MACHINE LEARNING FOR SYSTEMS BIOSCIENCES

Sašo Džeroski
Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

ABSTRACT

The above title currently provides the best single label for the topics covered by my research group. Here I provide a brief summary of our current research and some of the research directions we intend to pursue.

Systems biosciences study biological systems in a holistic manner, focusing on the interactions among system components rather than the components themselves. Large amounts of data of increasing complexity are generated in these sciences (which include systems ecology and systems biology): The use of machine learning to make sense of these data is thus a necessity. We discuss two machine learning tasks that appear in this context, predicting structured outputs and automated modeling of dynamic systems. We describe some techniques for solving these tasks and some example application of these techniques in systems ecology and systems biology.

1 THE DATA DELUGE IN SYSTEMS BIOSCIENCES

Systems biosciences study biological systems in a holistic manner, trying to understand their overall behavior. They focus on the interactions among system components rather than the individual components themselves. Systems biosciences include systems ecology and systems biology, both of which have a strong emphasis on systems modeling.

At the macro level, systems ecology studies the behavior of ecosystems, which comprise populations of different organisms. It is part of environmental sciences and strongly related to ecological modeling. At the micro level, systems biology studies the processes that happen in individual organisms, focusing on cellular processes at the molecular level. It is a novel branch of the life sciences and relies heavily on a variety of bioinformatics approaches.

Ecology is concerned with the distribution of species over space and time. Ecological modeling is concerned with constructing and using models of ecosystems, i.e., communities/ populations of different species. Common types of ecological models include habitat and population dynamics models. The former focus on space and predict the suitability of a location for a given species. The latter focus on time and model the changes in time for one or more populations in a given ecosystem.

Data in ecology and environmental sciences can come in large quantities and many varieties. Spatial data can come from remote sensing and geographical information systems. Temporal data describe the time courses of population densities. Communities of organisms present at a sampling site can be represented as hierarchies (subsets of the taxonomy of living organisms). Microscopy images of organisms present at a site are often taken. Graphs can be used to represent the chemical structure of pollutants.

The data deluge is of even larger proportions and variety in systems biology. Systems biology is a new branch of the life sciences that tries to understand organisms as a whole. It attempts to put together an integrated picture of the processes that happen in the system at all levels and their dynamics. The levels of organization of the systems range from the genome to the phenome, with high-throughput data being collected for each of them by the different "–omics" disciplines, such as genomics, transcriptomics, proteomics, metabolomics, and phenomics.

Among the different types of data, genomics data in the form of DNA/RNA sequences are increasingly common, with new genomes being sequenced at a rapidly increasing rate. While spatial data are still rare, temporal data in the form of time courses of gene expression levels, protein and metabolite concentrations are being collected. Hierarchies describe the structure of proteins (classified into folds and families) and protein/ gene function (described in terms of annotation schemes, such as the Gene Ontology). Microscopy images of cellular cultures (or videos thereof), collected within genomic and compound screens are gaining popularity. In the context of studying QSAR (quantitative structure-activity relationships), the structure of chemical compounds is typically represented in the form of graphs.

To make sense of data that come in such large quantity and variety, the use of machine learning (ML) techniques is a must. ML, an essential area of artificial intelligence, studies computer programs that learn from (automatically improve with) experience. A major part of ML (inductive learning) is concerned with learning from examples. This part has a large overlap with the area of data mining, which includes supervised learning (or predictive modeling) and unsupervised learning (including clustering). Machine learning also studies the topic of computational scientific discovery, which includes approaches for automated modeling of dynamic systems.

In the remainder of this paper, we first describe the general task of predictive modeling, concerned with predicting structured outputs, and some approaches to solving it. We then describe the task of automated modeling of dynamic systems and some approaches to solving it. Several example applications of these approaches in systems ecology and systems biology are discussed next. We finally outline some directions for further research.

2 ML FOR PREDICTING STRUCTURED OUTPUTS

The task of predicting structured outputs. The task of predictive modeling is to learn a predictive model from

examples. A predictive model returns an output (i.e., target) property of an example, given input properties (e.g., attribute) of the example. Typically, we need to predict the scalar value of a single variable: The task is called classification when the target is discrete and regression when it is real-valued.

However, there are many real life domains, such as image annotation, text categorization, predicting gene functions, etc., where the input and/or the output can be structured. We will focus here on the latter, namely, on predictive modeling tasks with structured outputs. The inputs will have the form of vectors of attribute values.

We consider three different classes of targets: tuples of discrete/real values, hierarchies of discrete values (classes), and (short) time series of real values. The corresponding tasks of structured output prediction are called multi-target prediction, hierarchical multi-label classification, and prediction of (short) time series.

Descriptive attributes						Target attributes					
Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2		Cladophora sp.	Gongrosira incrustans	Oedogonium sp.	Stigeoclonium tenue	Melosira varians	Nitzschia palea
0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1
2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1
3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0

Table 1: An example task of multi-target classification.

Multi-target prediction is the simplest extension of the classical predictive modeling task: Instead of a single discrete/continuous target variable, we need to predict several of these. Typically, all of the targets are discrete (resp. continuous) and we face the tasks of multi-target classification and regression. An example task of multi-target classification is given in Table 1. The attributes, targets, and three instances from a dataset are shown. The descriptive attributes are physical and chemical parameters of water quality, while the target variables denote the presence/absence of bioindicator organisms.

In the multi-target prediction task from Table 1, all targets are binary. In this case, each target can be considered a label, and each example can be assigned more than one label (if a value of one is predicted for the corresponding target). This variant of the multi-target prediction task is called multi-label classification (MLC).

Hierarchical multi-label classification (HMLC) can be considered an extension of multi-label classification. Like in MLC, in HMLC we need to predict a subset of the set of possible labels for each example. However, as the name implies, labels in HMLC are organized in a hierarchy. Labelings (predictions) in HMLC have to obey the hierarchy constraint: If a label is predicted, all of its parent labels have to be predicted as well.

Prediction of (short) time series is the last task we consider here. The (short) time series are sequences of real-valued measurements, taken at consecutive timepoints. The task is to predict such a sequence from the values of a set of descriptive attributes.

Approaches for predicting structured outputs. There are two groups of methods for solving the task of predicting structured outputs: (1) methods that predict component(s) of the output and then combine the components to get the overall prediction (called local methods) and (2) methods that predict the complete structure as a whole (called global methods). The latter group of methods has several advantages over the former. They exploit the dependencies that exist between the components of the structured output in the model learning phase and thus result in better predictive performance. They are also more efficient: the number of components in the output can be very large (e.g., hierarchies in functional genomics), in which case executing a basic method for each component is not feasible.

A variety of methods now exist for predicting structured outputs: An overview of these is beyond the scope of this paper. The majority of approaches to predicting structured outputs can handle only one type of output (e.g., multi-target classication). An exception to this is the approach of predictive clustering: Besides unifying supervised and unsupervised learning (prediction and clustering), predictive clustering also allows for structured prediction of different output types. In particular, all of the three tasks above can be handled by predictive clustering trees (PCTs). An example PCT for the task from Table 1 is given in Figure 1.

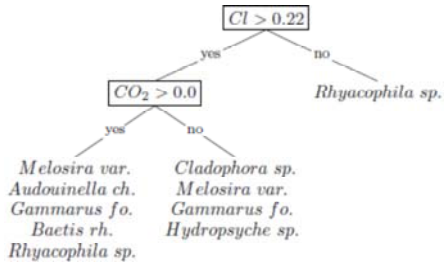

Figure 1: An example tree for multi-target classification.

Recent advances in predicting structured outputs. Among the recent developments, we will focus here on those related to predictive clustering. The predictive power of PCTs has been greatly enhanced by the introduction of ensembles of PCTs. Methods have also been developed for learning predictive clustering rules, as well as rule ensembles. Finally, methods for constrained (e.g. by size and error) induction of PCTs have been developed (cf. Chapters 7 and 15 of Džeroski et al. (2010)).

3 ML FOR MODELING DYNAMIC SYSTEMS

Computational scientific discovery (CSD) attempts to provide computational support for the process of discovery

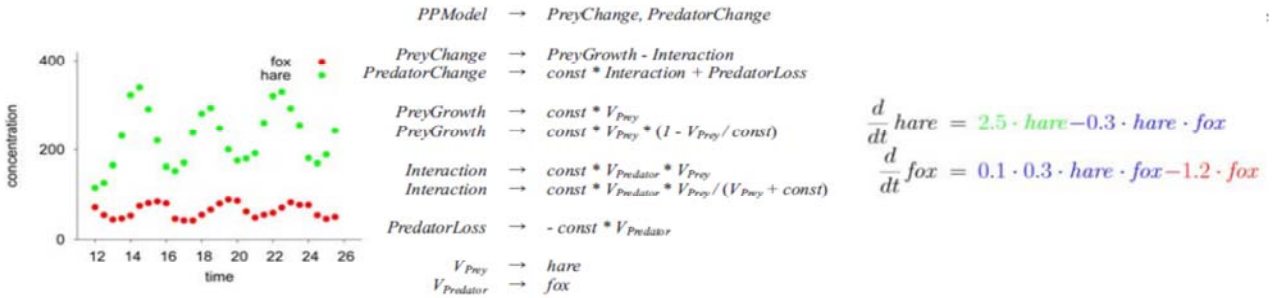

Figure 2: An example task of automated modeling.

of scientific knowledge. The input to CSD approaches includes observations (measured data) and existing scientific (domain) knowledge. The focus in CSD is on using standard formalisms for representing scientific knowledge that are accepted and routinely used by scientists.

The task of automated modeling of dynamic systems belongs to the CSD paradigm. For dynamic systems that change their state over time, models in the form of ordinary differential equations (ODEs) have to be learned, including both the structure of and the parameters in the equations. ODEs are well known, accepted and widely used in science.

The CSD paradigm facilitates the combination of both observational data and existing background/domain knowledge, where different types of domain knowledge can be used. We can start from existing ODE models for the system at hand (that are partial, incomplete, and/or inaccurate) and revise/improve them in light of observed data. We can also provide a set of basic components as building blocks from which ODE models can be built. The domain knowledge can be provided in different forms, including grammars, constraints, and process-based modeling formalisms.

An example input / output for a task of automated modeling of population dynamics with ODEs is given in Figure 2. At the far left, the observed data are given, consisting of the densities of two populations. In the middle, the domain knowledge is given in the form of a context-free grammar specifying the functional form of the ODEs to be considered. At the far right, the output is given, consisting of two ODEs describing three population dynamics processes (prey growth – first term in first equation, predator loss – second term in second equation, and predator-prey interaction – term shared by both equations).

The use of domain knowledge allows for the convergence of the two major modeling paradigms, theoretical (knowledge-driven) modeling and empirical (data-driven) modeling. In the first approach, a domain expert derives a proper model structure based on domain-specific modeling knowledge: Extensive knowledge and little data are needed. In the second approach, different model structures are explored to fit observed data in a trial and error process: Extensive data and little domain knowledge are necessary. The integrated approach to automated modeling allows us to trade between the quantity

Approaches for automated modeling of dynamic systems. These approaches use measured data to identify both the model structure and the values of the constant parameters in the model. They combine heuristic search methods with parameter estimation techniques: While the search methods explore the space of candidate model structures, the parameter estimation techniques find optimal values of a single structure and evaluate its fit against observations. The result of the evaluation in turn guides the search method towards a model (i.e., a structure/parameters combination) with a good fit. Alternative approaches differ in the type and representation of domain knowledge.

Recent advances in automated modeling of dynamic systems. The most recent methods that follow this paradigm use process-based formalisms for representing domain knowledge as well as models. The key concepts here are entities and processes. Entities, which can be of different types, describe the components of the modeled system: Their properties are typically system variables in the ODEs. Processes describe the interactions between entities and give rise to the terms that appear in the ODEs. Process-based domain knowledge lists the basic types of entities and processes in the domain at hand: The types can be organized hierarchically. For each type of process, alternative modeling templates are provided, which can be used in the ODEs. A representative for this class of approaches is LAGRAMGE2 (cf. Chapter 4 of (Džeroski and Todorovski 2007)).

4 ML FOR SYSTEMS ECOLOGY

There are many applications of machine learning to ecological and environmental modeling problems. These range from predicting earthquakes to assessing the state of the environment from remote sensing data: Forest height and density can be predicted in this fashion. Here we will focus on applications of the two classes of ML methods described above to the ecological modeling tasks of modeling habitat and modeling population dynamics.

Methods for structured output prediction can and have been applied to problems of relating environmental conditions to community structure. This is a generalization of the problem of modeling habitat for a single species, as the presence/absence or abundance has to be predicted for several species (or higher taxonomic units) rather than a single one. PCTs have been applied to predict community structure in Slovenian rivers (Table 1 and Figure 1), Lake Prespa in Macedonia and polder lakes in Belgium.

CSD methods for automated modeling of dynamic systems have been applied to many problems of modeling population dynamics. In fact, applications in automated modeling of population dynamics have been the driving force for the development of many of the methods mentioned above. The automated modeling of the dynamics of two populations is illustrated in Figure 2.

Automated modeling approaches have been mainly applied to modeling the population dynamics of aquatic ecosystems. The Lagoon of Venice was the first one to be modeled in this fashion, followed by a number of lakes, including Lakes Glumsoe (Denmark), Bled (Slovenia), and Kasumigaura (Japan), cf. Čerepnalkoski et al (2012), as well as Greifensee (Switzerland) and Kinneret (Israel).

5 ML FOR SYSTEMS BIOLOGY

In systems biology, ML methods for structured output prediction can be used for integrative analysis of high-throughput data being collected by the different "–omics" disciplines. In addition, a variety of knowledge (from different sources) collected by humans and stored in bioinformatics databases can be used in these analyses: These include for example annotations of genes with their functions in terms of the Gene Ontology. The learned model and especially their (in-silico) predictions are often further used by biologists as hypothesis that are examined and validated by experiments in the wet-lab.

Tasks addressed in this context include gene function prediction and relating time course profiles of gene expression level to gene properties (function). The former is an instance of HMLC and has been addressed for several model organisms, including yeast and water cress (Chapter 15 of Džeroski et al. (2010)), as well as the mouse (Schietgat et al. 2010). The latter is an instance of predicting short time series and has been used to relate the function of yeast genes and the time profiles of their response to different types of stress (Slavkov et al. 2010).

A major focus of systems biology is the study of the structure and dynamics of biological networks. After the structure of (links in) the network is determined, the kinetics of its behavior is described with ODE models, which can include different types of kinetics. Approaches for automated modeling of dynamic systems can be applied in this context. Džeroski and Todorovski (2008) give an overview of different methods of this kind, as well as the possibilities of their use in systems biology. They also give an example of using such methods for modeling the process of glycolysis. More recently, Tashkova et al. (2011) have addressed the task of modeling endocytosis, more precisely endosome maturation, an important part of the immune response manipulated by pathogens (e.g., *M. tuberculosis*).

6 FUTURE RESEARCH DIRECTIONS

Predicting structured outputs. Besides multi-target prediction, HMLC, and predicting short time series, we need to consider other task with different, possibly more complex targets (e.g., tuples of hierarchies or time series). Methods for learning PCTs can be extended in several directions, such as integration with semi-supervised lerning approaches or transfer learning. Handling imbalanced distributions is another major challenge in this context. Finally, we also need to explicitly take into account spatio-temporal information, e.g., by considering autocorrelation.

Approaches for automated modeling of dynamic systems. These approaches integrate search through the space of ODE structures with parameter fitting. After decades of using local optimization methods for the latter task, global optimization approaches (meta-heuristics) have been considered recently (Tashkova et al. 2011, Čerepnalkoski et al. 2012). However, much work remains to be done. The use of global optimization methods opens the door to using other quality criteria (and not just mean squared error) for selecting models: Multiple criteria could also be used with multi-objective optimization. This would broaden the possible application areas for CSD methods. Extensions in the direction of incorporating spatial aspects within the temporal models (e.g., by using compartments) are also of practical relevance. Finally, we will explore the avenue of learning ensembles of ODE models of dynamic systems and the use of such approaches in modeling .environmental and biological systems.

Applications in systems biosciences. We anticipate that the number and types of possible applications for the methods described above in systems biosciences will explode over the next decade. In ecology, one task on the horizon is predicting community structure by taking into account the taxonomy of living organisms (a HMLC task). In systems biology, the analysis of data from image screens, either genomic or compound, will need immediate attention: This is a task where data are complex, both in quantity and in quality (structure).

Acknowledgements: The author is supported by the FP7 EU projects SUMO and REWIRE.

References

[1] D. Čerepnalkoski, K. Taškova, L. Todorovski, N. Atanasova, S. Džeroski. The influence of parameter fitting methods on model structure selection in automated modeling of aquatic ecosystems. *Ecological Modeling*, In press, 2012.

[2] S. Džeroski, and L. Todorovski, eds. *Computational Discovery of Scientific Knowledge*. Springer, Berlin, 2007.

[3] S. Džeroski and L. Todorovski. Equation discovery for systems biology: finding the structure and dynamics of biological networks from time course data. *Current Opinion in Biotechnology*, 19(4): 360-368, 2008.

[4] S. Džeroski, B. Goethals, and P. Panov, eds. *Inductive Databases and Constraint-based Data Mining*. Springer, Berlin, 2010.

[5] L. Schietgat, C. Vens, J. Struyf, H.Blockeel, D. Kocev, S. Džeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics* 11 (2), 2010.

[6] I. Slavkov, V. Gjorgjioski, J. Struyf, S. Džeroski. Finding explained groups of time-course gene expression profiles with predictive clustering trees. *Molecular bioSystems*, 6 (4): 729-740, 2010.

[7] K. Tashkova, P. Korosec, J. Silc, L. Todorovski, S. Džeroski. Parameter estimation with bioinspired meta-heuristic optimization: modeling the dynamics of endocytosis. *BMC Systems Biology* 5:159, 2011.