Learning bagged models of dynamic systems

Nikola Simidjievski[1,2] **, Ljupco Todorovski**[3] **, Sašo Džeroski**[1,2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

[3] Faculty of Administration, University of Ljubljana, Slovenia

nikola.simidjievski@ijs.si

Abstract. In this paper, we present an ensemble learning method for modeling dynamic systems. The method is a combination of the bagging approach to ensemble learning and the approach of inductive process modeling of dynamic systems. We illustrate the use of the proposed method on the task of modeling phytoplankton growth in Lake Bled.

Keywords: dynamic systems, dynamic modeling, equation discovery ensembles, bagging, aquatic ecosystems

1 Introduction

Experts construct mathematical models to describe and predict the behavior of dynamic systems under various conditions. Such models are often formulated as ordinary differential equations (ODEs), which describe the change of the state of the system over time. Constructing a model is a cascade process using knowledge and measured data about the observed system and integrating these together with the laws of nature into an understandable pattern. Two major paradigms for constructing models of dynamic systems exist: (1) theoretical (knowledge-driven) and (2) empirical (data-driven) modeling. In the first approach, domain experts derive a mathematical (ODE) model structure of a system, the parameters of which are calibrated using the measured data. The second approach uses the measured data to search for the best combination of model structure and parameters which most adequately fits the task. This paradigm is used by machine learning approaches to modeling dynamic systems.

Within the area of computational scientific discovery, equation discovery systems [1] have emerged that use observation data to determine both the model structure and parameter values. The state-of-the-art equation discovery techniques [6] integrate both the theoretical and the empirical approaches to modeling dynamic systems. They use process-based domain knowledge, which describes the entities and processes that may occur in the particular system. The ability to use such knowledge has proven to be the key factor for the success of such systems.

Another machine learning paradigm that we draw upon is the paradigm of ensemble learning. In machine learning, ensemble learning methods [3] build combinations of multiple predictive models (ensembles) to maximize the predictive performance of the system. An ensemble learning method is a supervised learning technique which combines several base models and improves the overall predictive outcome representing a single model and minimizes the over-fitting of the data. Ensembles are employed mostly in the context of solving classification and regression tasks. In this paper, we focus on adapting them to the task of modeling dynamic systems. To the best of our knowledge ensembles of dynamic system models have not yet been considered in the machine learning community and addressed in this context.

We adapt the traditional approach to ensemble learning, i.e. bootstrap aggregation [4], to solve the task of learning ensembles of dynamic models from observation data. We also aim to evaluate the performance of this kind of ensembles of ODE models in comparison to "classical" individual ODE models. The individual models and the components of the ensembles derived with the process-based modeling automated approach. The remainder of this paper is organized as follows. Section 2 describes the bootstrap aggregation ensemble learning algorithm and its adaptation to the context of building ensemble models of dynamic systems. Section 3 gives an introduction to automated modeling of dynamic systems, focusing on inductive process modeling and recent contribution in this area, i.e., the ProBMoT system. Section 4 describes the experimental setup and discusses the results of the experimental evaluation. For the experimental evaluation, a case study from the domain of modeling aquatic ecosystems domain is considered, where the aim is modeling phytoplankton concentration in Lake Bled, Slovenia. Finally, Section 5 summarizes the conclusions of this paper.

2 Bagging of ODE models

Several practical and theoretical studies clearly state that an ensemble of models gives better performance than the best single model [2]. While ensembles of models for other machine learning tasks are widely used [3], learning ensembles of ODE models has not been considered. So far several approaches can be considered to address this challenge, using state-of–the-art techniques for learning different models from different data samples. For example, such approach considers using contiguous sections of the observed data with randomly selected starting points and duration. Adapting existing sampling methods requires more care and it is discussed below.

BAGGING (Bootstrap aggregation) [4] developed by Breiman et.al is one of the first and simplest ensemble learning techniques. This technique uses bootstrap sampling. Data instances are sampled uniformly with-replacements to obtain bootstrap replicates of the training data. Each base model is learned from one bootstrap replicate. The models are combined afterwards by averaging the output (regression) or by voting (classification). This method successfully overcomes the over-fitting problem, but is not useful for linear models or in general with base models that only changes a little for small changes in the training data. Bagging has no memory, and can be parallelized to handle different replicates on different CPUs, which performance-wise is very useful [11].

Unlike the traditional methods for data sampling, where part of the data is used for training the different models, which are combined into an ensemble afterwards; here we consider time point error weighting as a method for sampling. The classical approach of generating bootstrap samples for a building ensemble models for common classification task, includes generating samples from random data instances in the training set. By uniformly sampling with replacements, it is very likely that some of the instances will repeat, and there of generating m different training sets with size same as the original training set. Furthermore, every sample is used in the learning algorithm, thereof learning m different model classifiers which will participate in the ensemble.

On the other hand, our approach differs in the sampling process from the classical approach. Mainly, for performing bagging, we choose uniformly random time points

from the time series, and by assigning weights to each of them we are adapting the technique of generating bootstrap replicas in the context of time-series. The values in the weighting set vary in the interval from 0 to any positive integer, as long as the sum of the values in the set normalizes to the size of the dataset. By using this technique we can emphasize different time-points, thus replicating the effects of bootstrap replicas. In the learning phase this technique is used along with a particular objective function Weighted Root Mean Squared Error (WRMSE):

$$WRMSE = \frac{1}{\sum_{t=0}^{n} w_t} * \sqrt{\frac{\sum_{t=0}^{n} w_t * (y_t - \hat{y}_t)^2}{n}} \quad (1)$$

, where y_t and $\hat{y}_t$ denote the observed and simulated values of the system variables, n denotes number of time-points in the observed data, w_t denote the weight at time point t. More precisely, penalizing (encouraging) different time points from each bootstrap replica.

3 Inductive Process Modeling & ProBMoT

Equation discovery is a machine learning sub-field, which by using observations, aims to determine the scientific laws that govern the dynamics of a given system and induce them in form of equations. Moreover, Inductive Process Modeling (IPM) [5] is an equation discovery approach, which takes into consideration the domain-knowledge along to the observations, and thus producing an explanatory process-based model. These models consist of two basic components: entities and processes. The entities represent the state/subject of the system, whereas the processes represent the interaction between the entities thus modeling the system dynamics.

From the mathematical point of view, the entities represent the variables involved in an interaction represented by the processes, which results in a set of differential and algebraic equations. This set of equations is the model of the dynamic system. The domain-knowledge is embodied in a form of a library of generic entities and generic processes that represent vague concepts of the particular domain. A search is performed through all possible model variants described in the library, resulting in a set of candidate models. Each candidate model is then used in the process of estimating parameter values. In this step, each of the candidate models is simulated

with a set of parameter estimates. The set for which the simulated trajectory is most adequate to the observed data is chosen to be the model parameters. Each of the models is sorted by their estimation function, i.e. sum of squared errors (SSE), and the model with the minimal error/best fit is chosen as an output of the algorithm.

ProBMoT [6] is software tool for complete modeling, parameter estimation and simulation of process-based models. ProBMoT follows the basic IPM paradigm and employs domain-specific modeling knowledge formulated in a library. Overlooking the definition of specific model structures, conceptual models are formulated from these templates, representing an abstraction of a whole class of models. Significantly, the main feature of the ProbMoT is the use of more specific constraints and rules which gives the ability of constructing more feasible models by allowing another (higher) level of conceptualization of the domain knowledge. The conceptual models are significant due to their transparency to the domain experts and the ubiquity among domains. Generating conceptual models, represents a task which involves spatial segregation of a system into a variety of discrete fragments (layers) with a different role in the system, as well as aggregation or differentiation among layer These models define the system as set of components related between them, where each connection represents rule or an exchange of information, giving more logical representation of the modeled system. Using this approach, candidate models are created by applying all possible substitution rules that will represent descendant functional processes from the lowest level of the hierarchy.

The parameter estimation process is based on meta-heuristic optimization framework jMetal 4.3. [7] that implements a number of global optimization algorithms [8], this is to, avoid the fast convergence to a local-optimum when using local optimization algorithms . For simulating the candidate ODE models, ProBMoT employs the CVODE solver from SUNDIALS [9].

Basically, the process of parameter estimation resembles the one in IPM and follows the least-square approach for parameter identification, and tends to optimize a variety of quantitative objective functions. The most competitive advantage of the ProBMoT is the implementation of custom quantitative objective functions such as sum of squared errors (SSE), mean squared error (MSE), root mean squared error

(RMSE) and weighted RMSE (WRMSE) used for generating ensembles. Moreover this process has the ability to adapt to scenarios when models should are fitted on multiple observation datasets. This means, every chosen set of parameter estimates most sufficiently corresponds to every of the observation datasets. This process can be considered as a multi-objective optimization problem, but for most cases we translate it into a single-objective. The parameter set to be fitted can also include the initial values of the system.

In order to adequately meet the purpose of learning ensemble models of ODEs additional modifications of the ProBMoT were made. Basically, by tempering with the learning data, ProBMoT is able to generate diverse sets of output models which are considered as the members of the ensembles. Each of the models is fitted using the WRMSE function during the learning phase. The simulations of the member models are then combined into one by techniques such as averaging or weighted median, which represents the output of the ensemble model. The output model is again evaluated with the RMSE error function, thus providing a comparable measure of the performance of the ensemble to the single models in the evaluation process.

4 Experimental Evaluation

In this paper we address the task of learning ensemble models of nonlinear dynamic systems in the domain of aquatic ecosystems. More precisely, our case study is modeling the food-web dynamics of Lake Bled in Slovenia. The model includes three ecological variables represented as ordinary differential equations, i.e. phytoplankton concentration, dissolved phosphorus and zooplankton concentration. According to Atanasova et. al. [10] the dynamics of this ecosystem are severely complex, thus modeling this kind of system and obtaining acceptable results has proven to be a challenging task.

To completely reveal the effects of ensembles of ODE models and additionally to minimize the complexity of the modeled system, we focus on modeling just phytoplankton concentration, whereas the other two variables are assumed to be exogenous.

4.1 Datasets

The datasets used for these experiments were obtained from the Slovenian Environment Agency. The measurements consist of physical, chemical and biological data for the years from 1996-2002. All data was depth-averaged for the upper ten meters of the lake. Measurements where performed with a monthly frequency, and interpolated with cubic spline algorithm, thus obtaining supposable daily measurements. Each of the datasets consists of around 300 time points. For our experiment we used the datasets 1996-2001 for training the models and the 2002 dataset for evaluation.

4.2 Experimental setup

Using the process-based modeling formalism of ProBMoT we define a specific library describing the entities and processes involved in the dynamics of phytoplankton concentration variance. The library used is deduced from the complete library of aquatic ecosystem presented in [10], resulting in total of 128 candidate models. For the parameter estimation procedure we used Differential Evolution with rand/1/bin strategy, 1000 evaluations over a population space of 50 individuals. Our problem stated 20 parameters and 1 initial value eligible for estimation. We perform two experiments. In the first, we train single and use this as a baseline. In the second, we train an ensemble using the proposed bagging method with ten replicas and averaged output.

To evaluate the performance of the ensembles for this experiment we used the following setup, similar to the learning curve analysis, used in machine learning. At each iteration, we train a model on a combination of train data sets corresponding to measurements in the period from 1996-2001. The train model is then evaluated on a single data set corresponding to the 2002 measurements. In the first iteration, we use 2001 dataset for training. In each consequent iteration, we add one more preceding year to the combination of data sets used in the previous iteration (see the first column of Table 1). The performance of the models was evaluated using RMSE qualitative comparison.

4.3 Results

Table 1 summarizes the RMSE values for both the single model and ensemble model evaluation. As is stated before, several learning iterations were used for obtaining the Learning Curve (Figure1), each of them on a different subset of the training set. The results show the evaluation from the best models obtained. Mainly, presented here are the predictions from the best models in each iteration of the single model scenario, respectively only best models were considered in the process of generating the ensemble output. The ensemble output is generated, by averaging the predictions of each of the ten bootstrap replicas.

Table 1. Comparison of the test errors of a single model and an ensemble model obtained with bagging ten models

Train datasets (year)	Test dataset (year)	Single model	Ensemble Bagging/10/Average
'01	2002	1.252	2.359
'00,'01	2002	0.951	0.977
'99,'00,'01	2002	1.618	1.510
'98,'99,'00,'01	2002	1.624	1.536
'97,'98,'99, '00,'01	2002	1.672	1.635
'96,'97,'98,'99,'00,'01	2002	1.695	1.677

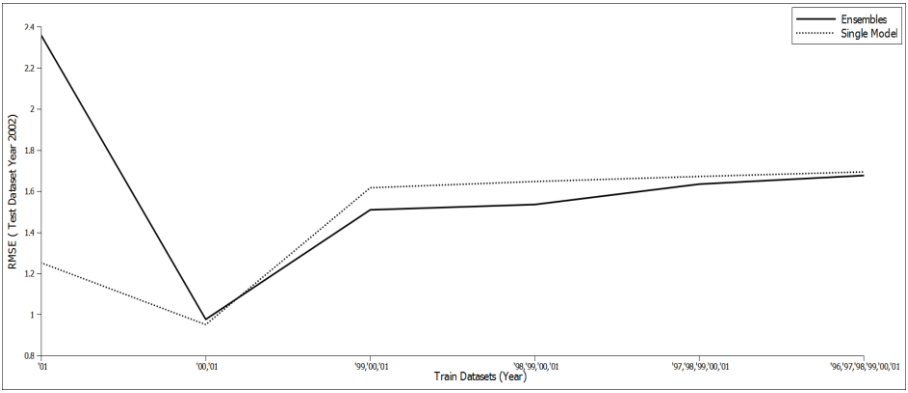

Figure 1. Learning Curve - Test errors of a single model and a bagged ensemble of ten models

4.4 Discussion

The results of the experiments show that the ensemble model performs slightly better on the given test data set than the single model. This slight increase in predictive performance comes at the cost of substantial increase in computational complexity.

This imposes two possible directions for improving the ensemble approach. We can overcome the high computational complexity of the ensemble approach by developing a parallel implementation which will bring the time required to build the ensemble close to the time required to build a single model. Second, we consider focusing on improving the ensemble model accuracy. According to Kunceva et.al [12], model diversity can be a key-factor of the performance of the ensembles. Having this in mind, we will consider a different technique for sampling the training data. Instead of selecting random time points, we can select random time windows consisting of several contiguous time points. This sampling approach will preserve the dynamics present in the input data which is lost when sampling with random time points.

The shape of the learning curve can be explained by the fact that the complexity of the problems by year varies, i.e., the observation data sets are of different lengths and the dynamics of the modeled state differs from year to year. An interesting phenomenon that can be spotted from this experimental setup, is adding additional training data beyond the second year does not improve the predictive performance. Adding just one preceding year is sufficient for improving the performance of both the ensemble model and the single model on this particular case study.

These results uncover important characteristics of the ensembles approach. Moreover, they clearly point to important directions for future improvement of the performance of this methodology.

5 Concluding remarks and Further work

We address the task of learning ensembles of ordinary differential equation models of aquatic ecosystem dynamics. To this end, we extend the state-of-the art approach for inductive process modeling. In particular, we adapted the ProBMoT, tool for automated process-based modeling, to the ability for generating diverse set of ODE models and their combination to an ensemble model. Additionally, we added the ability of leaning ODE models with custom designed objective functions and optimization techniques. These modifications are used to properly generate ensembles of dynamic models and evaluate their performance.

The case study considered is a model of phytoplankton growth in the Lake Bled in Slovenia, a complex nonlinear dynamic process part of the food-web dynamics of the lake. Our concern was generating an ensemble of diverse ODE models with the adapted bagging technique. The model considers one ecological variable from multiple observation data sets. Having this in mind, we considered a learning curve evaluation scenario. In summary, the results of our study show that ensemble models slightly improve the overall predictive performance.

The following possible directions for further work have been identified. First, the development of a methodology for increasing the diversity of the generated models by investigating different techniques for sampling the training set. Second, studying the model selection methods as a possible direction for improving the performance of the ensembles. In addition, experiments on other ecological domains can be performed, thus reaffirming the conclusions drawn in this paper. Moreover, adapting other state-of-the-art ensemble techniques in the context of ODE models, such as boosting, are to be investigated. Finally, we intent to examine domains other than ecological, thus extend the generality of our approach.

References:

[1] L.Todorovski and S.Dzeroski. Integrating Domain Knowledge in Equation Discovery, *Computational Discovery of Scientific Knowledge* vol. 4660, (Springer Berlin, Heidelberg, 2007), pp. 69-97.

[2] G. Seni and J.Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions.* Morgan and Claypool Publishers, 2009.

[3] O.Okun, G. Valentini and M.Re. *Ensembles in Machine Learning Applications* vol. 373, Springer, 2011.

[4] L.Breiman. Bagging predictors. *Machine Learning* 24, 2 (1996a), pp. 123-140.IPSSC

[5] W.Bridewell, P.Langley, L.Todorovski and S.Dzeroski. Inductive Process Modeling. *Machine Learning* 71 (2008), pp. 1-32.

[6] D. Cerepnalkoski, K.Taskova, L. Todorovski, N.Atanasova and S. Dzeroski. The influence of parameter fitting methods on model structure selection in automated modelling of aquatic ecosystems. *Ecological Modeling* 245:136-165,2012

[7] J.J.Durillo and A.J.Nebro. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software* 42:760-771,2011

[8] R.Storn and K.Price. Differential Evolution: A simple and efficient heuristic for global optimization over continuous space. *Journal of Global Optimization* 11(34),341-359, 1997

[9] S.Cohen and A.Hindmarsh. CVODE, a stiff/nonstiff ODE solver in C. *Computers in Physics*,10:138-143,1996

[10] N.Atanasova, L.Todorovski, S.Dzeroski, S. Reker-Remec, F.Reckangel and B.Kompare. Automated modelling of a food web in Lake Bled using measured data and a library of domain knowledge. *Ecological modelling,* vol 194. no.1-3,pp.37-48, 2006

[11] B. Panda, J. Herbach, S. Basu and R. Bayardo. PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce. *In Proceedings of International Conference on Very Large DataBases* (VLDB 2009), pages 1426-1437, Lyon, France, 2009.

[12] L.Kuncheva and C. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51, pp. 181-207, 2003

For wider interest

"A model of a system is a tool used to answer questions about the real systems without having to do an experiment" – Lennart Ljung

Engineers and mathematicians construct mathematical model to describe and predict the behavior of real dynamic systems under various conditions. These kinds of models normally are formulated with ordinary differential equations (ODEs), which represent a change of the state of the system over time. This process can either be theoretical or empirical. In the theoretical approach, experts use their knowledge of the domain to derive a mathematical equation of some process in the nature. On the other hand, the empirical paradigm uses measured data and tries to find the model that best fits the observed data in a trial and error process. This paradigm has been recently used to develop machine learning approaches to constructing ODE models from observed data. The state-of-the-art approaches have emerged that combine both the domain-knowledge and measured data to identify both the model structure and the values of the models parameters.

Our research is concerned with extending the existing machine leaning paradigms of automated modeling towards learning ensembles of ODE models. Our primary objective is adapting the existing and developing new techniques for combing ODE models, thus improving the descriptive and predictive performance of nonlinear and chaotic dynamic systems.