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.

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

2   summary of model challenges for networks

3   introduction to the confusion matrix

4   list of problems to solve - baseline values and response to bias - effect of training set bias on performance -

5   which models need the least amount of interactions to work

6   summary of the results

7   Baseline values

8   In this section, we will assume a network of connectance ρ, *i.e.* having ρS^2 interactions (where S is the

9   species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the

10 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

11 confusion matrix, which ends up expressing *relative* values).

12 In order to write the values of the confusion matrix for a hypothetical classifier, we need to define two

13 characteristics: its skill, and its bias. Skill, here, refers to the propensity of the classifier to get the correct

14 answer (*i.e.* to assign interactions where they are, and to not assign them where they are not). A no-skill

15 classifier guesses at random, *i.e.* it will guess interactions with a probability ρ. The predictions of a no-skill

16 classifier can be expressed as a row vector $\mathbf{p} = [\rho(1 - \rho)]$. The confusion matrix $\mathbf{M}$ for a no-skill classifier

17 is given by the element-wise product of these vectors $\mathbf{o} \odot \mathbf{p}$, *i.e.*

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

18 In order to regulate the skill of this classifier, we can define a skill matrix $\mathbf{S}$ with diagonal elements equal

19 to s, and off-diagonal elements equal to $(1 - s)$, and re-express the skill-adjusted confusion matrix as

20 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

$$\begin{pmatrix} \rho^2 & \rho(1-\rho) \\ (1-\rho)\rho & (1-\rho)^2 \end{pmatrix} \odot \begin{pmatrix} s & (1-s) \\ (1-s) & s \end{pmatrix}.$$

21 Note that when $s = 0$, $\mathbf{Tr}(\mathbf{M}) = 0$ (the classifier is *always* wrong), when $s = 0.5$, the classifier is no-skill

22 and guesses at random, and when $s = 1$, the classifier is perfect.

23 The second element we can adjust in this hypothetical classifier is its bias, specifically its tendency to

24 over-predict interactions. Like above, we can do so by defining a bias matrix $\mathbf{B}$, where interactions are

25 over-predicted with probability b, and express the final classifier confusion matrix as $\mathbf{M} \odot \mathbf{S} \odot \mathbf{B}$, *i.e.*

$$
\begin{pmatrix} \rho^2 & \rho(1-\rho) \\ (1-\rho)\rho & (1-\rho)^2 \end{pmatrix} \odot \begin{pmatrix} s & (1-s) \\ (1-s) & s \end{pmatrix} \odot \begin{pmatrix} b & b \\ (1-b) & (1-b) \end{pmatrix}.
$$

26 The final expression for the confusion matrix in which we can regulate the skill and the bias is

$$
\mathbf{C} = \begin{pmatrix} s \times b \times \rho^2 & (1-s) \times b \times \rho(1-\rho) \\ (1-s) \times (1-b) \times (1-\rho)\rho & s \times (1-b) \times (1-\rho)^2 \end{pmatrix}.
$$

27 Numerical experiments

28 Effect of training set on performance

29 Required amount of positives to get the best performance

30 Guidelines for prediction

31 References