Guidelines for the supervised learning of species interactions

Timothée Poisot [1,2]

[1] Université de Montréal [2] Québec Centre for Biodiversity Sciences

Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

Last revision: *December 8, 2021*

1. The prediction of species interaction networks is gaining momentum as a way to circumvent limitations in data volume. Yet, ecological networks are challenging to predict because they are typically small and sparse. Dealing with extreme class imbalance is a challenge for most binary classifiers, and there are currently no guidelines as to how predictive models can be trained.

2. Using simple mathematical arguments and numerical experiments in which a variety of classifiers (for supervised learning) are trained on simulated networks, we develop a series of guidelines related to the choice of measures to use for model selection, and the degree of unbiasing to apply to the training dataset.

3. Classifier accuracy and the ROC-AUC are not informative measures for the performance of interaction prediction. PR-AUC is a fairer assesment of performance. In some cases, even standard measures can lead to selecting a more biased classifier because the effect of connectance is strong. The amount of correction to apply to the training dataset depends as a function of the classifier and the network connectance.

4. These results reveal that training machines to predict networks is a challenging task, and that in virtually all cases, the composition of the training set needs to be experimented on before performing the actual training. We discuss these consequences in the context of the low volume of data.

example on diagnostic test: rare events are hard to detect even with really good models

summary of model challenges for networks

introduction to the confusion matrix

list of problems to solve - baseline values and response to bias - effect of training set bias on performance - which models need the least amount of interactions to work

summary of the results

Baseline values

In this section, we will assume a network of connectance ρ, *i.e.* having ρS^2 interactions (where S is the species richness), and $(1-\rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the network is a column vector $\mathbf{o}^T = [\rho(1-\rho)]$ (we can safely drop the S^2 terms, as we will work on the confusion matrix, which ends up expressing *relative* values).

In order to write the values of the confusion matrix for a hypothetical classifier, we need to define two characteristics: its skill, and its bias. Skill, here, refers to the propensity of the classifier to get the correct answer (*i.e.* to assign interactions where they are, and to not assign them where they are not). A no-skill classifier guesses at random, *i.e.* it will guess interactions with a probability ρ. The predictions of a no-skill classifier can be expressed as a row vector $\mathbf{p} = [\rho(1-\rho)]$. The confusion matrix $\mathbf{M}$ for a no-skill classifier is given by the element-wise product of these vectors $\mathbf{o} \odot \mathbf{p}$, *i.e.*

$$\mathbf{M} = \begin{pmatrix} \rho^2 & \rho(1-\rho) \\ (1-\rho)\rho & (1-\rho)^2 \end{pmatrix}$$

Numerical experiments

Effect of training set on performance

Required amount of positives to get the best performance

Guidelines for prediction

References