

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

This work is released by its authors under a CC-BY 4.0 license



Last revision: *December 10, 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 assessment 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 - Strydom et al. (2021) importance of drawing on traits +
3 validation is challenging - Whalen et al. (2021) machine learning from genomics
4 Binary classifiers are usually assessed by measuring properties of their confusion matrix, *i.e.* the
5 contingency table reporting true/false positive/negative hits. A confusion matrix is laid out as

$$\begin{pmatrix} \text{tp} & \text{fp} \\ \text{fn} & \text{tn} \end{pmatrix},$$

6 wherein tp is the number of interactions predicted as positive, tn is the number of non-interactions
7 predicted as negative, fp is the number of non-interactions predicted as positive, and fn is the number of
8 interactions predicted as negative. Almost all measures based on the confusion matrix express rates of
9 error or success as proportions, and therefore the values of these components matter in a *relative* way. At a
10 coarse scale, a classifier is *accurate* when the trace of the matrix divided by the sum of the matrix is close
11 to 1, with other measures focusing on different ways in which the classifier is wrong.

12 The same approach is used to evaluate *e.g.* species distribution models (SDMs). Indeed, the training and
13 evaluation of SDMs as binary classifiers suffers from the same issue of low prevalence. In a previous work,
14 Allouche et al. (2006) suggested that κ was a better test of model performance than the True Skill Statistic
15 (TSS), which we will refer to as Youden's informedness (or J); these conclusions were later criticized by
16 Somodi et al. (2017), who emphasized that informedness' relationship to prevalence depends on
17 assumptions about bias in the model. Although this work offers recommendations about the comparison
18 of models, it doesn't establish baselines or good practices for training on imbalanced ecological data.
19 Within the context of networks, there are three specific issues that need to be addressed. First, what values
20 of performance measures are we expecting for a classifier that has poor performance? This is particularly
21 important as it can evaluate whether low prevalence can lull us into a false sense of predictive accuracy.
22 Second, independently of the question of model evaluation, is low prevalence an issue for *training*, and
23 can we remedy it? Finally, because the low amount of data on interaction makes a lot of imbalance
24 correction methods (see *e.g.* Branco et al., 2015) hard to apply, which indicators can be optimized with the
25 least amount of positive interaction data?

26 We establish that due to the low prevalence of interactions, even poor classifiers applied to food web data

will reach a high accuracy; this is because the measure is dominated by the accidental correct predictions of negatives. The F_1 score and positive predictive values are less sensitive to bias, but **TODO** Chicco et al. (2021) - MCC maximizes other measures, other measures do not maximize MCC, except notably when prevalence is low, or the baseline guessing level is uncertain, which applies to networks. In this cases, informedness should be used as a maximization criterion.

Baseline values

Confusion matrix with skill and bias

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.*

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

In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal to s , and off-diagonal elements equal to $(1 - s)$, and re-express the skill-adjusted confusion matrix as $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

Note that when $s = 0$, $\text{Tr}(\mathbf{M}) = 0$ (the classifier is *always* wrong), when $s = 0.5$, the classifier is no-skill and guesses at random, and when $s = 1$, the classifier is perfect.

The second element we can adjust in this hypothetical classifier is its bias, specifically its tendency to over-predict interactions. Like above, we can do so by defining a bias matrix \mathbf{B} , where interactions are over-predicted with probability b , and express the final classifier confusion matrix as $\mathbf{M} \odot \mathbf{S} \odot \mathbf{B}$, *i.e.*

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

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

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

In all further simulations, the confusion matrix \mathbf{C} is transformed so that it sums to 1.

What are the baseline values of performance measures?

In this section, we will change the values of b and s for a given value of ρ , and see how the values of common performance measures for binary classification are affected. Specifically, we will focus on four quantities: the accuracy $((\text{tp} + \text{tn})/(\text{tp} + \text{tn} + \text{fp} + \text{fn}))$, the *balanced* accuracy $(\text{tp}/(2(\text{tp} + \text{fn})) + \text{tn}/(2(\text{tn} + \text{fp})))$, Youden's J $(\text{tp}/(\text{tp} + \text{fn}) + \text{tn}/(\text{tn} + \text{fp}) - 1)$, and the F_1 score $(2\text{tp}/(2\text{tp} + \text{fp} + \text{fn}))$.

Justification of why these 4

Assuming a no-skill unbiased classifier ($s = 0.5$, *i.e.* $\mathbf{C} = \mathbf{M}$ after normalization), the accuracy is $\rho^2 + (1-\rho)^2$, the balanced accuracy is 0.5, Youden's J is 0, and $F_1 = \rho$. In other words, given a connectance $\rho = 0.05$, we expect that a classifier guessing at random would still achieve an accuracy of 0.905; for a connectance of $\rho = 0.01$, this accuracy *increases* to over 0.98. In other words, networks with fewer interactions have inherently higher accuracy, because it is easy to predict the overwhelming majority of non-interactions right.

67 In order to examine how these values change w.r.t. the skill and bias, we performed a grid exploration of
68 the values of $\text{logit}(s)$ and $\text{logit}(b)$ linearly from -10 to 10 , and visualize the result for a connectance of
69 $\rho = 0.15$, which is within the range of usually observed connectance values for empirical food webs.

70 [Figure 1 about here.]

71 Numerical experiments

72 In the following section, we will generate random networks, and train four binary classifiers (as well as an
73 ensemble model using the sum of the outputs) on 30% of the interaction data. Networks are generated by
74 picking random generality g and vulnerability v traits for $S = 200$ species uniformly on the unit interval,
75 and assigning an interaction from species i to species j if $0.2g_i - \xi \leq v_j \leq 0.2g_i + \xi$, where ξ is a constant
76 regulating the connectance of the networks, and varies uniformly in $[5 \times 10^{-3}, 10^{-1}]$. This model gives
77 fully interval networks that are close analogues to the niche model (Williams & Martinez, 2000), but has
78 the benefit of only relying on two features (g_i, v_j) , and having the exact same rule for all interactions. It is,
79 therefore, a simple case which most classifiers should be able to learn.

80 The training sample is composed of 30% of the 4×10^4 possible entries in the network, *i.e.* $n = 12000$. Out
81 of these interactions, we pick a proportion ν (the training set bias) to be positive, so that the training set
82 has νn interactions, and $(1 - \nu)n$ non-interactions. We vary ν uniformly in $]0, 1[$. This allows to evaluate
83 how the measures of binary classification performance respond to artificially rebalanced dataset for a
84 given network connectance. Note that both ξ and ν are sampled from a distribution rather than being
85 picked on a grid; this is because there is no direct relationship between the value of ξ and the connectance
86 of the simulated network, and therefore the precise value of ξ is not relevant for the analysis of the results.

87 The dataset used for numerical experiments is composed of 20000 such (ξ, ν) pairs, on which four learners
88 are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random forest
89 regressor. All models were taken from the `MLJ.jl` package (Blaom et al., 2020; Blaom & Vollmer, 2020) in
90 Julia 1.7 (Bezanson et al., 2017). In order to pick the best adjacency matrix for a given learner, we
91 performed a thresholding approach using 500 steps on predictions from the testing set, and picking the
92 threshold that maximized Youden's informedness, which is usually the optimized target for imbalanced
93 classification. During the thresholding step, we measured the area under the receiving-operator

characteristic (ROC-AUC) and precision-recall (PR-AUC) curves, as measures of overall performance over the range of returned values. We report the ROC-AUC and PR-AUC, as well as a suite of other measures as introduced in the next section, for the best threshold. The ensemble model was generated by summing the predictions of all component models on the testing set (ranged in $[0, 1]$), then put through the same thresholding process. The complete code to run the simulations is given as an appendix.

After the simulations were completed, we removed all runs (*i.e.* pairs of ξ and ν) for which at least one of the following conditions was met: the accuracy was 0, the true positive or true negative rates were 0, the connectance was larger than 0.2. This removes both the obviously failed model runs, and the networks that are more densely connected compared to the connectance of empirical food webs (and are therefore less difficult to predict, being less imbalanced).

Effect of training set bias on performance

Required amount of positives to get the best performance

Guidelines for prediction

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>

Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, 5(55), 2704. <https://doi.org/10.21105/joss.02704>

Blaom, A. D., & Vollmer, S. J. (2020, December 31). *Flexible model composition in machine learning and its implementation in MLJ*. <http://arxiv.org/abs/2012.15505>

118 Branco, P., Torgo, L., & Ribeiro, R. (2015, May 13). *A Survey of Predictive Modelling under Imbalanced*
 119 *Distributions*. <http://arxiv.org/abs/1505.01658>

120 Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable
 121 than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix
 122 evaluation. *BioData Mining*, 14, 13. <https://doi.org/10.1186/s13040-021-00244-z>

123 Somodi, I., Lepesi, N., & Botta-Dukát, Z. (2017). Prevalence dependence in model goodness measures with
 124 special emphasis on true skill statistics. *Ecology and Evolution*, 7(3), 863–872.
 125 <https://doi.org/10.1002/ece3.2654>

126 Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz,
 127 N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap
 128 towards predicting species interaction networks (across space and time). *Philosophical Transactions of*
 129 *the Royal Society B: Biological Sciences*, 376(1837), 20210063.
 130 <https://doi.org/10.1098/rstb.2021.0063>

131 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine
 132 learning in genomics. *Nature Reviews Genetics*, 1–13.
 133 <https://doi.org/10.1038/s41576-021-00434-9>

134 Williams, R., & Martinez, N. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.
 135 <http://userwww.sfsu.edu/>

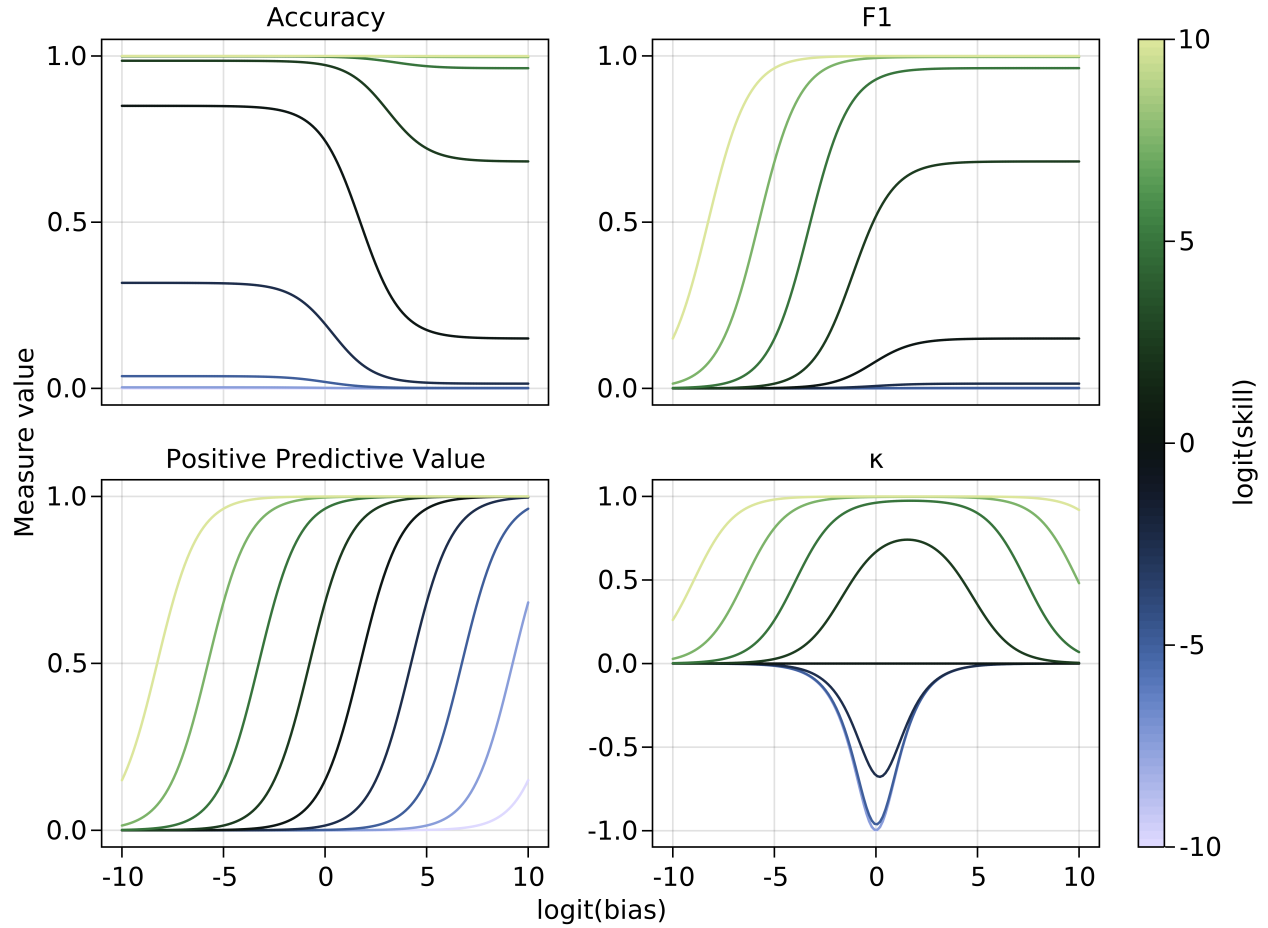


Figure 1: Consequences of changing the classifier skills (s) and bias (b) for a connectance $\rho = 0.15$, on accuracy, F_1 , positive predictive value, and κ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions. The F_1 score increases with skill but also increases when the bias tends towards estimating *more* interactions; PPV behaves in the same way. Interestingly, κ responds as expected to skill (being negative whenever $s < 0.5$), and peaks for values of $b \approx 0.5$; nevertheless, the value of bias for which κ is maximized is *not* $b = 0.5$, but instead increases with classifier skill. In other words, at equal skill, maximizing κ would lead to select a *more* biased classifier.