Automatic Modeling of Dynamical Interactions Within Marine Ecosystems

Omar Iken[1], Maxime Folschette[1], and Tony Ribeiro[2]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[2] Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France

Abstract. Marine ecology models are used to study and anticipate population variations of plankton and microalgae species. These variations can have an impact on ecological niches, the economy or the climate. Our objective is the automation of the creation of such models. Learning From Interpretation Transition (LFIT) is a framework that aims at learning the dynamics of a system by observing its state transitions. LFIT provides explainable predictions in the form of logical rules. In this paper, we introduce a method that allows to extract an influence graph from a LFIT model. We also propose an heuristic to improve the model against noise in the data.

Keywords: logical modeling · dynamic systems · heuristics · interaction graph

1 Introduction

Marine ecosystems represent the majority of aquatic systems on Earth. These ecosystems have an important impact, such as regulating climate change, providing food and maintaining biodiversity, etc. The understanding of these systems is therefore particularly interesting, but their complexity makes it difficult to create new models. A key component of these ecosystems is the *phytoplankton*, which are microscopic organisms present at the surface of all aquatic ecosystems. They are the basis of aquatic food as they are at the lowest level of the oceanic food chain, in addition to being responsible for the production of a large part of the planet's oxygen. Since 1992, the SRN network has gathered samplings of sea water in order to measure phytoplankton population variations along with environmental factors [3].

Learning From Interpretation Transition (LFIT) [2] aims to automate the construction of models of dynamic systems from their state transitions. LFIT produces an explainable model that describes the whole system dynamics and can also predict its probable future variations.

In this work, we use LFIT to learn causal models of marine ecosystems from the dataset of the SRN network. The method we propose in this paper can help biologists to identify which factors and species are in interaction. These models are also useful to predict future population changes in plankton species, and thus may help to predict the evolution of climate change.

2 State Of The Art

LFIT The LFIT framework [2] is based on the notion of *dynamical transition*, that is, an atomical transition in a discrete time series. Given a set of such transitions, LFIT builds a model of the system dynamics under the form of a set of rules as follows:

$$\underbrace{\mathrm{v}_0^{\mathrm{val}_0}}_{\text{head}} \leftarrow \underbrace{\mathrm{v}_1^{\mathrm{val}_1} \wedge \cdots \wedge \mathrm{v}_m^{\mathrm{val}_m}}_{\text{body}}$$

Such a rule is said to *match* a given state if, for each $i \in \{1, 2, \ldots, m\}$, the variable v_i has the value val_i in the current state. Intuitively, the rule above means that the variable v_0 can take the value val_0 in the next state if the rule matches the current state. We distinguish two types of rules on a given matched state: (1) *likeliness* rules, where the conclusion is observed in at least one transition from this state, and (2) *unlikeliness* rules, where the conclusion is never observed in any transition from this state. Each rule of both type is weighted by the number of states it matches in the observations.

DATASET In this work we focus on the SRN dataset [3]. This dataset includes long-term time series of marine phytoplankton and physical-chemical measurements. Water samples were collected along the eastern English Channel coast every 15 to 30 days from 1992 to 2020, at different depths and locations. Each sample in the dataset is characterized by its sampling location and depth, sampling date, name of the measured environmental factor or phytoplankton species, and value of measurement. There are several hydrological factors like temperature, oxygen or salinity, associated to different units, while for each phytoplankton species, the number of individuals is counted. Here we focus on one location of interest: the coastal station 1 of Boulogne-sur-Mer, and focus only on 11 hydrological factors and 12 phytoplankton species only sampled at sea surface level, as it was done in [1].

3 Preprocessing

Data sampling being irregular, we re-sample the measurements for the different factors and species on a monthly basis to minimize missing data. Furthermore, since LFIT only works with abstract discrete values for state transitions, we also need to discretize the measurements. This step has a significant impact on the results provided by the LFIT algorithms. Here, hydrological factors are discretized in two values: *below average* and *above average*; phytoplankton species in three values: *low*, *below average* and *above average*, where *low* encodes a population too low to be counted.

After discretization, we have a time series of 308 data point from which are extracted the 307 atomic transitions to be processed by LFIT. These transitions are divided into 80% (253 transitions) for the learning set and 20% (54 transitions) for the test set. The fitted model has 1683 likeliness rules and 1981 unlikeliness rules, and its accuracy is about 67.0% on the test set.

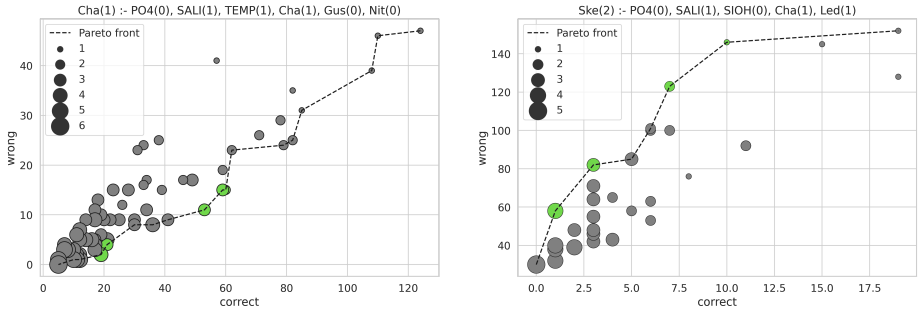

Fig. 1: Distribution and Pareto frontier of weighted combinations for a likeliness rule (left) and an unlikeliness rule (right). Correct weight is in abscissa and wrong weight is in ordinate. The size of circles corresponds to the number of conditions in the combination, green circles being the "best" alternatives to consider.

4 Rules Improvement

Training data being incomplete and noisy, the output of LFIT is not perfect, i.e., rules may be improved and some could be removed. Indeed, LFIT only learns rules that are true on every single transition they match in the training set. This is why we propose a heuristic to simplify rules, in order to improve their quality and reduce their quantity, thus improving improve model accuracy and readability.

For that, we simply generate all subsets of each rule body, and then count how many transitions validate/falsify the new rule, thus giving us two weights: a *correct weight*, and a *wrong weight*. For a likeliness rule, we want to maximize the correct weight and minimize the wrong weight, and conversely for an unlikeliness rule. We can then compute the *Pareto frontier* (of correct/wrong weights) to find the best subsets of a given body, as shown in Figure 1.

From the Pareto front, we extract what we consider the best rules based on the ratio of correct and wrong weights. We chose an arbitrary ratio of 2 for likeliness rules and $1/2$ for unlikeliness rules.

By eliminating the redundant rules (subsumed by another rule) from the improved model, we obtain a set of 1609 likeliness rules and 1405 unlikeliness rules. With this, we manage to obtain a new accuracy of 71.6% on the test set.

5 Influence Graph

From the detailed interactions in the form of rules, a more abstract view of the system can be extracted in the form of an influence graph containing the mutual influences between variables. This allows to identify whether an environmental factor or a species is a strict activator or inhibitor of another. The point is to identify which variable need to be controlled in order to impose a desired behavior on a targeted species. We propose to extract such model from LFIT

$$\mathrm{a}^0 \leftarrow \mathrm{a}^0 \wedge \mathrm{b}^1 \wedge \mathrm{c}^2 \implies \begin{cases} a \xrightarrow{+} a \\ b \xrightarrow{?} a \\ c \xrightarrow{-} a \end{cases}$$

$$\mathrm{a}^0 \leftarrow \mathrm{a}^1 \implies a \xrightarrow{-} a$$

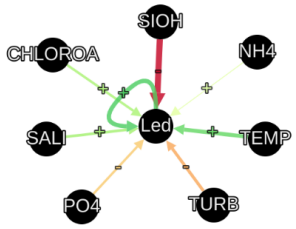

Fig. 2: Left: an example of influence extraction on variables with 3 expression levels (0, 1 and 2). Right: an example of the final influence graph of the phytoplankton species Led, where negative influences are in red, positive influences in green, and color intensity corresponds to the certainty.

output rules by comparing the value of the head of each rule with the value of each feature of the body of the same rule. Figure 2 (left) shows an example of the process on two rules of the same variable a. For instance, a positive influence is considered when a given body value makes the head evolve the same way (such as a^0 in the example which produces a^0) and conversely for a negative influence. This produces a score that allows to compute the influence between each pair of variables. Figure 2 (right) shows the resulting influence graph restricted to one phytoplankton species (Led) once all corresponding rules have been processed.

6 Conclusion

In this paper, we presented new methods to both improve a LFIT model and extract an underlying influence graph that can be used by biologists to get new insights of the studied data. The accuracy of the method could be improved with a stricter pre-processing of the data, a fine-tuning of the rule improvement from the Pareto frontier, and exploring other ways of building the influence graph. Yet, it consists in a first encouraging step towards the automation of marine ecosystem models.

Acknowledgments The authors would like to thank Sébastien Lefebvre and Stéphane Karasiewicz for providing the data and for their help.

References

1. Karasiewicz, S., Breton, E., Lefebvre, A., Hernández Fariñas, T., Lefebvre, S.: Realized niche analysis of phytoplankton communities involving hab: Phaeocystis spp. as a case study. Harmful Algae **72**, 1–13 (2018)
2. Ribeiro, T., Folschette, M., Magnin, M., Roux, O., Inoue, K.: Learning dynamics with synchronous, asynchronous and general semantics. In: International Conference on Inductive Logic Programming. pp. 118–140. Springer (2018)
3. SRN – Regional Observation and Monitoring program for Phytoplankton and Hydrology in the eastern English Channel: SRN dataset. 1992-2016. (2017)