Multi-label Classification for Species Distribution Modeling

Jun Yu, Weng-Keen Wong, Tom Dietterich {YUJU,WONG,TGD}@EECS.OREGONSTATE.EDU
School of EECS, Oregon State University, Corvallis, OR 97331 USA

Julia Jones JONESJ@GEO.OREGONSTATE.EDU
Dept of Geosciences, Oregon State University, Corvallis, OR 97331 USA

Matthew Betts, Sarah Frey {MATTHEW.BETTS,SARAH.FREY}@OREGONSTATE.EDU
Susan Shirley {SUSAN.SHIRLEY}@OREGONSTATE.EDU
Dept of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331 USA

Jeffrey Miller JEFFREY.MILLER@OREGONSTATE.EDU
Dept of Rangeland Ecology and Management, Oregon State University, Corvallis, OR 97331 USA

Matt White MATT.WHITE@DSE.VIC.GOV.AU
Arthur Rylah Institute, Dept of Sustainability and Environment, Heidelberg, Victoria 3084 Australia

Abstract

Species distribution models play a key role in creating reserves for species conservation, predicting the effects of ecological change, and testing ecological theory. Although many methods have been developed for models of individual species, ecologists are recognizing the advantages of predicting multiple species simultaneously. This problem of multiple species prediction can be addressed by machine learning algorithms from the area of multi-label classification. However, to date, multi-label classification has been applied primarily to problems in text and image annotation. The goal of this paper is to introduce species distribution modeling as a new domain for multi-label classification, present preliminary results illustrating the advantages of multi-label algorithms, and discuss new research directions presented by this domain.

1. Introduction

Environmental change is a major challenge for researchers and policymakers. Many ecologists are studying how species distributions respond to environmental change. Species distribution modeling (SDM) involves predicting the occurrence of a species at a site given a set of environmental features (such as climate, elevation, vegetation, and land use) (Elith et al., 2006). SDM plays an important role in species conservation (Ferrier et al., 2002), predicting the effects of ecological change on species (Thomas et al., 2004) and testing ecological theory (Austin, 2002). Publicly accessible datasets for SDM, such as eBird (Munson et al., 2009), are now available.

Much work in SDM has focused on species-level modeling, which predicts the distribution of a single species. Methods for modeling individual species include Bioclimatic Envelopes (Araujo et al., 2005), Genetic Algorithms (Stockman et al., 2006), regression approaches such as GAMs and GLMs, and machine learning approaches such as Maximum Entropy models (Phillips et al., 2004) and Boosted Regression Trees (Elith et al., 2008). Machine learning approaches are consistently among the top performers for single-species models (Elith et al., 2006). Community-level predictions, which predict the occurrence of multiple species, or species assemblages, may provide better ecological interpretations and more accurate predictions, especially for rare species, than single-species models (Ferrier & Guisan, 2006). Communities may be structured by environmental conditions or by biological interactions among species, or both. As a result, approaches to multi-species modeling place different priority upon assembling (finding groups of co-occurring species) and

Appearing in the *ICML 2011 Workshop on Machine Learning for Global Challenges*, Bellevue, WA, USA, 2011. Copyright 2011 by the author(s)/owner(s).

Table 1. Species datasets used in our experiments.

Dataset	**Plants**	**HJA Moths**	**HJA Birds**	**HB Birds**	**eBird**
Instances	15327	256	364	371	5985
Features	81	52	21	40	22
Labels	50	422	23	34	34

predicting (finding relationships of a species to environmental factors). These approaches are: 1) assemble first, predict later, 2) predict first, assemble later, and 3) assemble and predict together (Ferrier & Guisan, 2006). The assemble and predict together strategy has the potential to reveal the relative importance of biological interactions between species compared to environmental controls on communities, a topic of major theoretical and practical importance in ecology. However, this strategy has received little attention, perhaps because ecologists are unfamiliar with algorithms for multiple response prediction.

Multiple response prediction is precisely the task addressed by multi-label classification algorithms. In multi-label classification, the goal is to learn a classifier that maps from a vector of features $\mathbf{x}$ to a vector of output labels $\mathbf{y} = (y_1, \ldots, y_L)$, where each label y_l is typically binary valued. Numerous techniques for multi-label classification have been proposed in the literature, including graphical models (Ghamrawi & McCallum, 2005), ensemble methods (Read et al., 2009), dimensionality reduction (Ji et al., 2010) and kernel learning (Elisseeff & Weston, 2002). Multi-label classification has been primarily applied to the task of text labeling (Ghamrawi & McCallum, 2005) and image annotation (Ji et al., 2010).

Our goal is to introduce SDM as a new domain for multi-label classification. We include empirical results comparing multi-label algorithms versus algorithms that predict each label independently (known as *binary relevance* algorithms in the multi-label literature) on five different species datasets. Although SDM could be viewed as a typical multi-label classification problem, we discuss several unique challenges in this domain, along with directions for future work.

2. Methodology

Table 1 summarizes characteristics of the five species datasets used in our experiments. Each data instance corresponds to species observations at a particular site. The features associated with each data instance consist of environmental characteristics at that site. The first four datasets in the table were collected and compiled by trained scientists. These four datasets include the Plants dataset from the Arthur Rylah Institute in Victoria, Australia, the HJA Moths and Birds datasets from the H. J. Andrews Experimental Forest in Oregon, and the HB Birds dataset from the Hubbard Brook Experimental Forest in New Hampshire.

The fifth dataset was from the eBird project (Munson et al., 2009), which is a citizen science program that allows bird watchers to upload bird species observations. Although the eBird dataset has known sources of bias due to issues such as the expertise level of the citizen scientists (Yu et al., 2010), we will not address these issues in our experiments. Instead, we will use the eBird dataset purely for evaluating multi-label algorithms. In our experiments, we used a subset of the eBird data consisting of eBird checklists from New York state during the breeding season (May-June) in 2006-2008.

We split each dataset into training, validation, and testing partitions. To alleviate the effects of spatial autocorrelation, we placed a black and white checkerboard over the entire dataset. Data points falling into white cells formed the test set while those falling into the black cells were further divided into the training and validation sets.

We compared the performance of two binary relevance classifiers against their multi-label classifier extensions. We applied binary relevance SVM (BR-SVM) with a linear kernel, tuning the C parameter over the validation set, and binary relevance boosted regression trees (BR-BRT), tuning the number of trees and the depth of the trees. As the multi-label algorithm, we employed the Ensemble of Classifier Chains (ECC) algorithm (Read et al., 2009). We chose ECC because it consistently performed well on species data as compared to other publicly-available multi-label classifiers. A classifier chain in ECC is a sequence of base classifiers that predict the label y_j given the input feature $\mathbf{x}$ and all output labels $y_1, \ldots, y_{j-1}$ already predicted in the chain. The ensemble of classifier chains is created by 20 random orderings of the labels and by training each chain on a bootstrap replicate of the original training data. We used linear SVMs and BRTs as the base classifiers, resulting in multi-label algorithms ECC-SVM and ECC-BRT. The base classifiers shared the same tunable parameters, which were tuned over the validation set. In order to report confidence intervals, we generated 1000 bootstrapped samples from the test data and computed mean and 95% bootstrap confidence intervals for each metric.

We compared the performance of our algorithms using Hamming loss and Ranking loss, which are both standard multi-label metrics. We also reported two metrics that are relevant to the SDM domain: site-based and species-based AUC. In site-based AUC, the AUC is computed for the present vs. absent species at each site $\mathbf{x}_i$ and then averaged across sites. In species-based AUC, the AUC is computed on each species over all sites and then averaged across species.

3. Results and Discussion

Table 2 summarizes the results of our experiments, which demonstrate the benefit of multi-label approaches. ECC-SVM outperformed BR-SVM in all rows while ECC-BRT outperformed BR-BRT in 13 out of 20 rows. Since linear SVMs were unable to capture non-linear interactions between features, the gain in performance by ECC-SVM over BR-SVM came from the correlation between labels. BR-BRT, on the other hand, could capture non-linear interactions between the input features but, being a binary relevance algorithm, it was unable to exploit label correlations. However, once the label correlations were modeled by ECC-BRT, the gains over BR-BRT were not as substantial as the gains by ECC-SVM over BRT-SVM. Further analysis suggested that the large gains by ECC-SVM over BR-SVM were due to the presence/absence of other species acting as a proxy for the non-linear interactions between environmental features. In addition, for situations in which ECC-BRT outperformed BR-BRT, we found that these gains were due to more accurate predictions for rarer species. We are actively investigating these findings further.

Having shown the importance of modeling interactions between the output labels, we now discuss some of the unique properties of species data that can produce new research directions for multi-label classification. Species datasets, as shown in Table 1 are fundamentally different from many of the common multi-label datasets used in the literature. Unlike text, the feature space is dense and lower dimensional, with fewer than 50 features in a typical dataset. The output space, on the other hand, is often large, ranging from tens to thousands of species, and sparse, with few species present at each site. In our experiments, we used a subset of the species in the data. Almost none of the current multi-label algorithms are well-suited for the full species datasets. New algorithms need to be developed for data of this form and for accurately predicting rarer species, which are of great interest to ecologists.

Second, in SDM, interactions between species are complex, with communities of species forming due, for instance, to cooperation or competition. Many of the existing multi-label algorithms only model pairwise correlations between labels and are thus inadequate for helping ecologists discover these complex interactions. Algorithms that can discover more complex correlations between labels are often focused purely on predictive performance and do not produce comprehensible models. This need for interpretability is important to provide insight into the ecological principles governing interactions between species. In particular, ecologists are interested in comparing the relative importance of two major forces that structure ecological communities – environmental factors and biological interactions between species.

4. Conclusion

Using species distribution data, we have demonstrated that multi-label algorithms that captured correlations between species outperformed algorithms that predicted each species individually. These results on species data also point to new research directions for multi-label algorithms. These directions include developing new algorithms for datasets with sparse high dimensional label space but lower dimensional feature space, modeling more complex interactions beyond pairwise label correlations, and producing comprehensible multi-label models. These research directions can provide valuable contributions not only to machine learning but also to ecology.

Acknowledgments

This work was supported in part by NSF grants ARC 0941748 and CCF 0832804.

References

Araujo, M. B., Pearson, R. G., Thuiller, W., and Erhard, M. Validation of species-climate impact models under climate change. *Global Change Biology*, 11:1504–1513, 2005.

Austin, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecol Model*, 157:101–118, 2002.

Elisseeff, A. and Weston, J. A kernel method for multi-labelled classification. In *Advances in NIPS 14*, pp. 681–687, 2002.

Elith, J., Graham, C. H., Anderson, R. P., et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129–151, 2006.

Elith, J., Leathwick, J. R., and Hastie, T. A working guide to boosted regression trees. *J Anim Ecol*, 77: 802–813, 2008.

Table 2. Performance of various algorithms in terms of Hamming loss, Ranking loss, Site-based AUC and Species-based AUC. 95% bootstrap confidence intervals are also shown. The ● symbol indicates a significant improvement by ECC over BR using SVM or BRT as the base classifier.

Dataset	**BR-SVM**	**ECC-SVM**	**BR-BRT**	**ECC-BRT**
Hamming Loss				
Plants	0.0655 ± 0.0001	0.0653 ± 0.0001 ●	**0.0650 ± 0.0001**	**0.0650 ± 0.0001**
HJA Moths	0.1350 ± 0.0011	0.1331 ± 0.0011	0.1335 ± 0.0011	**0.1327 ± 0.0011**
HJA Birds	0.1690 ± 0.0005	**0.1645 ± 0.0005 ●**	0.1676 ± 0.0005	0.1651 ± 0.0005
HB Birds	0.1033 ± 0.0003	0.1003 ± 0.0003	**0.0978 ± 0.0003**	0.0983 ± 0.0003
eBird	0.2140 ± 0.0001	0.2095 ± 0.0001 ●	0.1855 ± 0.0001	**0.1844 ± 0.0001**
Ranking Loss				
Plants	0.1803 ± 0.0001	0.1695 ± 0.0001 ●	**0.1491 ± 0.0001**	0.1520 ± 0.0001
HJA Moths	0.2248 ± 0.0008	0.2093 ± 0.0007 ●	0.2014 ± 0.0008	**0.2013 ± 0.0007**
HJA Birds	0.1312 ± 0.0006	0.1236 ± 0.0006 ●	0.1207 ± 0.0005	**0.1174 ± 0.0005 ●**
HB Birds	0.1205 ± 0.0006	0.1053 ± 0.0005 ●	**0.1013 ± 0.0005**	0.1043 ± 0.0006
eBird	0.2134 ± 0.0002	0.1794 ± 0.0002 ●	0.1763 ± 0.0002	**0.1714 ± 0.0002 ●**
Site-based AUC				
Plants	0.8197 ± 0.0001	0.8304 ± 0.0001 ●	**0.8508 ± 0.0001**	0.8479 ± 0.0001
HJA Moths	0.7751 ± 0.0008	0.7907 ± 0.0007 ●	0.7986 ± 0.0008	**0.7991 ± 0.0007**
HJA Birds	0.8688 ± 0.0006	0.8764 ± 0.0006 ●	0.8793 ± 0.0005	**0.8832 ± 0.0005 ●**
HB Birds	0.8796 ± 0.0006	0.8947 ± 0.0005 ●	**0.8987 ± 0.0005**	0.8958 ± 0.0005
eBird	0.7866 ± 0.0002	0.8206 ± 0.0002 ●	0.8237 ± 0.0002	**0.8285 ± 0.0002 ●**
Species-based AUC				
Plants	0.7750 ± 0.0002	0.7852 ± 0.0002 ●	**0.8204 ± 0.0001**	0.8047 ± 0.0002
HJA Moths	0.5173 ± 0.0001	0.6090 ± 0.0001 ●	0.6232 ± 0.0001	**0.6261 ± 0.0001**
HJA Birds	0.6498 ± 0.0005	0.6847 ± 0.0005 ●	0.6956 ± 0.0004	**0.7067 ± 0.0004 ●**
HB Birds	0.6419 ± 0.0006	0.7048 ± 0.0006 ●	0.7093 ± 0.0006	**0.7157 ± 0.0006 ●**
eBird	0.6111 ± 0.0003	0.6369 ± 0.0003 ●	0.6979 ± 0.0003	**0.7150 ± 0.0003 ●**

Ferrier, S. and Guisan, A. Spatial modelling of biodiversity at the community level. *J Appl Ecol*, 43: 393–404, 2006.

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. ii. community-level modelling. *Biodiversity and Conservation*, 11(12):2309–2338, 2002.

Ghamrawi, N. and McCallum, A. Collective multilabel classification. In *Proceedings of CIKM*, pp. 195–200, 2005.

Ji, S., Tang, L., Yu, S., and Ye, J. A shared-subspace learning framework for multi-label classification. *ACM TKDD*, 4(2):8:1–8:29, 2010.

Munson, M. A., Webb, K., Sheldon, D.l, et al. The ebird reference dataset, version 1.0. Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY, June 2009.

Phillips, S. J., Dudik, M., and Schapire, R. E. A maximum entropy approach to species distribution modeling. In *Proceedings of ICML*, pp. 83–91, 2004.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. In *Proceedings of ECML-PKDD*, pp. 254–269, 2009.

Stockman, A. K., Beamer, D. A., and Bond, J. E. An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate. *Diversity and Distribution*, 12:81–89, 2006.

Thomas, C. D., Cameron, A., Green, R. E., et al. Extinction risk from climate change. *Nature*, 427 (6970):145–148, 2004.

Yu, J., Wong, W-K., and Hutchinson, R. Modeling experts and novices in citizen science data for species distribution modeling. In *Proceedings of ICDM*, pp. 1157–1162, 2010.