Applications of Conditional Topic Models to Species Distribution Prediction

by

Paul Christopher Wilkins

A PROJECT

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented July 20th, 2010
Commencement June 2010

Chapter 1 – Introduction

The topic of species distribution modelling has been on of increasing interest in recent years. As climate change is becoming of even more interest to researchers, more tools are needed to better analyze and predict various climate change scenarios. One particular area of interest is that of species distribution modeling. Species distribution modelling addresses the problem of determining either the fundamental or the realized niche of a species, either at the current time or projecting into the past or future. Species distribution models (SDMs) are seen as a potentially powerful tool both for applied policy decisions like reservation design and theorectical understanding, discovering what factors are most important in determining the fundamental niche of a species, as well as the extent to which various factors determine how much of that niche is realized.

Currently, almost all SDMs focus on a single species at a time. For any given species, a model is developed and trained for that particular species. An advantage of this approach is that is keep computational costs down relative to a broader model. There is, however, potential in the idea that by modeling multiple species at once, mutual information between species can be leveraged to provide more accurate modeling while offering insights into the nature of the relationships between specific species. This paper examines the attempt to use one such model for doing species distribution modeling on several species at once.

1.1 Background

Multilabel Prediction (MLP) is a class of problems in which a number of input features are used to predict not a single output but rather a vector of outputs. MLP methods fall along a spectrum between two extremes. At one end, any such problem could be treated as n independent binary classification problems. This yields an output by simply applying your choice of binary classifier to each of the subproblems and concatenating the results. At the other end of the spectrum, one could consider simply treating this as one large single output classification problem. As in the independent problems case, this allows one to simply use any standard single output prediction algorithm they wish. Of course, there are serious drawbacks to this approach. In particular, the size of the output space is exponential in the number of labels.

Most MLP methods respresent a compromise between these two extremes. For example, cluster-first methods first partition the labels into cluster with high co-occurrence, then learn a binary classifier for each of these clusters. The original multilabel prediction is the done by running each of these cluster classifiers and concatenating the results. Labeling the training data for the cluster classifiers can be an issue: a cluster may be classified as present is a single member is present, if all the members are present, or some weighted majority vote of the members are present.

Another approach to the problem is to use ensemble methods. One possible ensemble approach is to learn a series of classifiers that classify jointly over a subset

of the labels, then take a weighted votes of these classifiers to determine overall classification [7]. Data sparcity is less of an issue in this case than in the fully joint case because the output space for each classifier has size 2^k, where k is the number of labels in each subset, instead of 2^n in the fully joint classification schema. As an alternative ensemble approach bagging has also been applied, where a subset of the training data is selected and for each subset of the data classification is learned only over those sets of labels the co-occur more frequently than some fixed threshold [5].

Multi-variate decision trees are an additional major approach to the multilabel prediction problem. Decision trees have many attractive features for classification: they are highly expressive but still readily interpretable. A number of extensions or alterations to multivariate regression trees have been proproposed [2, 3, 6, 8]. In general, multivariate decision trees tend to share many of the same strengths and weaknesses as their single variable counterparts. Particularly appealing is their high degree of interpretability, and moderately sized decision trees can generally be well understood without any special technical understanding. Their expressiveness is a bit of a double-edge sword. On one hand, they can theoretically represent any hypothesis. However, in practice the granularity of the hypotheses the can represent is limited by the amount of training data available, and is further reduced by the pruning which is necessary to prevent overfitting. In practice this can make interactions between input variables difficult to capture unless such interaction terms are explicitly added as additional input features.

1.2 Data Description

For this work, we used two data sets from entirely different domains. Each data set has somewhat different characteristics, but they both feature a fairly limited set of covariate data and are track species in the for mof presence/absence data.

1.2.1 Moth Data

The first dataset used in this study is from a database of moth traps throughout the H. J. Andrews Experimental Forest. These data were collected and compiled by Oregon State ecologist Dr. Jeffrey Miller, between the years of 1986 and 2008. Four different environmental covariate values were4 collected for each site: slope, aspect, elevation, and the vegetation type. Because this study involves predicting species occurrence related to the environment, the quantity of moths recorded at individual traps was discarded, and a 1 was recorded if the species occurred at that trap site over the course of the entire 23-year trapping period. Thus a vector of zeroes (absences) and ones (presences) was created for each species, with each position corresponding to a trap site. These were assembled as columns in a matrix, preceded by columns corresponding to each covariate recorded for each trap site. Each row then contained first the value for each covariate at that trap site, followed by a 1 or a 0 for each species of moth. The data used consisted of 256 traps and 606 different moth species.

After compiling the data into the described format used for modeling, it was split into training and test sets. This was done by sampling the data set without

replacement until a division was found with approximately equal distributions of species. The original data contains 32,352 individual moth records, however, it is important to note that, more than half of the included species occurred fewer than 18 times over the course of the entire trapping period, while one sixth of the species account for over half of the recorded moth occurrences. The low resolution of the majority of the moth data was one motivation for attempting multiple response modeling. As, we hypothesized that overall habitat trends would influence the accuracy of predictions for species for which there is comparatively little known (i.e. almost all present or all absent) if there was a trend of covariance between species. For the purposes of this paper, the Northwestern North American moth dataset will be referred to as "moth data".

1.2.2 Hubbard Brook Data

The second dataset provides information on the distribution of forest birds in a mountain valley at Hubbard Brook Experimental Forest (HBEF), New Hampshire, USA. Bird data were collected across a survey grid consisting of 431 points along 15 north-south transects established throughout the HBEF. Transects were 500 m apart. Points spaced at either 100-m or 200-m intervals along each transect were visited 3 times during the peak breeding season (late May through June) of 2008. During each visit the bird abundance was surveyed for 10 min using fixed 50-m radius point counts. Surveys were performed between 0530-1000 by multiple trained observers in order to limit error in observer accuracy. At each bird census

site, there is associated environmental covariate data for elevation, slope, aspect and vegetation type. The latter was measured using satellite imagery (Landsat ETM scenes). As with the moth data, information on the abundance of birds was reduced to presence/absence at all sites. The bird dataset is structured identically to moth data with columns representing the response variable and associated covariates and rows representing census sites.

Chapter 2 – Methodology

2.1 Conditional LDA

In this thesis, I examine an extension of the original LDA model[1]. In that model, all documents (also referred to in the SDM context as "sites") are drawn from the same Dirichlet prior. When data about a site ("features") are available however, we can use that information to develop a more informative prior over the topics. In the conditional latent Dirichlet allocation model (cLDA), the Dirichlet prior is taken as a function of such features[4]. Specifically, the Dirichlet prior parameter for topic t in document d is given as:

$$\alpha_{dt} = e^{\mathbf{x}_d^T \lambda_t} \tag{2.1}$$

A normal prior is placed on the λ parameters, resulting in the following generative process for cLDA:

1. For each topic t,
   (a) Draw $\lambda_t \sim N(0, \sigma^2 I)$
   (b) Draw $\phi_t \sim D(\beta)$
2. For each document d,
   (a) For each topic t let $\alpha_{dt} = e^{\mathbf{x}_d^T \lambda_t}$.

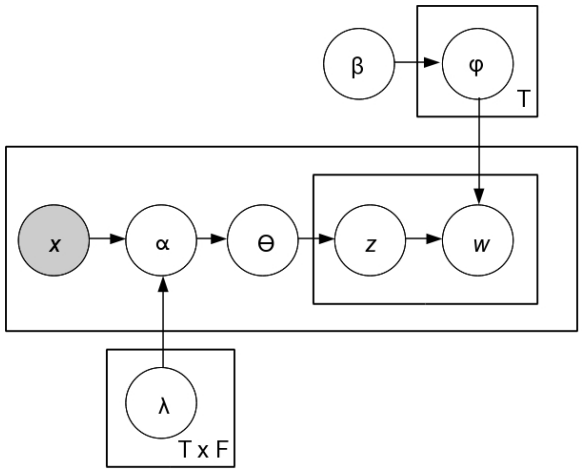

Figure 2.1: The cLDA topic model

(b) Draw $\theta_d \sim D(\alpha_d)$.

(c) For each word i,

i. Draw $z_i \sim M(\theta_d)$.

ii. Draw $w_i \sim M(\phi_{z_i})$.

In Mimno's original paper on cLDA the context was text documents, specifically machine learning papers. Here, the problems at issue were determining the likelihood of a given document and predicting the author (one of the features) of the document. In the case of species distribution prediction, we are interested in different questions. One question of great interest is that of species prediction: given the features of a particular site, predict what set of species will be present at a site. This question may be interpreted as calculating the most likely document given a set of features, that is:

$$\arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{x}, \lambda, \phi) \tag{2.2}$$

2.2 Bernoulli Conditional LDA

This formulation does not quite capture the problem being examined, however, as the above formulation does not answer the question of document length. In the above maximization problem as given, the number of words With the species distribution problem, we are not interested in specific counts of individual species, but simply whether or not each species is present or absent.

Indeed, the notion of document length is of questionable value to the species prediction problem, as the prediction being made will always be very sensitive to the document length selected. Furthermore, as the data is based on presence/absence metrics, many different documents may map to the same species prediction, making it impossible to create a one to one map from species predictions to documents.

To address these issues, we have modified the cLDA model to use multivariate Bernoulli topics instead of multinomial topics. Under the multivariate Bernoulli cLDA model, the generative process is as follows:

1. For each topic t,
   (a) Draw $\lambda_t \sim N(0, \sigma^2 I)$
   (b) For each vocabulary word v,
      i. Draw $\phi_{tv} \sim Gamma(\beta_v)$

2. For each document d,
   (a) For each topic t let $\alpha_{dt} = e^{\mathbf{x}_d^T \lambda_t}$.
   (b) Draw $\theta_d \sim D(\alpha_d)$.
   (c) For each vocabulary word v,
      i. Draw $z_i \sim M(\theta_d)$.
      ii. Draw $w_i \sim Bernoulli(\phi_{z_i,v})$.

Note that in this model, the length of every the document is equal to the number of words in the vocabulary. Conceptually, instead of repeatedly reaching into a bag of words and pulling out a word each time, under the Bernoulli cLDA model a single draw is made for each vocabulary word, with the result that that word either is or is not in the document. Such an approach, however, does require an alteration of the Gibbs sampling technique used in training the model as in Mimno.

As with the standard cLDA model, in the Gibbs Sampling step we are sampling from $P(z_i = j|\mathbf{z}_{-i}, \mathbf{w})$. As in Griffiths, we again begin with:

$$P(z_{di} = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_{di}|z_{di} = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_{di} = j|\mathbf{z}_{-i}) \tag{2.3}$$

The first term can be broken down as

$$P(w_{di}|z_{di} = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int P(w_{di}|z_{di} = j, \phi_{ji})P(\phi_{ji}|\mathbf{z}_{-i}, \mathbf{w}_{-i})d\phi_{ji} \tag{2.4}$$

Here we see the the derivation begins to differ from multinomial LDA, because the index of the word determines the particular ϕ parameter used, rather than a vector defining a multinomial distribution. From this we get

$$P(w_{di}|z_{di} = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{\#[(d') : z_{d'i} = j \wedge w'_d i = 1] + \beta}{\#[(d') : z_{d'i} = j] + 2\beta} \tag{2.5}$$

The second term can be broken down as

$$P(z_{di} = j|\mathbf{z}_{-di}) = \int P(z_{di} = j|\theta_d)P(\theta_d|\mathbf{z}_{-i})d\theta_d \tag{2.6}$$

This term resolves identically to the multinomial case:

$$P(z_{di} = j|\mathbf{z}_{-di}) = \frac{\#[i' : z_{di'} = j] + \alpha_d j}{V - 1 + K\alpha} \tag{2.7}$$

2.3 Parameter tuning

In order to test the model, there are several parameters that should first be tuned. We first tuned them by setting aside a portion of the training data as a validation set. In testing the models, a default set of parameters was chosen, and several versions of the model were trained by varying a single parameter. From this series of experiments, we took the best value of the tested parameter in each experiment and used that for our final testing. Here is an overview of the parameters tuned:

2.3.1 Gibbs Sampling

In the EM algorithm used to train this model, the E step involves sampling the z's (unobserved topic variables), and using those to calculate estimates of θ and ϕ. We optimized the following variables:

- Burn-in time: the number of iterations of Gibbs sampling done before beginning to take samples
- Number of samples: the number of samples taken during a round of Gibbs sampling
- Sample distance: the number of iterations of Gibbs sampling done between each sample taken

2.3.2 LBFGS Optimization

The maximization step of the EM algorithm involves the use of an LBFGS optimizer to optimize the lambdas given the values sampled in the E step.

- Max Iterations: maximum number of iterations the optimizer will run before exiting regardless of convergence
- Δx: threshold for convergence based on change in the lambda values
- Δf: threshold for convergence based on change in the optimization function

2.3.3 LDA parameters

The LDA model itself has some parameters that may be tuned

- β: By default, LDA has used a uniform Dirichlet prior on topic composition (that is ϕ). We considered trying asymmetric priors, but preliminary results confirmed the observation by Wallach/Mimno that using a symmetric β produced better results
- Number of topics: the number of topics to be used is highly domain dependent. With too few topics, the model may not be expressive enough. With too many, there may not be enough data to produce meaningful topics.

Chapter 3 – Results

For each experiment that was run, the results were analyzed in a number of ways. To determine presence or absence, a threshold must be set, as the model produces real-valued outputs. Three possible methods of thresholding were tried. Macro averaging or global thresholding involes setting a single threshold for all species and all sites. This approach may help avoid overfitting for individual species. Micro averaging, on the other hand, sets thresholds on a per site basis. Finally, the multi-tash average set thresholds on a per species basis.

Once the thresholds had been set and the confusion matrices had been calculated, the following metrics were calculated for each matrix:

- F1 score - This is the harmonic mean of precision and recall, which ranges from a worst value of 0 to a best of 1.
- AUC - This score measures the probability that a site where the species is present will be ranked more highly than a site where the species was absent.
- Precision - the number of true positives divided by the total number of examples labeled as positive. This is useful for determining the accuracy of the algorithm at identifying the total potential range of the species. Precision was calculated at 90% and 100% recall.

- Cohen's Kappa - This measures the level of relative agreement between the model's predictions and the actual presence/absence data. As a heuristic, a Cohen's Kappa > 0.4 tends to indicate at least a moderate level of agreement.

3.1 Tuning Results

As mentioned earlier, tuning was accomplished by means of searching over a sginel parameter at a time using manually defined steps, then taking the best result for each parameter and using that for the final evaluation. With this approach, I found that there was very little pattern or trend to the results as I varied the parameters, and the differences were usually very small. I have included here data from one of the more varied parameters, the number of topics used by the model, to illustrate the sort of results produced. As I will discuss more later, I believe this difficulty in tuning the model may be a mjor factor in the overall results achieved.

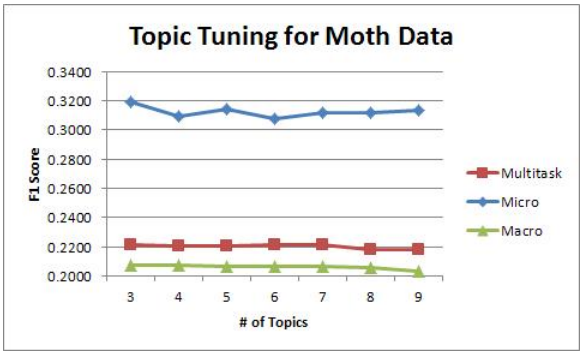

Figure 3.1: Topic number tuning for the Andrews data

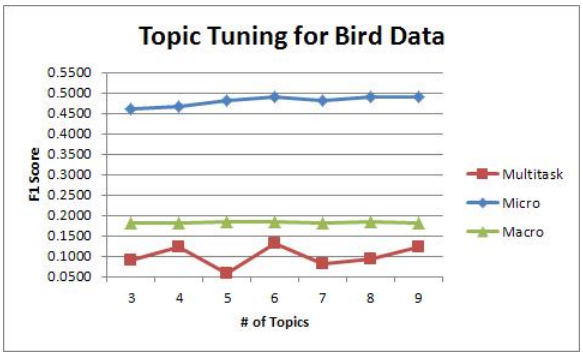

Figure 3.2: Topic number tuning for the Hubbard Brook data

3.2 Comparison to Other Models

For evaluation, the best of these results were compared against several other models that were applied to the same multi-label SDM problem. The following models were used for comparison:

- Logistic Regression (lr) - A separate model was run for each species with regularized covariates
- GLMnet - Another per species model, with a mixture of l1 and l2 regularization penalties
- Neural Network (nnet) - multi-layer perceptron with a tuned number of hidden units. This was a joint model across all species
- Random Forests - Each model consisted of 500 randomized bagged decision trees. A different model was learned for each species
- Hybrid-LDA (lr+tm) - This apporach combined LDA with logistic regression,

using the logistic regerssion to start the annealing process and help score it along the way. This was a joint model acroos all species

Table 3.1: Comparison of SDM methods on Hubbard Brook bird data (micro)

Metric	ctm	gbm	glmnet	nnet	rf	lr+tm
AUC	0.87	0.89	0.9	0.83	0.86	0.87
F1	0.49	0.54	0.52	0.43	0.53	0.47
Kappa	0.42	0.47	0.46	0.36	0.45	0.41
prec90	0.15	0.17	0.17	0.12	0.13	0.16
prec100	0.15	0.17	0.17	0.12	0.2	0.16

Table 3.2: Comparison of SDM methods on Andrews moth data (micro)

Metric	ctm	gbm	nnet	rf
AUC	0.77	0.8	0.75	0.78
F1	0.32	0.33	0.26	0.3
Kappa	0.23	0.25	0.17	0.21
prec90	0.13	0.11	0.04	0.09
prec100	0.13	0.13	0.08	0.13

Chapter 4 – Discussion

As can be seen from these results, the conditional LDA model performed on par with the other methods tested, but not particularly better. This result was not quite what we expected, as it was hoped that the conditional LDA model would perform signficantly better than the single species models, due to the potential for leveraging correlations between species to aid prediction. Tuning the model proved very difficult, and most parameters seemed to have little clear trend when they were varied while leaving other parameters at a fixed value. I did not explore the option of doing some sort of gradient search over the parameter space, as such an approach would be computationally prohibitive using the methods outlined in this paper.

Computational constraints were a continual obstacle in training the conditional LDA model, and ultimately proved to be a serious barrier to trying the algorithm on larger data sets. Both the Hubbard Brook and Andrews data sets are relatively small, so it is possible that there simply was not enouogh data available to gain a signficant advantage from cross-species information. There is general interest in the field in developing faster methods for training conditional LDA and similar models, and perhaps with advances computational efficiency running on larger data sets would be feasible and might produce more noticable gains over single species methods.

Bibliography

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Hendrik Blockeel, Maurice Bruynooghe, Saso Dzeroski, Jan Ramon, and Jan Struyf. Hierarchical multi-classification, 2002.

[3] S. Keon Lee. On generalized multivariate decision tree by using GEE. *Computational Statistics and Data Analysis*, 49(4):1105–1119, 2005.

[4] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet multinomial regression.

[5] Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Multi-label classification using ensembles of pruned sets. *Data Mining, IEEE International Conference on*, 0:995–1000, 2008.

[6] E. Suzuki, M. Gotoh, and Y. Choki. Bloomy decision tree for multi-objective classification. *Lecture notes in computer science*, pages 436–447, 2001.

[7] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Lecture Notes in Computer Science*, 4701:406, 2007.

[8] H. Zhang. Classification Trees for Multiple Binary Responses. *Journal of the American Statistical Association*, pages 180–193, 1998.