the optimal number of ensemble constituents (25) as well as the optimal way to select (using a separate validation set) and combine them (using simple average). Furthermore, the evaluation results show that ensemble models have significantly better predictive performance than single models. Finally, the ensembles of process-based models accurately simulate the current and predict the future behaviour of the three aquatic ecosystems.

1. Introduction

Mathematical models are widely used to describe the structure and predict the behaviour of dynamic systems under various conditions. Constructing such a model is a process that uses both expert knowledge and measured data about the observed system. The main challenge is integrating these two into an understandable model within the laws of nature.

Two major paradigms for constructing models of dynamic systems exist: theoretical (knowledge-driven) and empirical (data-driven) modelling. Following the first paradigm, domain experts establish an appropriate structure of the model and calibrate its parameters in an automatic fashion using measured data. The second approach uses measured data to search for such a combination of model structure and parameter values that leads to simulated behaviour that fits the measurements well. In both approaches, the models are often formulated as ordinary differential equations (ODEs).

Within the area of computational scientific discovery (Langley et al., 1987), a sub-field of equation discovery has emerged that studies methods for learning the model structure and parameter values of dynamic systems from observations (Džeroski and Todorovski, 2003; Bridewell et al., 2008). The state-of-the-art approaches in this area, referred to as process-based modelling (Bridewell et al., 2008; Čerepnalkoski et al., 2012), integrate the theoretical and the empirical paradigm to modelling dynamic systems. A process-based model (PBM) provides an abstraction of the observed system at two levels: qualitative and quantitative.

At the qualitative level, a process-based model comprises entities and processes. Entities correspond to agents involved in the modelled system, whereas processes represent the relations and interactions between the entities. This results in an interpretable model of a system, explaining the structure of the observed system. On the other hand, at the quantitative level, the entities define a set

* Corresponding author at: Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. Tel.: +386 1 477 3635.
E-mail addresses: nikola.simidjievski@ijs.si (N. Simidjievski), ljupco.todorovski@fu.uni-lj.si (L. Todorovski), saso.dzeroski@ijs.si (S. Džeroski).

http://dx.doi.org/10.1016/j.ecolmodel.2014.08.019
0304-3800/

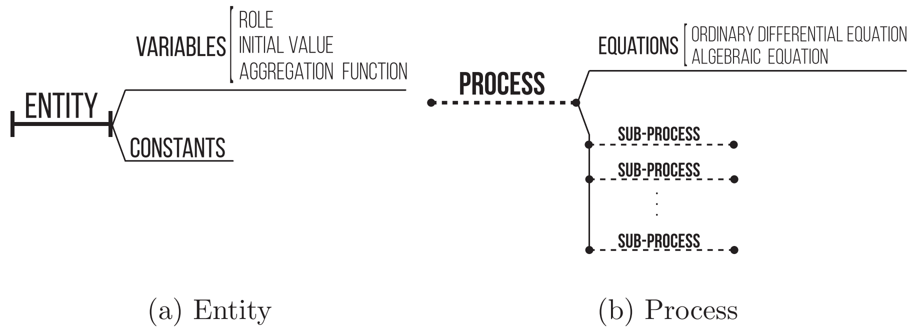

Fig. 1. The internal structure of entities and processes in process-based models.

of variables and constants, and the processes are annotated with equations modelling the underlying relations and interactions. At this level, we can transform a process-based model to a system of ODEs and simulate its behaviour.

Following the process-based modelling approach, we can generalize specific entities and processes into template entities and processes in a given modelling domain. A collection of such template entities and processes is called a library of process-based domain-specific knowledge. In modelling aquatic ecosystems, such a library of model components has been proposed by Atanasova et al. (2006b). The library defines a set of template entities, i.e., nutrients, primary producers, animals and environment, that typically occur in aquatic ecosystems (Luenberger, 1979). These entity templates are used to define template processes that provide recipes for modelling food-web interactions between the aquatic ecosystem entities. The knowledge encoded within the template entities and processes allow for automated modelling of population dynamics in aquatic ecosystems from measurements of system states (e.g., nutrients and species concentrations) through time. Process-based modelling software can then integrate the encoded knowledge with the measured system behaviour into a PBM of the observed system.

In our previous work, we have shown the utility of the process-based modelling approach for modelling population dynamics in a number of natural lakes (Čerepnalkoski et al., 2012) and marine ecosystems (Bridewell et al., 2008). Note however, that these studies focused on establishing descriptive, explanatory models of the population dynamics in aquatic ecosystems and the obtained models were analyzed and simulated on the same data that were used for learning them. In particular, they aimed to identify the limiting factors of the phytoplankton growth in the observed systems that are evident from the qualitative level of the learned process-based models. The generalization power of the obtained process-based models in terms of their ability to predict the future behaviour of the observed systems was not investigated in these studies.

In this study, we shift our focus towards the predictive performance of process-based models. The results of the preliminary experiments indicate the tendency of process-based models to overfit: While focusing on the provision of detailed and accurate descriptions of the observed systems, PBMs fail to accurately predict future system behaviour. To address this limitation of process-based models, we propose here a standard method for improving the predictive performance of models in machine learning, the use of ensembles. The idea of ensembles is to learn a set of predictive models (instead of a single one) and then combine their predictions. The prediction obtained with the ensemble is expected to be more accurate than the one obtained with a single model (Maclin and Opitz, 1999; Rokach, 2010).

The main contribution of this paper is a novel method for learning ensembles of process-based models. For tasks such as modelling the behaviour of ecosystems, the ensembles are usually employed in the context of learning tasks for classification and regression (Crisci et al., 2012; Knudby et al., 2010). However, to the best of our knowledge ensembles of process-based models have not yet been addressed in this context, and considered for tasks for modelling ecosystems.

We test the utility of the newly developed method for predictive modelling of population dynamics in lakes. To this end, we conjecture that ensembles of PBMs, similarly to other types of ensembles in machine learning, will improve the predictive performance of single models and lead to satisfactory prediction of future behaviour of the observed aquatic ecosystems. To test this hypothesis, we experiment on a series of tasks of modelling population dynamics in three lakes: Lake Bled, Lake Kasumigaura and Lake Zurich. From each lake we use seven yearly data sets, using six for learning and one for testing the predictive performance of the learned models. The aim of the experiments is two fold: Beside validating our central hypothesis (that ensembles perform better than single models), we also seek appropriate design choices related to our method for building ensembles of process-based models.

The remainder of this paper is organized as follows. Section 2 introduces the novel approach to learning ensembles of process-based models by discussing the task of automated modelling of dynamic systems – the process-based modelling approach, and focuses on a recent contribution to the area of automated process modelling, i.e., the ProBMoT tool. Section 3 describes ensemble methods in general and their adaptation for process-based modelling in particular. The design of the experiments, the evaluation measures and the data sets used are described in Section 4. Section 5 presents the results obtained the experimental evaluation. In Section 6, we discuss the contributions of this paper and overview the related work. Finally, Section 7 summarizes the work presented in this paper and discusses directions for further work.

2. Process-based modelling and ProBMoT

Equation discovery is the area of machine learning that aims at developing methods for learning quantitative laws, expressed in the form of equations, from collections of observed data. Recently, equation discovery methods have been used in the context of learning models of dynamic systems (Todorovski and Džeroski, 2007; Džeroski and Todorovski, 1993). The state-of-the-art equation discovery methods for modelling dynamic systems, referred to as process-based modelling (Bridewell et al., 2008; Džeroski and Todorovski, 2003) integrate domain-specific modelling knowledge and data into explanatory models of the observed systems. In the rest of this section, we briefly introduce the process-based modelling approach and then describe its particular implementation within the ProBMoT software platform.