Proceedings of the Twenty-First International Conference on Machine Learning, pages 655-662, 2004.

A Maximum Entropy Approach to Species Distribution Modeling

Steven J. Phillips PHILLIPS@RESEARCH.ATT.COM
AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932

Miroslav Dudík MDUDIK@CS.PRINCETON.EDU
Robert E. Schapire SCHAPIRE@CS.PRINCETON.EDU
Princeton University, Department of Computer Science, 35 Olden Street, Princeton, NJ 08544

Abstract

We study the problem of modeling species geographic distributions, a critical problem in conservation biology. We propose the use of maximum-entropy techniques for this problem, specifically, sequential-update algorithms that can handle a very large number of features. We describe experiments comparing maxent with a standard distribution-modeling tool, called GARP, on a dataset containing observation data for North American breeding birds. We also study how well maxent performs as a function of the number of training examples and training time, analyze the use of regularization to avoid overfitting when the number of examples is small, and explore the interpretability of models constructed using maxent.

1. Introduction

We study the problem of modeling the geographic distribution of a given animal or plant species. This is a critical problem in conservation biology: to save a threatened species, one first needs to know where the species prefers to live, and what its requirements are for survival, i.e., its ecological niche (Hutchinson, 1957).

The data available for this problem typically consists of a list of georeferenced occurrence localities, i.e., a set of geographic coordinates where the species has been observed. In addition, there is data on a number of environmental variables, such as average temperature, average rainfall, elevation, etc., which have been measured or estimated across a geographic region of interest. The goal is to predict which areas within the region satisfy the requirements of the species' ecological niche, and thus form part of the species' *potential distribution* (Anderson & Martínez-Meyer, 2004). The potential distribution describes where conditions are suitable for survival of the species, and is thus of great importance for conservation. It can also be used to estimate the species' *realized distribution*, for example by removing areas where the species is known to be absent because of deforestation or other habitat destruction. Although a species' realized distribution may exhibit some spatial correlation, the potential distribution does not, so considering spatial correlation is not necessarily desirable during species distribution modeling.

It is often the case that only *presence* data is available indicating the occurrence of the species. Natural history museum and herbarium collections constitute the richest source of occurrence localities (Ponder et al., 2001; Stockwell & Peterson, 2002). Their collections typically have no information about the *failure* to observe the species at any given location; in addition, many locations have not been surveyed. In the lingo of machine learning, this means that we have only positive examples and no negative examples from which to learn. Moreover, the number of sightings (training examples) will often be very small by machine-learning standards, say a hundred or less. Thus, the first contribution of this paper is the introduction of a scientifically important problem as a challenging domain for study by the machine learning community.

To address this problem, we propose the application of maximum-entropy (maxent) techniques which have been so effective in other domains, such as natural language processing (Berger et al., 1996). Briefly, in maxent, one is given a set of samples from a distribution over some space, as well as a set of features (real-valued functions) on this space. The idea of maxent is to estimate the target distribution by finding the distribution of maximum entropy (i.e., that is closest to uniform) subject to the constraint that the expected value of each feature under this estimated distribution matches its empirical average. This turns out to be equivalent, under convex duality, to finding the maximum likelihood Gibbs distribution (i.e., distribution that is exponential in a linear combination of the features). For species distribution modeling, the occurrence localities of the species serve as the sample points, the geographical region of interest is the space on which this distribution is defined, and the features are the environmental variables (or functions thereof). See Figure 1 for an example.

In Section 2, we describe the basics of maxent in greater detail. Iterative scaling and its variants (Darroch & Ratcliff, 1972; Della Pietra et al., 1997) are standard algorithms for computing the maximum entropy distribution. We use our own variant which iteratively updates the weights on

Appearing in *Proceedings of the 21^{st} International Conference on Machine Learning*, Banff, Canada, 2004.

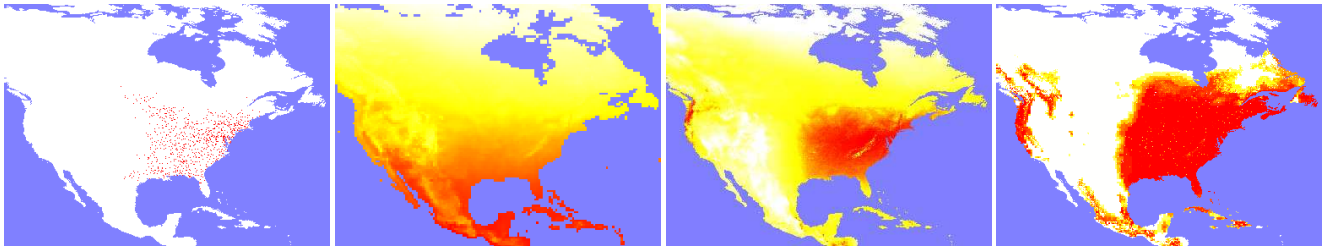

Figure 1. Left to right: Yellow-throated Vireo training localities from the first random partition, an example environmental variable (annual average temperature, higher values in red), maxent prediction using linear, quadratic and product features, and GARP prediction. Prediction strength is shown as white (weakest) to red (strongest); reds could be interpreted as suitable conditions for the species.

features sequentially (one by one) rather than in parallel (all at once), along the lines of Collins, Schapire and Singer (2002). This sequential approach is analogous to AdaBoost which modifies the weight of a single "feature" (usually called a base or weak classifier in that context) on each round. As in boosting, this approach allows us to use very large feature spaces.

One would intuitively expect an oversize feature space to be a problem for generalization since it increases the possibility of overfitting, leading others to use feature selection for maxent (Berger et al., 1996). We instead use a regularization approach, introduced in a companion theoretical paper (Dudík et al., 2004), which allows one to prove bounds on the performance of maxent using finite data, even when the number of features is very large or even uncountably infinite. Here we investigate in detail the practical efficacy of the technique for species distribution modeling. We also describe a numerical acceleration method that speeds up learning.

In Section 3, we describe an extensive set of experiments we conducted comparing maxent to a widely used existing distribution modeling algorithm; results of the experiments are described in Section 4. Quite a number of approaches have been suggested for species distribution modeling including neural nets, genetic algorithms, generalized linear models, generalized additive models, bioclimatic envelopes and more; see Elith (2002) for a comparison. From these, we selected the Genetic Algorithm for Ruleset Prediction (GARP) (Stockwell & Noble, 1992; Stockwell & Peters, 1999), because it has seen widespread recent use to study diverse topics such as global warming (Thomas et al., 2004), infectious diseases (Peterson & Shaw, 2003) and invasive species (Peterson & Robins, 2003); many further applications are cited in these references. GARP was also selected because it is one of the few methods available that does not require absence data (negative examples).

We compare GARP and maxent using data derived from the North American Breeding Bird Survey (BBS) (Sauer et al., 2001), an extensive dataset consisting of thousands of occurrence localities for North American birds and used previously for species distribution modeling, in particular for evaluating GARP (Peterson, 2001). The comparison suggests that maxent methods hold great promise for species distribution modeling, often achieving substantially superior performance in controlled experiments relative to GARP. In addition to comparisons with GARP, we performed experiments testing: (1) the performance of maxent as a function of the number of sample points available, so as to determine the all important question of how much data is enough; (2) the effectiveness of regularization to avoid overfitting on small sample sizes; and (3) the effectiveness of our numerical acceleration methods.

Lastly, it is desirable for a species distribution model to allow interpretation to deduce the most important limiting factors for the species. A noted limitation of GARP is the difficulty of interpreting its models (Elith, 2002). We show how the models generated by maxent can be put into a form that is easily understandable and interpretable by humans.

2. The Maximum Entropy Approach

In this section, we describe our approach to modeling species distributions. As explained above, we are given a space X representing some geographic region of interest. Typically, X is a set of discrete grid cells; here we only assume that X is finite. We also are given a set of points $x_1, \ldots, x_m$ in X, each representing a locality where the species has been observed and recorded. Finally, we are provided with a set of environmental variables defined on X, such as precipitation, elevation, etc.

Given these ingredients, our goal is to estimate the range of the given species. In this paper, we formalize this rather vague goal within a probabilistic framework. Although this will inevitably involve simplifying assumptions, what we gain will be a language for defining the problem with mathematical precision as well as a sensible approach for applying machine learning.

Unlike others who have studied this problem, we adopt the view that the localities $x_1, \ldots, x_m$ were selected independently from X according to some unknown probability distribution π, and that our goal is to estimate π. At the foundation of our approach is the premise that the distribution π (or a thresholded version of it) coincides with the biologists' concept of the species' potential distribution. Superficially, this is not unreasonable, although it does ignore the fact that some localities are more likely to have been visited than others. The distribution π may therefore exhibit sampling bias, and will be weighted towards areas and environmental conditions that have been better sampled, for example because they are more accessible.

That being said, the problem becomes one of *density estimation*: given $x_1, \ldots, x_m$ chosen independently from some unknown distribution π, we must construct a distribution $\hat{\pi}$ that approximates π.

Common name	Abbreviation	# examples
Gray Vireo	GV	78
Hutton's Vireo	HV	198
Plumbeous Vireo	PV	256
Philadelphia Vireo	PhV	325
Bell's Vireo	BV	419
Cassin's Vireo	CV	424
Blue-headed Vireo	BhV	973
White-eyed Vireo	WeV	1271
Yellow-throated Vireo	YV	1611
Loggerhead Shrike	LS	1850
Warbling Vireo	WV	2526
Red-eyed Vireo	RV	2773

Table 1. Studied species, with number of presence records

In constructing $\hat{\pi}$, we also make use of a given set of *features* $f_1, \ldots, f_n$ where $f_j : X \to \mathbb{R}$. These features might consist of the raw environmental variables, or they might be higher level features derived from them (see Section 3.3). Let $\boldsymbol{f}$ denote the vector of all n features.

For any function $f : X \to \mathbb{R}$, let $\pi[f]$ denote its expectation under π. Let $\tilde{\pi}$ denote the *empirical distribution*, i.e., $\tilde{\pi}(x) = |\{1 \le i \le m : x_i = x\}|/m$. In general, $\tilde{\pi}$ may be quite distant, under any reasonable measure, from π. On the other hand, for a given function f, we do expect $\tilde{\pi}[f]$, the empirical average of f, to be rather close to its true expectation $\pi[f]$. It is natural, therefore, to seek an approximation $\hat{\pi}$ under which f_j's expectation is equal (or at least very close) to $\tilde{\pi}[f_j]$ for every f_j. There will typically be many distributions satisfying these constraints. The *maximum entropy principle* suggests that, from among all distributions satisfying these constraints, we choose the one of maximum entropy, i.e., the one that is closest to uniform. Here, as usual, the entropy of a distribution p on X is defined to be $\mathrm{H}(p) = -\sum_{x \in X} p(x) \ln p(x)$.

Thus, the idea is to estimate π by the distribution $\hat{\pi}$ of maximum entropy subject to the condition that $\hat{\pi}[f_j] = \tilde{\pi}[f_j]$ for all features f_j. Alternatively, we can consider all *Gibbs distributions* of the form $q_{\boldsymbol{\lambda}}(x) = e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)}/Z_{\boldsymbol{\lambda}}$ where $Z_{\boldsymbol{\lambda}} = \sum_{x \in X} e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)}$ is a normalizing constant, and $\boldsymbol{\lambda} \in \mathbb{R}^n$. Then, following Della Pietra, Della Pietra and Lafferty (1997), it can be proved that the maxent distribution described above is the same as the maximum likelihood Gibbs distribution, i.e., the distribution $q_{\boldsymbol{\lambda}}$ that minimizes $\mathrm{RE}(\tilde{\pi} \parallel q_{\boldsymbol{\lambda}})$ where $\mathrm{RE}(p \parallel q) = \sum_{x \in X} p(x) \ln(p(x)/q(x))$ denotes *relative entropy* or *Kullback-Leibler divergence*. Note that the negative log likelihood $\tilde{\pi}[-\ln(q_{\boldsymbol{\lambda}})]$ (also called log loss) only differs from $\mathrm{RE}(\tilde{\pi} \parallel q_{\boldsymbol{\lambda}})$ by the constant $\mathrm{H}(\tilde{\pi})$; we therefore use the two interchangeably as objective functions.

2.1. A sequential-update algorithm

There are a number of algorithms for finding the maxent distribution, especially iterative scaling and its variants (Darroch & Ratcliff, 1972; Della Pietra et al., 1997) as well as the gradient and second-order descent methods (Malouf, 2002; Salakhutdinov et al., 2003). In this paper, we used a sequential-update algorithm that modifies one weight λ_j at a time, as explored by Collins, Schapire and Singer (2002) in a similar setting. We chose this coordinate-wise descent procedure since it is easily applicable when the number of features is very large (or infinite).

Specifically, our very simple algorithm works as follows. Assume without loss of generality that each feature f_j is bounded in $[0, 1]$. On each of a sequence of rounds, we choose the feature f_j to update for which $\mathrm{RE}(\tilde{\pi}[f_j] \parallel q_{\boldsymbol{\lambda}}[f_j])$ is maximized, where $\boldsymbol{\lambda}$ is the current weight vector (and where $\mathrm{RE}(p \parallel q)$, for $p, q \in \mathbb{R}$, is binary relative entropy). We next update $\lambda_j \leftarrow \lambda_j + \alpha$ where

$$\alpha = \ln\left(\frac{\tilde{\pi}[f_j](1 - q_{\boldsymbol{\lambda}}[f_j])}{(1 - \tilde{\pi}[f_j])q_{\boldsymbol{\lambda}}[f_j]}\right). \tag{1}$$

The output distribution $\hat{\pi}$ is the one defined by the computed weights, i.e., $q_{\boldsymbol{\lambda}}$. Essentially, this algorithm works by altering one weight λ_j at a time so as to greedily maximize the likelihood (or an approximation thereof). This procedure is guaranteed to converge to the optimal maximum entropy distribution. The derivation of this algorithm, along with its proof of convergence are given in a companion paper (Dudík et al., 2004) and are based on techniques explained by Della Pietra, Della Pietra and Lafferty (1997) as well as Collins, Schapire and Singer (2002).

To accelerate convergence, we do a line search in each iteration: evaluate the log loss when λ_j is incremented by $2^i\alpha$ for $i = 0, 1, \ldots$ in turn, and choose the last i before the log loss decreases. This is similar to line search methods described in (Minka, 2001).

2.2. Regularization

The basic approach described above computes the maximum entropy distribution $\hat{\pi}$ for which $\hat{\pi}[f_j] = \tilde{\pi}[f_j]$. However, we do not expect $\tilde{\pi}[f_j]$ to be *equal* to $\pi[f_j]$ but only close to it. Therefore, in keeping with our motivation, we can soften these constraints to have the form

$$|\hat{\pi}[f_j] - \tilde{\pi}[f_j]| \le \beta_j \tag{2}$$

where β_j is an estimate of how close $\tilde{\pi}[f_j]$, being an empirical average, must be to its true expectation $\pi[f_j]$. Maximizing entropy subject to Eq. (2) turns out to be equivalent to finding the Gibbs distribution $\hat{\pi} = q_{\boldsymbol{\lambda}}$ which minimizes

$$\mathrm{RE}(\tilde{\pi} \parallel q_{\boldsymbol{\lambda}}) + \textstyle\sum_j \beta_j |\lambda_j|. \tag{3}$$

In other words, this approach is equivalent to maximizing the likelihood of the sought after Gibbs distribution with (weighted) ℓ_1-regularization. This form of regularization also makes sense because the number of training examples needed to approximate the "best" Gibbs distribution can be bounded when the ℓ_1-norm of the weight vector $\boldsymbol{\lambda}$ is bounded. (See (Dudík et al., 2004) for details.) In a Bayesian framework, Eq. (3) corresponds to a negative log posterior given a Laplace prior. Other priors studied for maxent are Gaussian (Chen & Rosenfeld, 2000) and exponential (Goodman, 2003). Laplace priors have been studied in the context of neural networks by Williams (1995).

The regularized formulation can be solved using a simple modification of the above algorithm. On each round, a feature f_j and value α are chosen so as to maximize the change in (an approximation of) the regularized objective function in Eq. (3). This works out to be

$$-\alpha\tilde{\pi}[f_j] + \ln(1 + (e^{\alpha} - 1)q_{\boldsymbol{\lambda}}[f_j]) + \beta_j(|\lambda_j + \alpha| - |\lambda_j|).$$

	GARP threshold = 1						GARP threshold = 10					
Bird	Area	L	LQ	LQP	T	GARP	Area	L	LQ	LQP	T	GARP
GV	0.307	0.000	0.000	0.000	0.003	0.000	0.144	0.046	0.079	0.018	0.079	0.085
HV	0.595	0.028	0.003	0.004	0.000	0.000	0.314	0.139	0.019	0.030	0.015	0.034
PV	0.428	0.004	0.005	0.002	0.006	0.003	0.149	0.063	0.030	0.036	0.027	0.067
PhV	0.545	0.096	0.000	0.000	0.004	0.000	0.199	0.423	0.036	0.034	0.055	0.069
BV	0.668	0.000	0.000	0.000	0.000	0.000	0.301	0.036	0.010	0.004	0.012	0.048
CV	0.430	0.060	0.018	0.008	0.015	0.067	0.225	0.242	0.123	0.092	0.088	0.149
BhV	0.563	0.060	0.006	0.005	0.008	0.009	0.226	0.336	0.122	0.103	0.086	0.110
WeV	0.433	0.008	0.000	0.001	0.001	0.001	0.141	0.216	0.045	0.036	0.029	0.067
YV	0.472	0.008	0.000	0.000	0.000	0.005	0.201	0.306	0.049	0.043	0.040	0.086
LS	0.724	0.005	0.000	0.000	0.001	0.000	0.356	0.135	0.080	0.063	0.071	0.112
WV	0.780	0.013	0.000	0.000	0.001	0.003	0.437	0.355	0.053	0.046	0.049	0.121
RV	0.667	0.057	0.003	0.001	0.003	0.006	0.326	0.250	0.104	0.084	0.074	0.109
Avg	0.551	0.028	0.003	0.002	0.004	0.008	0.252	0.212	0.063	0.049	0.052	0.088

Table 2. Omission rates in the equalized area test for GARP threshold of 1 (left) and 10 (right). "Area" column is area of species' potential distribution, as produced by GARP; other predictions are thresholded to give the same predicted area. The predictions analyzed are: maxent with linear (L); linear and quadratic (LQ); linear, quadratic and product (LQP); and threshold (T) features; and GARP.

Bird	L	LQ	LQP	T	GARP
GV	0.946	0.962	0.973	0.959	0.919
HV	0.870	0.957	0.955	0.963	0.835
PV	0.940	0.952	0.955	0.951	0.916
PhV	0.775	0.937	0.941	0.934	0.888
BV	0.857	0.932	0.936	0.937	0.840
CV	0.846	0.916	0.929	0.924	0.831
BhV	0.789	0.910	0.916	0.919	0.862
WeV	0.897	0.942	0.945	0.947	0.920
YV	0.849	0.925	0.928	0.929	0.882
LS	0.789	0.837	0.850	0.847	0.794
WV	0.644	0.836	0.840	0.840	0.742
RV	0.761	0.858	0.865	0.869	0.805
Avg	0.854	0.910	0.914	0.919	0.862

Table 3. AUC values averaged over 10 random partitions of occurrence localities. Predictions analyzed are as in Table 2.

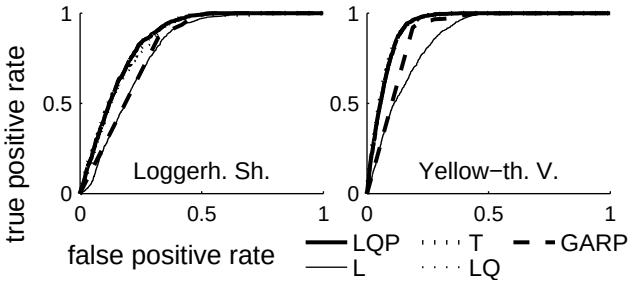

Figure 2. ROC curves for the first random partition of occurrence localities of the Loggerhead Shrike and the Yellow-throated Vireo. In both cases, maxent with linear features is the lowest curve, GARP is the second lowest, and the remaining three are very close together. Portions of LQ and T curves are obscured by the LQP curve.

The maximizing α, must be either $-\lambda_j$ or Eq. (1) with $\tilde{\pi}[f_j]$ replaced by $\tilde{\pi}[f_j] - \beta_j$ (provided $\lambda_j + \alpha \geq 0$) or $\tilde{\pi}[f_j] + \beta_j$ (provided $\lambda_j + \alpha \leq 0$). Thus, the best α (for a given f_j) can be computed by trying all three possibilities. Once f_j and α have been selected, we only need update $\lambda_j \leftarrow \lambda_j + \alpha$. As before, this algorithm can be proved to converge to a solution to the problem described above.

Throughout our study we reduced the β_j to a single regularization parameter β as follows. We expect $|\pi[f_j] - \tilde{\pi}[f_j]| \approx \sigma[f_j]/\sqrt{m}$, where $\sigma[f_j]$ is the standard deviation of f_j under π. We therefore approximated $\sigma[f_j]$ by the sample deviation $\tilde{\sigma}[f_j]$ and used $\beta_j = \beta\tilde{\sigma}[f_j]/\sqrt{m}$.

3. Experimental Methods

3.1. The Breeding Bird Survey

The North American Breeding Bird Survey (Sauer et al., 2001) is a data set with a large amount of high-quality location data. It is good for a first evaluation of maxent for species distribution modeling, as the generous quantities of data allow for detailed experiments and statistical analyses. It has also been used to demonstrate the utility of GARP (Peterson, 2001). Roadside surveys are conducted on standard routes during the peak of the nesting season. Each route consists of fifty stops located at 0.5 mile intervals. A three-minute count is conducted at each stop, during which the observer records all birds heard or seen within 0.25 mile of the stop. Data from all fifty stops are combined to obtain the set of species observed on the route. There are 4161 routes within the region covered by the environmental coverages described below.

3.2. Environmental Variables

The environmental variables (coverages) use a North American grid with 0.2 degree square cells, and are all included with the GARP distribution, available at http://www.lifemapper.org/desktopgarp. Some coverages are derived from weather station readings during the period 1961 to 1990 (New et al., 1999). Out of these we use annual precipitation, number of wet days, average daily temperature and temperature range. The remaining coverages are derived from a digital elevation model for North America, and consist of elevation, aspect and slope. Each coverage is defined over a 386 $\times$ 286 grid, of which 58,065 points have data for all coverages.

3.3. Experimental Design

We chose 12 out of the 421 species included in the Breeding Bird Survey to model, and considered a route to be an occurrence locality for a species if it had a presence record for any year of the survey. The chosen species and the number of routes where each has occurrence localities

are shown in Table 1. The occurrence data was divided into ten random partitions: in each partition, 50% of the occurrence localities were randomly selected for the training set, while the remaining 50% were set aside for testing.

We chose four feature types for maxent to use: the raw environmental variables (*linear* features); squares of environmental variables (*quadratic* features); products of pairs of environmental variables (*product* features); and binary features derived by thresholding environmental variables (*threshold* features). The latter features are equal to one if an environmental variable is above some threshold, and zero otherwise. Use of linear features constrains the mean of the environmental variables in the maxent distribution, linear plus quadratic features constrain the variance, while linear plus quadratic plus product features constrain the covariance of pairs of environmental variables.

On each training set, we ran maxent with four different subsets of the feature types: linear (L); linear and quadratic (LQ); linear, quadratic and product (LQP); and threshold (T). We also ran GARP on each training set.

The output of GARP and maxent have quite different interpretations. Nevertheless, each can be used to (partially) rank all locations according to their habitability. To compare these rankings, we used receiver operating characteristic (ROC) curves. For each of the runs, we calculated the AUC (area under the ROC curve), and determined the average AUC over the ten occurrence data partitions. See Section 3.5 for further discussion of this metric.

The AUC comparison is somewhat biased in maxent's favor, as a continuous prediction will typically have a higher AUC than a discrete prediction. We therefore do a second comparison, where we select operating thresholds for GARP that have been widely used in practice, and compare the algorithms only at those operating points. We call this an "equalized area test", and the details are as follows. We applied two thresholds to each GARP prediction, namely 1 and 10, corresponding to at least one, or all, best-subset models predicting presence (see Section 3.4 for GARP details). These are the most-often used GARP thresholds (Anderson & Martínez-Meyer, 2004). For each of the two resulting predictions, we set thresholds for the maxent models that result in prediction of the same area (geographic extent) as GARP. The predictions, now binary and with the same predicted area, are then simply compared using omission rates (fraction of test localities not predicted present). Again, averages were taken over the 10 random partitions of the occurrence data.

Most applications of species distribution modeling have much less data available than for North American birds. Indeed, species of conservation importance may have extremely few georeferenced locality records, often fewer than 10. To investigate the use of maxent in such limited data settings, we perform experiments using limited subsets of the Breeding Bird data. We selected increasing subsets of training data in each partition, ran all four versions of maxent, and took an average AUC over ten partitions.

In order to determine sensitivity of maxent to the value of β and its interaction with sample size, we varied β and the number of training examples and took an average AUC over ten partitions for all four versions of maxent. Lastly, to measure the effect of our acceleration method, we performed runs using the first random partition for the Loggerhead Shrike and the Yellow-throated Vireo, both with and without line search for α (as described in Section 2), and measured the log loss on both training and test data as a function of running time.

3.4. Algorithm implementations

For the maxent runs, we ran the iterative algorithm described in Section 2 for 500 rounds, or until the change in the objective function on a single round fell below 10^{-5}. For the regularization parameter β, to avoid overfitting the test data, we used the same setting of 0.1 for all feature types, except threshold features for which we used 1.0. In Section 4.4, we describe experiments showing how sensitive our results are to the choice of β.

To reduce the variability inherent in GARP's random search procedure, we made composite GARP predictions using the "best-subsets" procedure (Anderson et al., 2003), as was done in recent applications (Peterson et al., 2003; Raxworthy et al., 2004). We generated 100 binary models, using GARP version 1.1.3 with default parameter values, then eliminated models with more than 5% intrinsic omission (negative prediction of training localities). If at most 10 models remained, they then constituted the best subset; otherwise, we selected the 10 models whose predicted area was closest to the median of the remaining models. The composite prediction gives the number of best-subset models in which each point is predicted suitable (0-10). For Cassin's Vireo, the best subset was empty for most random partitions of occurrence localities, so we increased the intrinsic omission threshold to 10% for that species.

3.5. ROC curves

An ROC curve shows the performance of a classifier whose output depends on a threshold parameter. It plots true positive rate against false positive rate for each threshold. A point (x, y) indicates that for some threshold, the classifier classifies a fraction x of negative examples as positive, and a fraction y of positive examples as positive. The curve is obtained by "joining the dots".

The area under an ROC curve (AUC) has a natural statistical interpretation. Pick a random positive example and a random negative example. The area under the curve is the probability that the classifier correctly orders the two points (with random ordering in the case of ties). A perfect classifier therefore has an AUC of 1. However, to use ROC curves with presence-only data, we must interpret as "negative examples" all grid cells with no occurrence localities, even if they support good environmental conditions for the species. The maximum AUC is therefore less than one (Wiley et al., 2003), and is smaller for wider-ranging species.

4. Results

4.1. Equalized Area Test

The results of the equalized area test are in Table 2. With a threshold of 1, GARP predicts large areas as having suitable conditions for the species, and all algorithms have

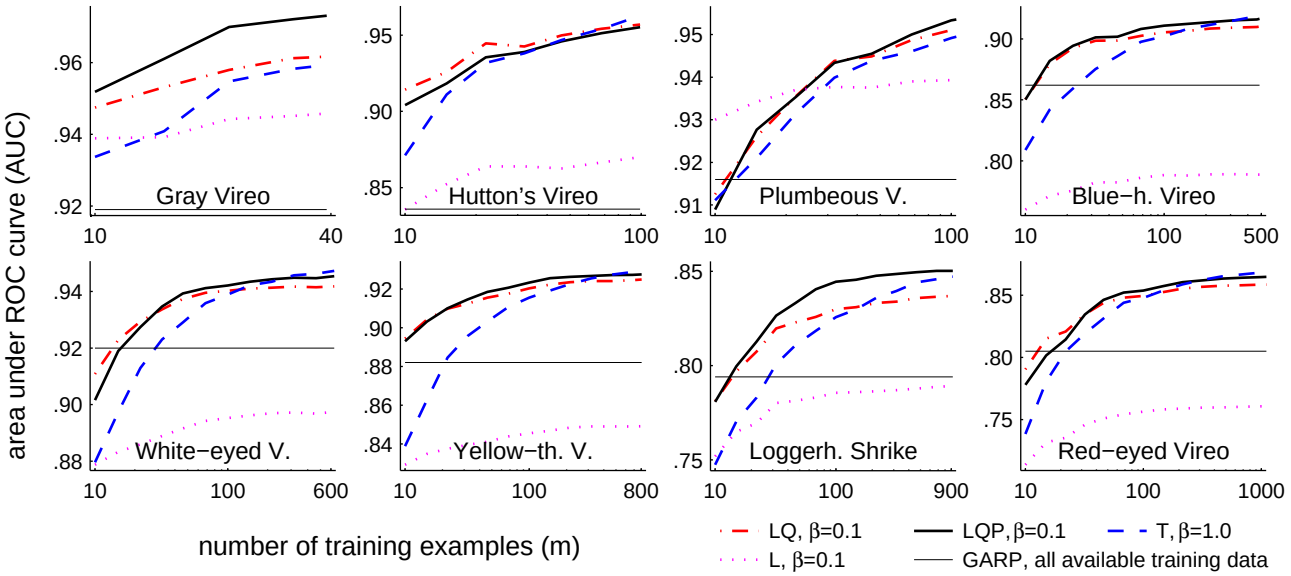

Figure 3. Learning curves. AUC averaged over 10 partitions for four versions of maxent (L, LQ, LQP and T) as a function of the number of training examples. Numbers of training examples are plotted on a logarithmic scale. We also include the average AUC for GARP on all training examples. Curves for the remaining species look qualitatively similar.

very low average omission (with the exception of GARP on Cassin's Vireo). A threshold of 10 causes less overprediction, and reveals more differences between the algorithms. The best results are obtained by maxent with two of the feature sets (LQP and T). These two are superior to GARP on all species, often very substantially; LQ is superior to GARP for all species but BhV.

4.2. ROC analysis

Table 3 shows the AUC for each species, averaged over the 10 random partitions of the occurrence localities. Example ROC curves used in computing the averages can be seen in Figure 2, which shows the performance of the algorithms on the first random partition for the Loggerhead Shrike and Yellow-throated Vireo.

The AUC for maxent improves dramatically going from linear (L) to linear plus quadratic (LQ) features, with a small further improvement when product features are added (LQP). The AUC for threshold features (T) is similar to LQP. For all species, the AUC for GARP is lower than for all maxent feature sets except sometimes L. Note that GARP is disadvantaged in AUC comparisons by not distinguishing between points in its highest rank (those points predicted present in all best-subset models), as can be seen in Figure 2, where GARP loses area at the left end of the ROC curve. Nevertheless, wherever GARP has data points, maxent with the better feature sets is quite consistently as good as or better than GARP.

4.3. Learning Curve Experiments

Figure 3 shows the AUC averaged over 10 partitions for an increasing number of training examples on eight of the species. We also include GARP results for full training sets as a base line. As expected, models with a larger number of features tend to overfit small training sets, but they give more accurate predictions for large training sets. Linear models do not capture species distribution very well and are included only for completeness. With the exception of the Plumbeous Vireo, three remaining versions of maxent outperform L models already for the smallest training sets. LQP models become better than LQ for 30-40 training examples; their performance, however, matches that of LQ already for smaller training sets. T models perform worse than both LQ and LQP for small training sets, but they slightly outperform LQP once training sets reach 400 examples. Learning curves for species with large numbers of examples indicate that for both LQ and LQP about 50-100 examples suffice for a prediction that is close to optimal for those models.

4.4. Sensitivity to Regularization

Figure 4 shows the sensitivity of maxent to the regularization value β for LQP and T versions of maxent. Due to the lack of space we do not present results for L and LQ versions, and give sensitivity curves for only four species. Curves for the remaining species look qualitatively similar. Note the remarkably consistent peak at $\beta \approx 1.0$ for threshold feature curves; theoretical reasons for this phenomenon require further investigation. For LQP runs, peaks are much less pronounced and do not appear at the same value of β across different species. Benefits of regularization in LQP runs diminish as the number of training examples increases (this is even more so for LQ and L runs, not presented here). This is because the relatively small number of features (compared with threshold features) naturally prevents overfitting large training sets.

4.5. Feature Profiles

Maxent as we have described it returns a vector $\boldsymbol{\lambda}$ that characterizes the Gibbs distribution $q_{\boldsymbol{\lambda}}(x) = e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(x)}/Z_{\boldsymbol{\lambda}}$ minimizing the (regularized) log loss. When each feature is derived from one environmental variable then the linear

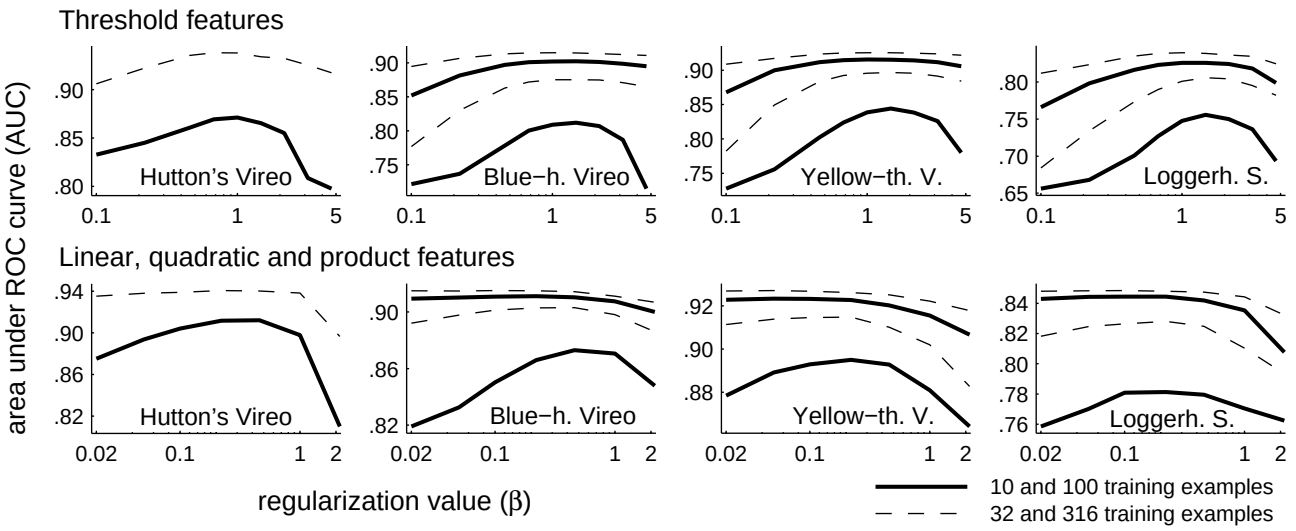

Figure 4. Sensitivity of maxent to regularization. AUC averaged over 10 partitions as a function of β for a varying number of training examples. For a fixed value of β, maxent finds better solutions (with a larger AUC) as the number of examples grows. We ran maxent with 10, 32, 100 and 316 training examples. Curves from bottom up correspond to these numbers; curves for higher numbers are missing where fewer training examples were available. Values of β are of the form $10^{i/6}$ and are plotted on a logarithmic scale.

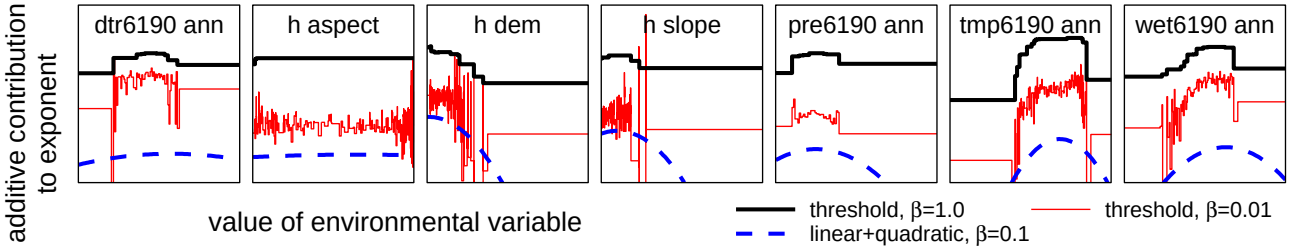

Figure 5. Feature profiles learned on the first partition of the Yellow-throated Vireo. For every environmental variable, its additive contribution to the exponent of the Gibbs distribution is given as a function of its value. This contribution is the sum of features derived from that variable weighted by the corresponding lambdas. Profiles for three types of maxent runs have been shifted for clarity — this corresponds to adding a constant in the exponent; it has, however, no effect on the resulting model since constants in the exponent cancel out with the normalization factor.

combination in the exponent of $q_{\boldsymbol{\lambda}}$ can be decomposed into a sum of terms each of which depends on a single environmental variable. Plotting the value of each term as a function of the corresponding environmental variable we obtain feature profiles for the respective variables. This decomposition can be carried out for L, LQ and T models, but not for LQP models. Note that adding a constant to a profile has no impact on the resulting distribution as constants in the exponent cancel out with $Z_{\boldsymbol{\lambda}}$. For L models profiles are linear functions, for LQ models profiles are quadratic functions, and for T models profiles can be arbitrary step functions. These profiles provide an easier to understand characterization of the distribution than the vector $\boldsymbol{\lambda}$.

Figure 5 shows feature profiles for an LQ run on the first partition of the Yellow-throated Vireo and two T runs with different values of β. The value of $\beta = 0.01$ only prevents components of $\boldsymbol{\lambda}$ from becoming extremely large, but it does little to prevent heavy overfitting with numerous peaks capturing single training examples. Raising β to 1.0 completely eliminates these peaks. This is especially prominent for the aspect variable where the regularized T as well as the LQ model show no dependence while the insufficiently regularized T model overfits heavily. Note the rough agreement between LQ profiles and regularized T profiles. Peaks in these profiles can be interpreted as intervals of environmental conditions favored by a species. However, from a flat profile we may not conclude that the species distribution does not depend on the corresponding variable since variables may be correlated and maxent will sometimes pick only one of the correlated variables.

4.6. Acceleration

For the LQP version of maxent, line search on α substantially accelerated convergence when meaured in terms of log loss both on training and on test data. Log loss on test data in the first partition decreased with running time (measured on a 1GHz Pentium) as follows:

Bird	Line search?	10s	50s	100s	300s
LS	no	10.424	10.205	10.131	10.068
LS	yes	10.130	10.054	10.047	10.040
YV	no	10.086	9.658	9.536	9.433
YV	yes	9.540	9.358	9.339	9.334

The observed acceleration is similar to that obtained by Goodman (2002). Line search made no discernible dif-

ference for threshold features. Indeed, while there is an approximation made in the derivation of α in Sections 2 and 2.2, the derivation is exact for binary features, hence line search is not needed. Maxent was much faster with threshold features: log loss was within .001 of convergence in at most 50 seconds for both species.

5. Conclusions

Species distribution modeling represents a scientifically important area that deserves the attention of the machine learning community while presenting it with some interesting challenges.

In this work, we have shown how to use maxent to predict species distributions. Maxent only requires positive examples, and in our study, is substantially superior to the standard method, performing well with fairly few examples, particularly when regularization is employed. The models generated by maxent have a natural probabilistic interpretation, giving a smooth gradation from most to least suitable conditions. We have also shown that the models can be easily interpreted by human experts, a property of great practical importance.

While maxent fits the problem of species distribution modeling cleanly and effectively, there are many other techniques that could be used such as Markov random fields or mixture models. Alternatively, some of our assumptions could be relaxed, mainly that of the independence of sampling. In our future work, we plan to address sampling bias and include it in the maxent framework in a principled manner. We leave the question of alternative techniques to attack this problem open for future research.

Acknowledgments

Thanks to Rob Anderson, Michael Collins, Ned Horning, Claire Kremen, Chris Raxworthy, Sacha Spector and Eleanor Sterling for numerous helpful discussions. R. Schapire and M. Dudík received support through NSF grant CCR-0325463. Part of this work was completed while R. Schapire was employed by AT&T Labs – Research. M. Dudík was partially supported by a Gordon Wu fellowship.

References

Anderson, R. P., Lew, D., & Peterson, A. T. (2003). Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling, 162*, 211–232.

Anderson, R. P., & Martínez-Meyer, E. (2004). Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biological Conservation, 116*, 167–179.

Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics, 22*, 39–71.

Chen, S. F., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *IEEE Trans. on Speech and Audio Processing, 8*, 37–50.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning, 48*, 253–285.

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Math. Statistics, 43*, 1470–1480.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*, 1–13.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2004). Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the 17th Annual Conference on Computational Learning Theory*.

Elith, J. (2002). Quantitative methods for modeling species habitat: Comparative performance and an application to Australian plants. In S. Ferson and M. Burgman (Eds.), *Quantitative methods for conservation biology*, 39–58. New York: Springer-Verlag.

Goodman, J. (2002). Sequential conditional generalized iterative scaling. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 9–16).

Goodman, J. (2003). *Exponential priors for maximum entropy models* (Technical Report). Microsoft Research. (Available from http://research.microsoft.com/~joshuago/longexponentialprior.ps).

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology, 22*, 415–427.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. *Proceedings of the Sixth Conference on Natural Language Learning* (pp. 49–55).

Minka, T. (2001). *Algorithms for maximum-likelihood logistic regression* (Technical Report). CMU CALD. (Available from http://www.stat.cmu.edu/~minka/papers/logreg.html).

New, M., Hulme, M., & Jones, P. (1999). Representing twentieth-century space-time climate variability. Part 1: Development of a 1961-90 mean monthly terrestrial climatology. *Journal of Climate, 12*, 829–856.

Peterson, A. T. (2001). Predicting species' geographic distributions based on ecological niche modeling. *The Condor, 103*, 599–605.

Peterson, A. T., Papes, M., & Kluza, D. A. (2003). Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science, 51*, 863–868.

Peterson, A. T., & Robins, C. R. (2003). Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology, 17*, 1161–1165.

Peterson, A. T., & Shaw, J. (2003). *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: ecological niche models, predicted geographic distribution, and climate change effects. *International Journal of Parasitology, 33*, 919–931.

Ponder, W. F., Carter, G. A., Flemons, P., & Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology, 15*, 648–657.

Raxworthy, C. J., Martinez-Meyer, E., Horning, N., Nussbaum, R. A., Schneider, G. E., Ortega-Huerta, M. A., & Peterson, A. T. (2004). Predicting distributions of known and unknown reptile species in Madagascar. *Nature, 426*, 837–841.

Salakhutdinov, R., Roweis, S. T., & Ghahramani, Z. (2003). On the convergence of bound optimization algorithms. *Uncertainty in Artificial Intelligence 19* (pp. 509–516).

Sauer, J. R., Hines, J. E., & Fallon, J. (2001). The North American breeding bird survey, results and analysis 1966–2000, Version 2001.2. http://www.mbr-pwrc.usgs.gov/bbs/bbs.html. USGS Patuxent Wildlife Research Center, Laurel, MD.

Stockwell, D., & Peters, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science, 13*, 143–158.

Stockwell, D. R. B., & Noble, I. R. (1992). Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation, 33*, 385–390.

Stockwell, D. R. B., & Peterson, A. T. (2002). Controlling bias in biodiversity data. In J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall and F. B. Samson (Eds.), *Predicting species occurrences: Issues of accuracy and scale*, 537–546. Washington, DC: Island Press.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., & Williams, S. E. (2004). Extinction risk from climate change. *Nature, 427*, 145–148.

Wiley, E. O., McNyset, K. M., Peterson, A. T., Robins, C. R., & Stewart, A. M. (2003). Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography, 16*, 120–127.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation, 7*, 117–143.