Combining Boosted Regression Trees and Hierarchical Species Occupancy Models

Rebecca A. Hutchinson and Thomas G. Dietterich
School of EECS, Oregon State University
rah@eecs.oregonstate.edu and tgd@eecs.oregonstate.edu

International Conference on
Computational Sustainability 2010

Abstract

We describe an extension to a standard model for the relationship between the occupancy pattern of a species on a landscape and imperfect observations of its presence or absence. The structure in the observation process is incorporated through a probabilistic model, and environmental inputs are incorporated using flexible tree-based methods. Our experiments on synthetic data compare the performance of the model with tree-based methods to the conventional parameterization which assumes a linear model for the covariates. Our results suggest that tree-based methods are helpful when the true relationship between the environmental inputs and the species occupancy does not fit the linear model.

Introduction

Goal: Predict the presence/absence Z_i of a bird species as a function of habitat features X_i at site i.

Problem #1: The training data does not observe Z_i directly. When an observer visits site i at time t, there is a detection probability d_{it} that depends on additional detection features W_{it}. The probability that the observer will report seeing the bird Y_{it} is d_{it} if $Z_i = 1$ (the bird is present) and 0 if $Z_i = 0$ (the bird is absent).

Solution #1: Fit a model of the form shown in Figure 1, that explicitly models separate contributions of the detection process (*d*) and the occupancy process (*o*). This model assumes that the true, latent occupancy Z is constant across the visits ($t = 1,\ldots,T$). This method has been developed in the ecology literature (e.g. [5]), where the standard methodology is to fit several models with different sets of covariates and choose one model according to a model selection criterion (e.g., AIC).

Problem #2: The sets of covariates X and W may be large, and we may not know which covariates are relevant and/or the correct model for their relationship to *o* and *d*.

Solution #2: Model the *o(X)* and *d(W)* using tree-based methods that can accommodate large numbers of covariates, interactions between covariates, and missing values in the covariates. One example of a tree-based method is boosted regression trees (BRTs, [4]), which have been successfully applied in ecology ([2]), but never as part of a model that accommodates uncertain detection.

Methods: Combine solutions 1 and 2 above. Learn tree-based models for occupancy and detection using functional gradient ascent ([3]).

References

[1] A.P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
[2] Jane Elith et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129-151, 2006.
[3] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189-1232, 2001.
[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Springer, 2nd edition*, 2009.
[5] D.I. MacKenzie et al. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence.* Academic Press, 2006.

This material is based upon work supported, in part, by the National Science Foundation under Grant NSF-0832804. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Model and Algorithms

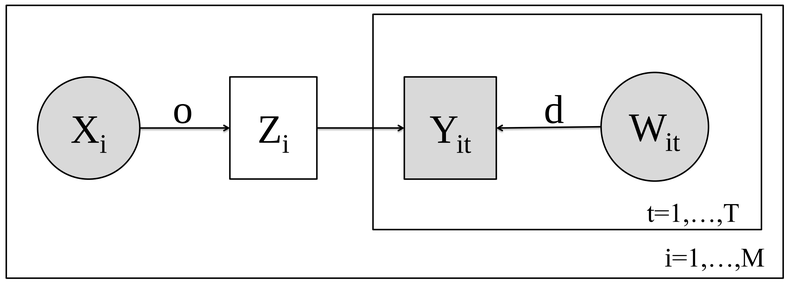

Figure 1: The graphical structure of the species occupancy model. The outer plate is repeated for each site *i*, and the inner plate is repeated for each visit *t*. *Y* is the observed presence or absence of the species, *Z* is the latent occupancy of the species, *W* contains detection covariates, and *X* contains occupancy covariates. Shaded nodes are observed; unshaded nodes are unobserved. Square nodes are binary; round nodes are continuous.

$$P(Z_i \mid X_i) = o(X_i) = o_i \qquad P(Y_{it} \mid W_{it}) = Z_i\, d(W_{it}) = d_{it}$$

$$l(\alpha, \beta | Y, X, W) = \sum_{i=1}^{M} [\log(o_i \prod_{t=1}^{T} [(d_{it})^{Y_{it}} (1 - d_{it})^{1 - Y_{it}}] + (1 - o_i) I(\sum_{t=1}^{T} Y_{it} == 0))]$$

In the log-likelihood function *l*, *I(x)* returns 1 if and only if *x* is true. α and β are the parameters of the tree methods (the split points and leaf values).

In functional gradient ascent, we train a tree at each stage to predict the following values:

$$\text{Occupancy:} \quad \frac{\delta l}{\delta o_j} = \frac{(\prod_{t=1}^{T} [d_{jt}^{y_{jt}} (1 - d_{jt})^{1 - y_{jt}}] - I(\sum_{t=1}^{T} y_{jt} == 0)) o_j (1 - o_j)}{o_j \prod_{t=1}^{T} [d_{jt}^{y_{jt}} (1 - d_{jt})^{1 - y_{jt}}] - (1 - o_j) I(\sum_{t=1}^{T} y_{jt} == 0)}$$

$$\text{Detection:} \quad \frac{\delta l}{\delta d_{js}} = \frac{\prod_{t=1}^{T} [d_{jt}^{y_{jt}} (1 - d_{jt})^{1 - y_{jt}}] o_j (y_{js} - d_{js})}{o_j \prod_{t=1}^{T} [d_{jt}^{y_{jt}} (1 - d_{jt})^{1 - y_{jt}}] - (1 - o_j) I(\sum_{t=1}^{T} y_{jt} == 0)}$$

The tree grown at each stage is weighted, and we iterate between updating the occupancy and detection models.

Alternatively, we can use Expectation-Maximization ([1]) to maximize the expected joint log-likelihood:

$$Q = \sum_{i=1}^{M} [P(Z_i = 1 | Y_{i\cdot}, X_i, W_{i\cdot})(\log(o_i) + \sum_{t=1}^{T} (Y_i \log(d_{it}) + (1 - Y_{it}) \log(1 - d_{it}))) + P(Z_i = 0 | Y_{i\cdot}, X_i, W_{i\cdot})(\log(1 - o_i) + \sum_{t=1}^{T} Y_{it}(-\infty))]$$

In the last line, we let the product of 0 and $-\infty$ be 0, so that the last term does not contribute to the likelihood if Y_{it} is 0 and causes the likelihood to be $-\infty$ if Y_{it} is 1. This is consistent with the assumption that if the site is truly unoccupied, the species will never be observed (i.e., no false positive observations), which is reasonable for expert observers.

The EM algorithm iterates between computing *P(Z|Y,X,W;α,β)* via Bayes rule in the E-step and using functional gradient ascent to update the trees in the M-step, based on derivatives of *Q* (not shown).

Experiments

Synthetic Data:
- 750 sites (M=250), each visited twice (T=2). 250 for training; 250 for validation, 250 for testing.
- 2 occupancy features and 2 detection features drawn from a standard normal distribution.
- Z_i and Y_{it} values sampled according to the generative model in Figure 1. The true values of Z were saved for evaluation purposes, even though Z is not available in real data.

Algorithms:
- LR has linear functions for *o* and *d*, trained on objective function *l*.
- LR+EM has linear functions for *o* and *d*, trained on objective function *Q*.
- BRT has tree ensembles for *o* and *d*, trained on objective function *l*.
- BRT+EM has tree ensembles for *o* and *d*, trained on objective function *Q*.

The validation set was used to select the regularization scheme for the LR models (L1 vs. L2 and a weight for each model component from {0,0.001, 0.01,0.1,0.5,1}). Validation set log-likelihood was also used to select the number of stages (trees) for BRT (from [1,1000]), the depth of the trees (from {1, 2, 3, 5}), and the shrinkage parameter (from {0.001, 0.01, 0.1}).

Evaluation:
We compare the methods based on test set log-likelihood, and the area under the ROC curve (AUC) on 2 different prediction tasks.
Synthetic data: *o* and *d* are linear functions of the covariates.

Method	Tuning parameters	Test set log-likelihood	AUC predicting Z	AUC predicting Y
True model		*-219*	*0.80*	*0.93*
LR	L2, wts (1, 0.1)	**-216**	**0.822**	**0.792**
LR+EM	L2, wts (1, 0.5)	**-217**	**0.822**	**0.792**
BRT	Depth 2, shrinkage 0.1, 1000 stages	-227	0.802	0.769
BRT+EM	Depth 2, shrinkage 0.1, 1000 stages	-264	0.766	0.729

Synthetic data: *o* and *d* are the XOR of the sign of the covariates.

Method	Tuning parameters	Test set log-likelihood	AUC predicting Z	AUC predicting Y
True model		*-154*	*0.812*	*0.922*
LR	L1, wts (1, 1)	-320	0.493	0.551
LR+EM	L1, wts (1, 1)	-317	0.488	0.533
BRT	Depth 5, shrinkage 0.1, 1000 stages	**-233**	**0.777**	**0.745**
BRT+EM	Depth 5, shrinkage 0.1, 15 stages	-249	0.773	0.717

Conclusions:
When the LR models can represent the true function generating the data, they have a slight edge over the BRT models. When the true function cannot be represented by the LR models, the BRT models perform better.

Questions for future experiments to explore:
(1) In real data, we do not observe Z so we cannot validate our models on the task of interest. How should we set tuning parameters and use validation data to do well at predicting Z using only the observed variables?
(2) How does the answer to (1) change with more covariates? Arbitrarily complex functions for *o* and *d*? Irrelevant covariates?
(3) Is there a systematic advantage/disadvantage to using EM?