Species Distribution Modeling and Prediction: A Class Imbalance Problem

Reid A. Johnson
Dept. of Computer Science & Engineering
University of Notre Dame
Notre Dame, Indiana 46556
rjohns15@nd.edu

Nitesh V. Chawla
Dept. of Computer Science & Engineering
University of Notre Dame
Notre Dame, Indiana 46556
nchawla@nd.edu

Jessica J. Hellmann
Dept. of Biological Sciences
University of Notre Dame
Notre Dame, Indiana 46556
hellmann.3@nd.edu

***Abstract*—Predicting the distributions of species is central to a variety of applications in ecology and conservation biology. With increasing interest in using electronic occurrence records, many modeling techniques have been developed to utilize this data and compute the potential distribution of species as a proxy for actual observations. As the actual observations are typically overwhelmed by non-occurrences, we approach the modeling of species' distributions with a focus on the problem of class imbalance. Our analysis includes the evaluation of several machine learning methods that have been shown to address the problems of class imbalance, but which have rarely or never been applied to the domain of species distribution modeling. Evaluation of these methods includes the use of the area under the precision-recall curve (AUPR), which can supplement other metrics to provide a more informative assessment of model utility under conditions of class imbalance. Our analysis concludes that emphasizing techniques that specifically address the problem of class imbalance can provide AUROC and AUPR results competitive with traditional species distribution models.**

I. Introduction

Forming knowledge of the factors that determine where species live and developing predictions about their distributions are important tasks for developing strategies in ecological conservation and sustainability. Often, however, there is insufficient biodiversity data to support these activities on a large scale. In response to this lack of data, many modeling techniques have been developed and used in an attempt to compute the potential distribution of species as a proxy for actual observations. Species distribution modeling is the process of combining occurrence data—locations where a species has been identified as being present or absent—with ecological and environmental variables—conditions such as temperature, precipitation, and vegetation—to create a model of a species' niche requirements.

As the number of actual observations are often quite small relative to the size of the geography that they occupy, the occurrences of a species (minority class) are often far outnumbered by the number of non-occurrences (majority class)—in other words, we can say that there is a "*class imbalance*." These non-occurrences can be either genuine absences or, more commonly, areas lacking occurrence information. Learning from these imbalanced datasets continues to be a pervasive problem in a large array of applications.

We posit that it may be useful to give the problem of species modeling a fresh look by investigating the potential utility of several general machine learning methods that are intended to address the problem of class imbalance, but that have not yet to found their way into the domain of species distribution modeling. As a detailed evaluation of the ecological realism of all models was not practical, we tested the performance of a representative selection of modeling methods for learning on presence/absence data, using datasets typical of the types of species and environmental data that are commonly employed. Our model comparison is reasonably broad, applying 8 methods to modeling the distributions of 9 species distributed variously across North America.

Contributions: While it produces significant difficulties in species distribution modeling, there has been little study of the effectiveness of methods that address the problem of class imbalance in predicting species' distributions. We address this area of study by introducing models that are particularly robust to class imbalance and applying them to the task of species distribution modeling. In addition, the effective evaluation of classifiers for species modeling also requires careful consideration. We consider both receiver operating characteristic and precision-recall curves, and discuss the merits of both vis-à-vis the extreme class imbalance in species distribution prediction. Finally, we provide recommendations for further investigation into the issue of class imbalance in species distribution modeling.

II. Problems of Class Imbalance

Class imbalance is encountered by inductive learning systems in domains for which one class is represented by many instances while the other is represented by only a few. Addressing the problem of class imbalance is particularly important because it often hinders the capability of traditional classification algorithms to identify cases of interest, such as species occurrences. The difficulties posed by class imbalance are relative, depending upon: (1) the imbalance ratio, i.e. the ratio of the majority to the minority instances, (2) the complexity of the concept represented by the data, (3) the overall size of the training dataset, and (4) the classifier involved [12].

In a typical supervised learning scenario, classifiers are trained and ranked by any of a large number of evaluation metrics. This situation is complicated by the presence of imbalance in the data. Not only can different evaluation metrics give conflicting rankings, but they may react to the presence of various levels of class imbalance in different ways. That is, not only can class imbalance make it difficult to develop effective classifiers, but it can make it difficult to develop evaluation metrics that effectively evaluate the performance of those classifiers.

In light of these problems, we focus on several evaluation metrics that have been shown to be robust to the effects of imbalance. The area under the Receiver Operating Characteristic curve (AUROC) is the de facto standard for performance in the class imbalance literature, while correlation (CORR) is frequently used in the evaluation of species distribution models. Our evaluations also include the area under the Precision-Recall curve (AUPR), a metric that we suggest may be useful for evaluating species models. These three metrics are commonly used as single representative numbers to describe classifier performance.

III. Material and Methods

A. Data for modeling

The environmental and species data used in our experiments were selected with the intent of facilitating useful comparisons while providing a reasonably wide scope of evaluation. The environmental data provides several bioclimatic variables over an environmentally heterogeneous study area. The selected species provide a range of class imbalance and spatial variation.

The environmental coverages constitute a North American grid with 10 arc-minute square cells. The coverages consist of 18 bioclimatic variables derived from the monthly temperature and rainfall values during the period 1950 to 2000. Each coverage is defined over a 302 x 391 grid, of which 67,570 points have data for all coverages [11].

As supplied, the environmental data required considerable grooming to generate datasets of consistent quality. Environmental coverages were altered so that projections, grid cell size, and spatial extent were consistent across all variables. We note that a limitation of our experiments are their restriction to a single spatial extent and grid cell size; it is a topic of current interest to evaluate how altering these factors might affect the various models studied, though also beyond the scope of this work.

All of our species data pertains to North America and is derived from the Global Biodiversity Information Facility (GBIF), an international government-managed data portal established to encourage free and open access to biodiversity data via the Internet. Some species data had more than one occurrence per grid cell, either because of repeat observations or sites in close proximity to each other; these duplicate occurrences were reduced to one record per grid cell.

Table I. Species occurrence information. The instances column corresponds to the number of species occurrences used in the study, while the imbalance column denotes the ratio of the total number of points in the study area to the number of instances.

Species	Instances	Imbalance
Vireo bellii	532	37:1
Vireo cassinii	654	30:1
Vireo flavifrons	1405	13:1
Vireo griseus	1492	13:1
Vireo huttoni	437	45:1
Vireo olivaceus	2924	6:1
Vireo philadelphicus	799	24:1
Vireo solitarius	1957	9:1
Vireo vicinior	121	167:1

The species studied are all small- to medium-sized birds belonging to the *Vireo* genus. Our study focuses on nine species prevalent in the Northeastern United States.

B. Modeling Methods

Eight models were used, some being algorithms trained in more than one way. Each model requires the use of presence and absence points. All background (i.e., non-presence) points were used as absences and all occurrence points were used as presences; accordingly, the models were developed using all of the points provided, each point distinguished as either present or absent.

The presence data were divided into random partitions: in each partition, 70% of the occurrence localities were randomly selected for the training set, while the remaining 30% were set aside for testing. This procedure was repeated ten times to create ten random 70/30 splits for each dataset. The models were then run on each of these datasets. Several statistics were computed and averaged over the ten runs for each model.

On each of these runs, we evaluated eight models: unpruned C4.5 decision trees (C4.5) [17], Classification and Regression Trees (CART) [1], logistic regression (LR), maximum entropy (MAXENT) [15], Naïve Bayes (NB) (with kernel density estimation), Hellinger Distance decision trees (HDDT) [4], Random Forests (RF) [2], and Random Forests with SMOTE applied (RF-SMT) [3]. Each model can be considered to use some rules or mathematical algorithms to define the ecological niche of the species based on the distribution of the species records in the environmental space. Once the predicted species niche is defined, the projection of the model into the geographical space produces a predictive map. MAXENT is a method based specifically on the concept of the ecological niche, while the other methods used have proven useful in other domains. Each method will be briefly explained in turn.

We used several decision trees in our experiments. A decision tree is a tree in which each branch node represents a choice between a number of alternative properties, wherein each leaf node represents a classification or decision. The important function to consider when building a decision tree is known as the splitting criterion, which defines how data should be split in order to maximize performance.

C4.5 builds decision trees from a set of training data using the concept of information entropy, a measure of uncertainty, to define the splitting criteria [17].

CART uses the Gini index, a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset [1]. While C4.5 and CART can be effective on datasets that have been sampled, they are considered to be "skew sensitive" [10]; that is, the methods can become biased as the degree of class imbalance increases.

HDDT, another decision tree, uses the measure of Hellinger distance to decide between alternative properties, a measure that quantifies the similarity between two probability distributions [4][19]. By using this measure, HDDT capitalizes on the importance of designing a decision tree splitting criterion that captures the divergence in distributions while being skew-insensitive [5].

Random Forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees [2]. In our experiments, we ran two implementations of Random Forest: one simply used Random Forest with the generated training sets, the other used **Random Forest with SMOTE** applied to the training sets.

SMOTE is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples. This is done by taking each minority class sample and introducing the synthetic examples along the line segments joining any/all of the nearest neighbors to minority class. SMOTE has proven effective at achieving better classifier performance than only under-sampling the majority class in imbalanced datasets [3].

Logistic regression is a type of regression analysis used for predicting the outcome of a categorical variable (a variable that can take on a limited number of categories) based on one or more predictor variables. In logistic regression, the regression coefficients represent the rate of change in the logit—the logarithm of the odds ratio—for each unit change in the predictor.

Naïve Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions. In general terms, a naïve Bayes classifier assumes that, given the class, the presence or absence of a particular feature is unrelated to the presence or absence of any other feature. Despite these strong assumptions, Naïve Bayes often performs quite well, especially as it only requires a relatively small amount of training data to estimate the parameters necessary for classification [18].

MAXENT estimates species' distributions by finding the distribution of maximum entropy (i.e. closest to uniform) subject to the constraint that the expected value of each environmental variable (or its transform and/or interactions) under this estimated distribution matches its empirical average). MAXENT is a leading model that has been shown to perform well in predicting species distributions, as evaluated by [9].

Table II. Modeling methods implemented.

Method	Class of model and explanation	Software
C4.5	decision tree using information gain	Weka
CART	decision tree using Gini coefficient	Weka
LR	logistic regression	Weka
MAXENT	maximum entropy	Maxent
NB	naïve Bayes classifier	Weka
HDDT	decision tree using Hellinger distance	C++
RF	decision tree ensemble	Weka
RF-SMT	decision tree ensemble with SMOTE	Weka

C. Metric Descriptions

The evaluation was focused on each model's predictive performance on the testing sets. Performance was evaluated using three statistics: the area under the Receiver Operating Characteristic curve (AUROC), the area under the Precision-Recall curve (AUPR), and correlation (CORR).

- AUROC has been used extensively in the species' distribution modeling literature, and provides a scalar measure of the ability of a model to discriminate between sites where a species is present versus those where it is absent [8]. If one picks a random positive example and a random negative example, then the area under the curve is the probability that the classifier correctly orders the two points (with random ordering in the case of ties). Accordingly, a perfect classifier has an AUROC of 1, while a score of 0.5 implies predictive discrimination that is no better than a random guess. The use of the AUROC metric with presence-only data indicates that we interpret all grid cells with no occurrence localities as "negative examples", even if they support good environmental conditions for the species. Therefore, in practice, the maximum AUROC is less than one, and is smaller for wider-ranging species. However, AUROC can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution [6]. In our experiments, AUROC was calculated using the trapezoid rule.
- AUPR provides a scalar measure of classifier performance by varying the decision threshold on probability estimations or scores, representing the degree to which positive examples are separated from negative

examples. AUPR captures the innate trade-off between successfully identifying positive class instances and remaining parsimonious in producing positive class predictions. AUPR is a skew-sensitive metric and is greatly affected by imbalances in the data. Thus when dealing with highly skewed datasets, AUPR can provide a especially informative picture of an algorithm's performance [6]. In the context of species modeling, AUPR significantly penalizes incorrect predictions of species occurrence (false positives). This benefits models that do not over-predict a species' range, but may over-penalize models that predict occurrences in locations where a species is assumed absent due to lack of occurrence information.

- CORR, the correlation between the observation in the occurrence dataset (a dichotomous variable) and the prediction, is known as the point biserial correlation, and can be computed as a Pearson correlation coefficient [20]. It is similar to AUROC, but also provides additional information, taking into account how far the prediction varies from the observation. This gives further insight into the distribution of the predictions.

Typically, AUROC is used in the species distribution literature for evaluation, while AUPR is generally omitted. However, by using both the AUPR and the AUROC metrics together, we gain a fuller characterization of the predictions than using either alone. For example, the AUPR curve reflects whether the first few occurrence predictions at the top of the prediction list are correct; the ROC curve does not provide this information. Additionally, AUROC and AUPR have different strengths: AUROC can be overly optimistic in cases of class imbalance while making fewer assumptions about misclassification costs than other metrics [7][16], while AUPR can be particularly useful when dealing with highly imbalanced datasets [13][14].

IV. Results

A. Performance Metrics

Model performance was evaluated using AUROC, AUPR, and CORR metrics.

Table III summarizes model performance on the AUROC metric. MAXENT obtained the highest AUROC on 4 of the 9 species datasets used for evaluation. These 4 datasets include the three most imbalanced datasets (fewest occurrences). C4.5 obtained the highest AUROC on the other 5 species datasets. Though it never obtains the highest AUROC performance on a given dataset, HDDT demonstrated the highest AUROC performance when averaged over all of the datasets, followed by C4.5 and MAXENT. The average AUROC of the top 4 models differs by less than 5%. HDDT averages about 15% higher AUROC performance than CART, which produced the lowest average AUROC of the models evaluated.

Table IV summarizes model performance on the AUPR metric. HDDT obtained the highest AUPR on 7 of the 9 species datasets, while MAXENT obtained the highest AUPR on the remaining 2. On average, HDDT received a 30% higher AUPR than MAXENT. The highest AUPR on the *V. vicinior* dataset—the most imbalanced dataset used for evaluation—was obtained by MAXENT. NB produced the lowest AUPR performance on all of the datasets used for evaluation.

Table V summarizes model performance on the CORR metric. HDDT obtained the highest CORR on 4 of the 9 datasets. HDDT also obtained the highest AUPR on each of these datasets. LR received the highest CORR on 2; and MAXENT, C4.5, and RF each obtained the highest CORR on a single dataset. HDDT obtained the highest CORR on average, followed by MAXENT and C4.5. NB demonstrated the lowest CORR performance on all of the datasets.

B. Distribution Maps

To display modeled results geographically, we show distributions predicted with several modeling methods, as shown in Figures 1 and 2. These maps illustrate variation in model predictions among techniques. The most obvious differences are in the proportion of the region that appears to be predicted most suitable for the species. The resultant distributions suggest a natural division into two groups: models that produce wide-ranging predictions, such as MAXENT and LR, and models that produce narrow, point-like predictions, such as HDDT and C4.5.

V. Discussion

Assessments of the model performance using AUROC, AUPR, and CORR indicate that the methods studied significantly differed in their performance.

A. Comparison of Methods

Though not dominant on any given species dataset, HDDT was shown in our evaluations to be the most stable classifier according to AUROC. That is, its average AUROC measure was the highest of all methods evaluated. MAXENT performance tended to perform best on datasets with relatively high levels of imbalance, while C4.5 perform best on datasets with relatively low levels of imbalance. However, MAXENT performance was significantly lower than HDDT on several datasets with low imbalance, while C4.5 showed similar characteristics on several datasets with high imbalance. According, though it never exceeded the highest performer on any given dataset, HDDT produced the most balanced performance across all of the datasets, performing well on datasets with both relatively high and relatively low imbalance.

As our results show, HDDT is handily the dominant performer with respect to AUPR. It outperforms the other methods—often by large amounts—on 7 of the 9 datasets,

Table III. Mean AUROC per method.

Species	**Mean AUROC**							
	MAXENT	HDDT	C4.5	LR	NB	RF	RF-SMT	CART
V. bellii	**0.909**	0.904	0.895	0.860	0.818	0.786	0.804	0.756
V. cassinii	0.940	0.937	**0.944**	0.930	0.910	0.878	0.828	0.819
V. flavifrons	0.898	0.918	**0.923**	0.888	0.886	0.853	0.841	0.812
V. griseus	0.922	0.946	**0.948**	0.938	0.935	0.847	0.834	0.848
V. huttoni	**0.972**	0.969	0.970	0.954	0.944	0.868	0.848	0.721
V. olivaceus	0.842	0.890	**0.891**	0.854	0.850	0.759	0.724	0.783
V. philadelphicus	**0.866**	0.851	0.853	0.820	0.805	0.783	0.782	0.750
V. solitarius	0.849	0.875	**0.880**	0.834	0.831	0.801	0.786	0.764
V. vicinior	**0.969**	0.951	0.904	0.947	0.921	0.714	0.755	0.776
Average	0.907	**0.916**	0.912	0.892	0.878	0.810	0.800	0.781

Table IV. Mean AUPR per method.

Species	**Mean AUPR**							
	MAXENT	HDDT	C4.5	LR	NB	RF	RF-SMT	CART
V. bellii	0.185	**0.212**	0.151	0.101	0.050	0.161	0.149	0.100
V. cassinii	0.320	**0.417**	0.321	0.343	0.062	0.248	0.224	0.150
V. flavifrons	0.258	**0.445**	0.347	0.330	0.147	0.325	0.293	0.197
V. griseus	0.346	**0.564**	0.412	0.450	0.132	0.342	0.315	0.265
V. huttoni	0.430	**0.512**	0.326	0.351	0.035	0.298	0.236	0.144
V. olivaceus	0.400	**0.604**	0.080	0.424	0.201	0.346	0.295	0.233
V. philadelphicus	**0.210**	0.192	0.204	0.153	0.094	0.180	0.174	0.128
V. solitarius	0.396	**0.468**	0.320	0.365	0.171	0.316	0.283	0.186
V. vicinior	**0.163**	0.107	0.086	0.077	0.011	0.085	0.100	0.047
Average	0.301	**0.391**	0.250	0.288	0.100	0.256	0.230	0.161

Table V. Mean CORR per method.

Species	**Mean CORR**							
	MAXENT	HDDT	C4.5	LR	NB	RF	RF-SMT	CART
V. bellii	0.291	0.282	0.294	0.205	0.063	**0.306**	0.291	0.191
V. cassinii	0.428	0.447	0.431	**0.461**	0.070	0.400	0.373	0.243
V. flavifrons	0.379	**0.460**	0.455	0.443	0.134	0.443	0.411	0.256
V. griseus	0.453	0.552	0.504	**0.569**	0.100	0.455	0.422	0.338
V. huttoni	0.524	**0.531**	0.472	0.489	0.043	0.464	0.404	0.275
V. olivaceus	0.442	**0.535**	0.391	0.481	0.092	0.367	0.302	0.177
V. philadelphicus	0.303	0.260	**0.314**	0.267	0.141	0.303	0.301	0.214
V. solitarius	0.428	**0.454**	0.386	0.427	0.121	0.395	0.362	0.196
V. vicinior	**0.253**	0.205	0.224	0.200	0.032	0.208	0.236	0.134
Average	0.389	**0.414**	0.386	0.394	0.088	0.371	0.345	0.225

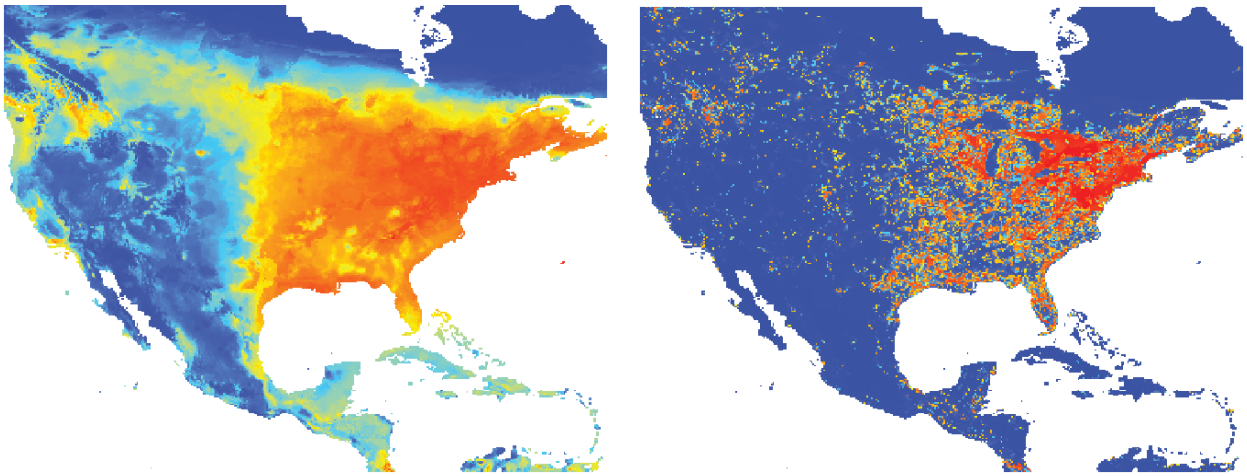

Figure 1. *V. olivaceus* distribution predicted by MAXENT (left) and HDDT (right).

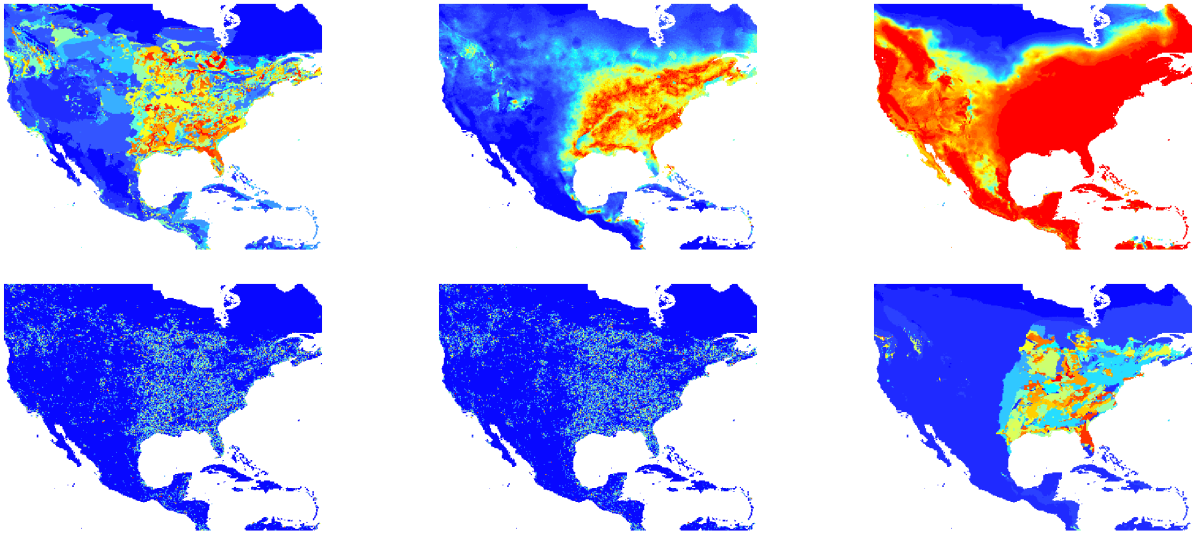

Figure 2. *V. olivaceus* distribution predicted by the other six evaluated models: C4.5 (top left), LR (top center), NB (top right), RF (bottom left), RF-SMT (bottom center), and CART (bottom right).

and boasts the highest average AUPR by a large margin. The dominant performance of HDDT with respect to AUPR is particularly important, as AUPR may a useful metric for evaluating imbalanced species distribution models. In particular, the AUPR scores reflects negatively upon classifiers that generate an imbalance between correctly predicted occurrences—true positives—and incorrectly predicted occurrences—false positives. When dealing with imbalanced datasets, classifiers that are sensitive to this imbalance may capture true positives at the cost of a proportionally high number of false positives, resulting in a lower AUPR score than a classifier that attempts to balance both measures. In general, if we consider only the top-k classifier predictions, we still find that MAXENT generates more false positive predictions than HDDT, resulting in a consistently lower AUPR score.

HDDT also performs well with respect to the CORR metric. While both AUROC and AUPR performance on individual datasets tended to be dominated by one or a few models, several different models dominate CORR on different datasets. That said, HDDT produced the highest CORR on 4 of the 9 datasets, and also demonstrated the highest average CORR.

Though the success that HDDT demonstrates with both the AUPR and CORR metrics suggests that the model is capturing useful information regarding the minority class, it is important that the method produces distribution maps that properly model the species in question.

The distributions produced by MAXENT and HDDT are strikingly different. HDDT is extremely conservative in its predictions, generally giving high probabilities only to a small region of the study area. This results in strong predictions only around individual occurrence localities. This may also explain its high AUPR performance, as by making fewer predictions than MAXENT, HDDT is less likely to produce false positive predictions, which are assigned a greater penalty by AUPR than by AUROC. In contrast, MAXENT is less particular, covering a large region of the study area with moderate probabilities.

Although the difference in predicted area might stem from predictions that are scaled in different ways, the variation in AUROC suggests actual differences between the methods. The conservative predictions allow HDDT to properly exclude regions of the Great Lakes, which MAXENT includes in its prediction. Additionally, as MAXENT tends to produce more false positive predictions than HDDT, it ultimately receives a lower AUPR score. However, as there is ambiguity as to which points represent true absence (and which represent assumed absence), we conjecture that AUPR may underpredict the true classifier performance.

B. Broad Patterns

Both MAXENT and HDDT performed relatively well according to all three evaluation measures. C4.5 and LR demonstrated slightly lower performance, followed by RF, RF-SMT, and CART. NB showed intermediate performance for AUROC, but the lowest AUPR and CORR performance. For most methods, predictive performance did not vary consistently with number of presence records available for modeling.

In general, we find that C4.5 tended to show the best AUROC performance on the datasets with the most occurrences per area (and hence the least imbalanced). However, when evaluated with respect to AUPR, C4.5 tended to perform poorly. Similarly, NB performed relatively well according to AUROC, yet was among the lowest performers with respect to AUPR and CORR. In contrast, the models that performed well according to AUPR also tended to perform well according to AUROC and CORR.

We find that, altogether, MAXENT and HDDT tend to outperform the other methods, though C4.5 occasionally boasts the highest AUROC. However, C4.5 tended to produce a high AUROC on datasets with low class imbalance, indicating that C4.5 is significantly skew-sensitive. Conversely, MAXENT demonstrated the highest AUROC for nearly all of the small, significantly imbalanced datasets. HDDT produced the most stable performance, when averaged over all of the datasets.

As AUROC is a discrimination metric that represents the likelihood that a presence will have a higher predicted value than an absence regardless of how well the predictions fit the data, it is possible that a poorly fitted model (overestimating or underestimating all the predictions) has a good discriminatory power. It is also possible that a well-fitted model has poor discrimination if, for example, probabilities for presences are only moderately higher than those for absences. In this way, it is possible for C4.5, which may not properly model any imbalance or incompleteness in the distribution, to nonetheless discriminate between presences and absences relatively well. In contrast, MAXENT, which attempts to model the occurrence localities as samples in a statistical distribution, may successfully model the imbalance and incompleteness of a species distribution while providing less discriminatory power. That is, it models the distribution more accurately, but is less accurate at properly distinguishing between points in that distribution.

This provides evidence of a fundamental drawback of AUROC in this domain: that it can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution. As species data are almost universally imbalanced, due typically to a relatively large study area with a comparatively few presence localities and incomplete sampling, this limitation of the metric is invariably extenuated. To partly address this issue, we propose the use of the AUPR metric to supplement the use of AUROC. AUPR can better capture this imbalance, allowing it to provide a more informative picture of an algorithm's performance under extreme imbalance. HDDT generally sees the highest boost in performance on datasets with less imbalance.

VI. Conclusions

We can draw two major conclusions from our results. First, the precision-recall metric can be a useful tool for evaluating species distribution models. It is a skew-sensitive metric that appears to have discriminatory power, and can complement the prevalent use of AUROC. Second, the HDDT model, which has been effective when used on imbalanced datasets in other domains, generally modeled species with performance competitive with MAXENT, an established species distribution model. Though HDDT originated in another discipline and has had little exposure in ecological analysis, it appears to offer considerable promise across a much broader range of species data, providing an exciting avenue for future research.

Class imbalance continues to be a significant problem for species distribution modeling, as it pertains to both the development and evaluation of these models. Though AUROC serves as a touchstone in the evaluation of species distribution models, it has commonly been assailed by allegations of theoretical limitations. Our experiments explore the utility of a similar metric, AUPR, finding that it may be capturing information ignored by AUROC. In turn, we suggest that AUROC and AUPR may provide a more informative measure of model utility when used together.

REFERENCES

[1] L. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Arxiv preprint arXiv:1106.1813*, 2011.

[4] D. Cieslak and N. Chawla. Learning decision trees for unbalanced data. *Machine Learning and Knowledge Discovery in Databases*, pages 241–256, 2008.

[5] D.A. Cieslak, T.R. Hoens, N.V. Chawla, and W.P. Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, pages 1–23, 2012.

[6] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[7] Chris Drummond and Robert C. Holte. Explicitly representing expected cost: an alternative to ROC representation. *KDD*, pages 198–207, 2000.

[8] J. Elith, C.H. Graham*, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, et al. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[9] J. Elith, C.H. Graham*, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

[10] P.A. Flach. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 194, 2003.

[11] RJ Hijmans, SE Cameron, JL Parra, PG Jones, and A. Jarvis. The worldclim interpolated global terrestrial climate surfaces. version 1.3. *Computer program available at website http://biogeo. berkeley. edu/[accessed April, 2006]*, 2004.

[12] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[13] Stanley Kok and Pedro Domingos. Learning the structure of Markov logic networks. In *ICML*, pages 441–448, New York, NY, USA, 2005. ACM.

[14] T.C.W. Landgrebe, P. Paclik, R.P.W. Duin, and A.P. Bradley. Precision-recall operating characteristic (p-roc) curves in imprecise environments. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 123–127. IEEE, 2006.

[15] S.J. Phillips, M. Dudík, and R.E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM, 2004.

[16] Foster J. Provost, Tom Fawcett, and Ron Kohavi. The Case Against Accuracy Estimation for Comparing Induction Algorithms. *ICML*, 1998.

[17] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[18] R. Raina, Y. Shen, A.Y. Ng, and A. McCallum. Classification with hybrid generative/discriminative models. *Advances in neural information processing systems*, 16, 2003.

[19] C.R. Rao. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, 19(1):23–63, 1995.

[20] R.F. Tate. Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, 25(3):603–607, 1954.