Expert Systems with Applications 38 (2011) 5268–5275

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Comparing machine learning classifiers in potential distribution modelling

Ana C. Lorena [a,*], Luis F.O. Jacintho [a], Marinez F. Siqueira [b], Renato De Giovanni [b], Lúcia G. Lohmann [c], André C.P.L.F. de Carvalho [d], Missae Yamamoto [e]

[a] CMCC – Universidade Federal do ABC, Santo André, SP, Brazil
[b] Centro de Referência em Informação Ambiental (CRIA), Campinas, SP, Brazil
[c] Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil
[d] ICMC, Universidade de São Paulo, São Carlos, SP, Brazil
[e] Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brazil

ARTICLE INFO

Keywords:
Ecological niche modelling
Potential distribution modelling
Machine learning

ABSTRACT

Species' potential distribution modelling consists of building a representation of the fundamental ecological requirements of a species from biotic and abiotic conditions where the species is known to occur. Such models can be valuable tools to understand the biogeography of species and to support the prediction of its presence/absence considering a particular environment scenario. This paper investigates the use of different supervised machine learning techniques to model the potential distribution of 35 plant species from Latin America. Each technique was able to extract a different representation of the relations between the environmental conditions and the distribution profile of the species. The experimental results highlight the good performance of random trees classifiers, indicating this particular technique as a promising candidate for modelling species' potential distribution.

1. Introduction

Potential distribution models are generated by retrieving the environmental conditions where a species is known to be present or absent and then providing these data as input to a modelling algorithm. This process allows to better understand the influence and relationship of environmental conditions in the distribution of species.

Among the ecological analysis currently carried out with the use of potential distribution models, it is possible to mention:

- Indicating priority areas for environmental conservation (Araújo, Williams, & Reginster, 2000; Loiselle, 2003; Ortega-Huerta & Peterson, 2004);
- Evaluating the risk of harmful proliferation of invasive species (Higgins, Richardson, Cowling, & Trinder-Smith, 1999; Peterson, 2003; Peterson, Papes, & Kluza, 2003; Thuiller et al., 2005; Williams, Hahs, & Morgan, 2008);
- Studying the impact of environmental changes in biodiversity (Pearson, Dawson, Berry, & Harrison, 2002; Peterson, Benz, & Papes, 2007; Peterson, Lash, Carroll, & Johnson, 2006; Peterson et al., 2002);
- Indicating the course of diseases spreading (Berry, 2002; Hannah et al., 2007; Hannah, Midgley, Hughes, & Bomhard, 2005).

Machine learning (ML) is a research area with roots in Artificial Intelligence and Statistics concerned with the development of techniques that can extract knowledge from datasets (Mitchell, 1997). This knowledge is represented in the form of a model, providing a compact description of the given data and allowing predictions for new data. ML algorithms are pointed as promising tools in modelling and prediction of species distribution (Elith et al., 2006).

This paper compares different supervised ML techniques by modelling the potential distribution of 35 Latin American plant species. This study aimed to evaluate the accuracy performance of a set of ML classifiers for possible future inclusion of their corresponding algorithms in *openModeller*,[1] an open source framework for potential distribution modelling. The techniques were statistically compared in a controlled set of experiments using diverse datasets.

This paper is organized as follows: Section 2 presents related work in potential distribution modelling. Section 3 presents the ML techniques employed in the experiments. Section 4 describes the datasets used. Section 5 shows and discusses the results obtained. Finally, Section 6 concludes this paper and presents future research directions.

* Corresponding author.
E-mail addresses: ana.lorena@ufabc.edu.br (A.C. Lorena), luis.jacintho@ufabc.edu.br (L.F.O. Jacintho).

[1] http://openModeller.cria.org.br.

0957-4174 © 2010 Elsevier Ltd. Open access under the Elsevier OA license.
doi:10.1016/j.eswa.2010.10.031

2. Potential distribution modelling

Potential distribution modelling has proven valuable for generating biogeographical information that can be applied in a broad range of fields, including conservation biology, ecology and evolutionary biology (Pearson, 2007). Lately, it became an important component of conservation planning, and a wide variety of modelling techniques have been developed for this purpose (Guisan & Thuiller, 2005). Some common uses of species' distribution models in conservation biology are: guiding field surveys to find populations of known species (Bourg, McShea, & Gill, 2005; Guisan et al., 2006); species' delimitation (Raxworthy, Ingram, Rabibosa, & Pearson, 2007); predicting species invasion (Higgins et al., 1999; Peterson, 2003; Peterson et al., 2003; Thuiller et al., 2005; Williams et al., 2008); exploring speciation mechanisms (Graham, Ron, Santos, Schneider, & Moritz, 2004; Kozak & Wiens, 2006); supporting conservation prioritization and reserve selection (Araújo et al., 2000; Ferrier, 2002; Leathwick, 2005; Loiselle, 2003; Ortega-Huerta & Peterson, 2004); testing ecological theory (Anderson, Laverde, & Peterson, 2002; Graham, Moritz, & Williams, 2006; Hugall, 2002; Peterson, Sobern, & Sanchez-Cordero, 1999); comparing paleodistributions and phylogeography (Hugall, 2002); reintroduction of endangered species (Pearce & Lindenmayer, 1998); detecting high species density (hot spots) important for conservation (Nelson & Boots, 2008); assessing disease risks (Peterson et al., 2007; Peterson et al., 2006); projecting potential impacts of climate change (Berry, 2002; Hannah et al., 2007; Hannah et al., 2005; Pearson et al., 2002; Peterson et al., 2002) and others.

Models to predict species' potential distributions are built by combining two kinds of data: species occurrence data and environmental data. Occurrence data are coordinates (pairs of longitude and latitude in a certain reference system) where the species was observed or collected. Environmental data is provided as a set of georeferenced rasters associated with variables that are known to influence the species distribution. The set of rasters is previously selected by a species expert or by generic pre-analysis tools. The main idea is that the spatial distribution of suitable environments for the species can be estimated from a selected "region of study" (Pearson, 2007).

In general, species occurrence data are significantly unbalanced in terms of the proportion of presence and absence data (Elith et al., 2006). In fact, it is common to have no examples of the absence class in these datasets – that is the case of the datasets employed in this paper. This happens because it is easier and more usual for specialists to record the presence of a species, resulting in few or no absence data. A strategy frequently used to overcome this limitation is to generate "pseudo-absence" points (Stockwell & Peters, 1999). In this paper the pseudo-absence points were generated using an algorithm which minimizes the risk of generating pseudo-absences in regions that are suitable for the species (described in Section 4).

The first step in the modelling process is to find the corresponding environmental conditions associated with each species occurrence point. The result from this step is a set of vector data which can used as input to a modelling algorithm. The algorithm tries to identify environmental conditions that are suitable for the species to survive and maintain populations.

Historically, a number of modelling algorithms have been applied to express the probability of a species to survive as a function of a set of environmental variables. The task is to identify potentially complex non-linear relationships in a multi-dimensional environmental space (Pearson, 2007). For a further conceptual discussion of these approaches see Sobern (2007).

Various studies have demonstrated that different modelling approaches have the potential to yield substantially different predictions (Brotons, 2004; Elith et al., 2006; Pearson et al., 2006; Segurado & Araújo, 2004; Thuiller, 2003; Thuiller et al., 2004). The most comprehensive model comparison of several modelling algorithms was provided by Elith et al. (2006). The present work focus on the comparison of the predictive behavior of nine ML classification techniques.

3. Machine learning techniques

Machine Learning (ML) techniques employ an inference principle named induction, in which general conclusions are obtained from a particular set of examples. One of the main approaches for induction is supervised learning. In supervised learning, the knowledge about the problem being modelled is presented by datasets composed of pairs in the form: input, desired output (Mitchell, 1997). The ML algorithm extracts the knowledge representation from these examples so that it can produce correct outputs for new inputs.

Therefore, given a dataset with n examples in the form $(\mathbf{x}_i, y_i)$, in which $\mathbf{x}_i$ represents an input and y_i denotes its label, a classifier, also named model, predictor or hypothesis, will be produced in a process named training. The obtained classifier can be regarded as a function f, which receives an input $\mathbf{x}$ and provides an output prediction y. This model also provides a compact description of the training data.

Next sections present a brief introduction to the ML techniques used in this work. Each technique employs a different approach or bias in the extraction of concepts from data and their choice was oriented toward promising representatives of different learning paradigms.

3.1. RIPPER

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a propositional rule learner (Cohen, 1995) based on the Incremental Reduced Error Pruning (IREP) algorithm (Furnkranz & Widmer, 1994). It produces a set of rules, one at a time, through two steps: growth and pruning.

In the growth phase, the rules are initially empty and conditions are added to the rules in order to maximize their coverage of the training data. To prevent the produced rules from overfitting, situation where they become too specific for the training data, the pruning step eliminates conditions from the rules that do not harm their classification accuracy. Substitution and reviewed rules are also created for each original rule of the training set. The best accuracy rules are maintained. This process is repeated until all classes are covered by the algorithm.

RIPPER models have the advantage of being easily interpretable, since the knowledge extracted from data is explicitly represented by a set of symbolic rules. Besides, they are flexible and incremental. New rules can be added or modified easily for new data.

3.2. GARP

Genetic Algorithm for Rule Set Production (GARP) (Stockwell & Peters, 1999) is an algorithm that was specifically developed in the context of ecological niche modelling. A GARP model consists of a set of mathematical rules based on environment conditions. Given a specific environment condition, if all rules are satisfied, the model predicts presence of the species. Four types of rules are possible: atomic, logistic regression, bioclimatic envelope, and negated bioclimatic envelope.

Each set of rules is an individual of a population in the typical context of a genetic algorithm (Mitchell, 1999). GARP generates an initial population which is evaluated in each iteration to check if the problem converged to a solution. Genetic operators (join,

cross-over and mutation) are applied in the beginning of each new iteration on the best individuals from previous steps, producing a new population of solutions.

GARP is therefore a non-deterministic algorithm that produces boolean responses (present/absent) for each different environment condition. This experiment used the GARP Best Subsets technique, which selects the 10 best GARP models out of 100 models. By aggregating 10 GARP models, a GARP Best Subsets model produces more outputs than just a boolean response. The probability of presence in a specific condition is proportional to the number of GARP models that predicts presence for the species.

3.3. Decision trees

Decision trees (DTs) organize the knowledge extracted from data in a recursive hierarchical structure composed of nodes and branches (Quinlan, 1986). Each internal node represents an attribute and is associated to a test relevant for data classification. Leaf nodes of the tree correspond to classes. Branches represent each of the possible results of the applied tests. A new example can be classified following the nodes and branches accordingly until a leaf node is reached.

The DT induction process aims to maximize the correct classification of all training data. To avoid overfitting, a pruning phase is usually applied to the trained tree. It prunes ramifications with low expressive power according to some criterion, like the expected error rate (Quilan, 1988).

One advantage of DTs is the comprehensibility of the classification structures generated. For each new data, it is possible to verify which attributes determined the final classification. DTs may be, nevertheless, not robust to inputs with high dimensions (with a large number of attributes).

3.4. Random trees

Random Forests (RFs) are combinations of tree predictors (Breiman, 2001). Each tree votes for its preferred class and the most voted class gives the final prediction.

Let T be a training dataset with n data items and where each item has m attributes. For each tree, a new training dataset T' is built by sampling T at random with replacement (bootstrap sampling). To determine a node split in the tree, a subset $m' \ll m$ of the attributes is chosen at random. The best split of these selected attributes is then used. The trees are grown in order to classify all data items from T' correctly and there is no pruning. The value m' can be chosen based on an out-of-bag error rate estimate.

RFs have been successful in a wide range of applications and are fast to train. Breiman (2001) also showed that RFs do not overfit, despite the number of trees employed in the combination.

3.5. k-Nearest neighbor

The k-Nearest Neighbor (kNN) algorithm is the simplest representative of instance-based ML techniques. It stores all training data and classifies a new data point according to the class of the majority of its k nearest neighbors in the given dataset. To obtain the nearest neighbors for each data, kNN uses a measure to compute the distance between pairs of data items. In general, the measure employed is the Euclidean distance.

kNN is able to build local approximations of the objective function, different for each new data point being classified. This characteristic may be advantageous when the objective function is complex, but may be described by several local approximations of low complexity. Another advantage of kNN is its simplicity. Nevertheless, prediction times are usually costly, since all training data must be revisited.

3.6. Naïve–Bayes

Naïve Bayes (NB) are probabilistic classifiers based on the Bayes theorem for conditional probabilities.

It builds a function, to be optimized, using a narrow (naïve) assumption that all attributes in a dataset are independent. Therefore, it assumes that the presence/absence of a characteristic describing a certain class is unrelated to the presence/absence of any other characteristic, which is not true for the majority of classification tasks. NB training is usually performed through the use of maximum likelihood algorithms.

Despite its simplicity, NB have been successful in complex practical applications, specially in text mining (McCallum & Nigam, 1998). It also shows low train and prediction times.

3.7. Logistic regression

Logistic Regression (LR) classifiers are statistical models in which a logistic curve is fitted to the dataset (Kleinbaum & Klein, 2005), modelling the probability of occurrence of a class. LR classifiers are also known as: logistic model, logit model and maximum-entropy classifiers.

The first step in LR consists of building a logit variable, containing the natural log of the odds of the class occurring or not. A maximum likelihood estimation algorithm is then applied to estimate the probabilities.

LR models are largely employed in Statistics and have demonstrated success in several real-world problems.

3.8. Support vector machines

Support Vector Machines (SVMs) are based on concepts from the Statistical Learning Theory (Vapnik, 1995). Given a dataset T composed of n pairs $(\mathbf{x}_i, y_i)$, in which $\mathbf{x}_i \in \Re^m$ and $y_i \in \{-1, +1\}$, SVMs seek for a hyper plane $\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}) + b = 0$ able to separate the data in T with minimum error maximizing the margin of separation between the classes. In this equation, $\mathbf{\Phi}$ represents a mapping function that maps the data in T to a space of higher dimension, such that the classes become linearly separable.

In SVM training and predictions, the mapping function appears as dot products in the form $\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j)$, which can be efficiently computed by Kernel functions, usually simpler than the mapping function. Some of the most used Kernel functions are the Gaussian or RBF (Radial-Basis Function) functions.

SVMs have good generalization ability. Besides, SVMs also stand out for their robustness to high dimensional data. Their main deficiency concerns the difficulty of interpreting the generated model and their sensibility to a proper parameter tuning.

3.9. Artificial neural networks

Artificial Neural Networks (ANNs) are computational systems based on the structure, processing method and learning ability of brain (Haykin, 1998). They are composed of simple processing units, which simulate the biological neurons. These artificial neurons, also named nodes, are disposed in one or more layers. Each node is connected to one or more nodes through weighted connections, which simulate the biological synapses. In this work Multilayer Perceptron (MLP) ANNs were employed.

The representation and knowledge about data is acquired and stored in an ANN by adjusting connections' weights. There are several algorithms for ANN training. They usually try to adjust the ANN weights to approximate the outputs of the ANN to the desired outputs known for training data. The algorithm named back-propagation is based on this error correcting concept.

In general, among the advantages of ANNs are their robustness to noisy data and their ability to represent linear and non-linear functions of various forms and complexities. Disadvantages include the need of parameter tuning and the difficulty in interpreting the concepts learned by the ANN, which are codified in the weights.

4. Biotic and abiotic data

Nine environmental layers were used in this study, all continuous-valued attributes, four of them were climatic variables and the other five topographic. The climatic variables were: mean temperature of wettest quarter, mean temperature of driest quarter, precipitation of wettest quarter, and precipitation of driest quarter. These datasets came from Worldclim (Hijmans, 2005), a set of global climate layers (climate grids) generated through the interpolation of average monthly climate data from weather stations on a 30 arc-second resolution. The topographic layers used were: elevation, slope, aspect and flow water direction and accumulation. These data came from HYDRO1k, a geographic database developed to provide comprehensive and consistent global coverage of topographically derived data from the USGS 30 arc-second digital elevation model of the world (Verdin & Greenlee, 1996).

Regarding biotic data, 35 plant species (*Bignoniaceae*), totalizing 3507 spatial points, were used in the modelling process (Fig. 1). All records came from specimens deposited in herbaria and were part of a recent taxonomic revision. Most records were georeferenced at municipality level based on location data from specimen labels. For this reason, all layers were resampled on a 10 min spatial resolution ($18.6 \times 18.6 = 344$ km^2 at the equator). Fig. 1(a) shows a table with the datasets used in the experiments, along with the number of points available for each species (♯Data column) and a "Number" column by which the each of the datasets will be further referred in the paper. Fig. 1(b) plots some points where the species data were sampled.

Since the algorithms employed in this paper required presence and absence data, pseudo-absences were generated in the same number of presences for each species. To minimize the risk of generating pseudo-absences in regions that are suitable for the species, they were randomly generated outside the inner Bioclim (Nix, 1986) envelope. Therefore, a Bioclim model was generated for each species based on all corresponding presence points. As can be seen on Fig. 2(a), Bioclim divides the environmental space in three regions: Suitable (inner envelope), Marginal and Unsuitable. In this work, suitable envelopes were calculated using a 0.95 cutoff. Therefore, pseudo-absences were always generated in marginal or unsuitable areas. Fig. 2(b) illustrates pseudo-absences generated outside the inner Bioclim envelope in a bidimensional space.

Therefore, each dataset described in Fig. 1 has additional pseudo-absence data, generated in the same number as the presence data, in order to avoid a class unbalance which could harm the ML induction algorithms. In total, each dataset has the double of data items from those presented in Fig. 1(a), 50% from class presence and 50% from class absence for each species.

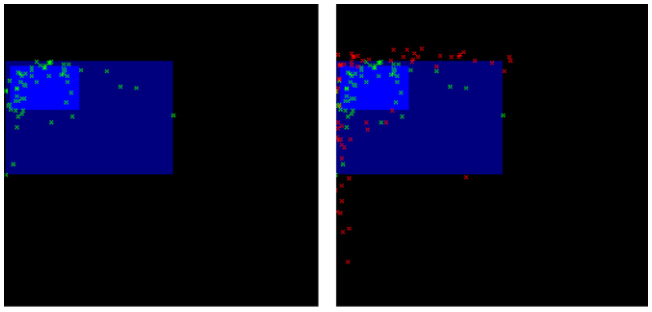
It must be noticed that Latin America has a vast biodiversity distributed in wide areas. As consequence, ecological monitoring atten-

Number	Dataset	♯Data
1	*Adenocalymma cladotrichum*	193
2	*Adenocalymma impressum*	124
3	*Adenocalymma purpurascens*	79
4	*Adenocalymma schomburgkii*	102
5	*Amphilophium aschersonii*	63
6	*Amphilophium granulosum*	60
7	*Anemopaegma karstenii*	90
8	*Anemopaegma paraense*	93
9	*Bignonia bracteomana*	107
10	*Bignonia lilacina*	91
11	*Bignonia nocturna*	153
12	*Bignonia prieurei*	76
13	*Bignonia uleana*	77
14	*Fridericia bracteolata*	106
15	*Fridericia cinnamomea*	89
16	*Fridericia japurensis*	115
17	*Fridericia nigrescens*	90
18	*Fridericia pearcei*	73
19	*Fridericia spicata*	99
20	*Fridericia trailii*	94
21	*Fridericia tuberculata*	96
22	*Lundia densiflora*	144
23	*Lundia spruceana*	88
24	*Manaosella cordifolia*	61
25	*Mansoa alliacea*	92
26	*Pleonotoma clematis*	82
27	*Pleonotoma jasminifolia*	90
28	*Pleonotoma melioides*	106
29	*Pyrostegia dichotoma*	179
30	*Tanaecium bilabiatum*	133
31	*Tanaecium xanthophyllum*	104
32	*Tynanthus polyanthus*	129
33	*Tynanthus schumannianus*	84
34	*Xylophragma platyphyllum*	67
35	*Xylophragma pratense*	78

(a) Datasets description

(b) Plots of species data

Fig. 1. Datasets employed.



(a) Bioclim envelopes based on presence data (in green)

(b) Pseudo-absence data (in red) based on Bioclim envelopes

Fig. 2. Generating pseudo-absence data.

dance for recording species' data can be considered costly and difficult. The resulting large data gaps can clearly benefit from the use of modelling techniques, which can help monitoring, preserving and developing strategies for the sustainable use of natural resources.

5. Experiments

The ML techniques from Section 3 were compared in a controlled set of experiments aiming to investigate their effectiveness in potential distribution modelling. This comparison will also guide future inclusion of new modelling algorithms in the *openModeller* framework. Only two of the ML algorithms tested in this paper are already available in *openModeller*: GARP and SVMs.

5.1. Experimental protocol

In order to improve performance estimate of the algorithms used in this paper, the datasets described in Section 4 were divided with the 10-fold stratified cross-validation methodology. Accordingly, each dataset was divided into 10 subsets of approximately equal size, with 50% of presence data and 50% of absence data. For each ML technique, the examples from 9 folds were then used to train a classifier, which was evaluated in the remaining fold. This process was repeated 10 times, using at each cycle a different fold for test. The performance of each classifier was given by the average of the performances observed in the test folds. The AUC (Area Under the ROC Curve) was used to evaluate the classifiers effectiveness in the classification of presence/absence data.

Experiments with SVMs were performed with the LibSVM tool (Chang & Lin, 2004) currently available from *openModeller*. GARP models were generated with the new GARP Best Subsets implementation available in *openModeller*. All other ML classifiers tested were induced with the Weka tool (Witten & Frank, 2005), which is a free collection of ML algorithms. For all techniques employed, default parameter values were used to allow a fair comparison of the different techniques.

After calculating AUC performances for each ML technique in the 35 datasets, a comparison approach suggested in Demsar (2006) was followed. This comparison considers the performance of the different techniques in multiple datasets as a whole. In this approach, a ranking matrix is built containing the datasets as rows and the techniques as columns. In this matrix, an element r_{ij} represents the rank of technique j performance in dataset i. Row by row, a rank of the algorithms performance is made. The highest AUC technique receives the first place in the rank, the second higher AUC technique receives the second place and so on. If a tie occurs, the authors looked for the standard deviation values. This approach is not followed originally in Demsar (2006), but it can be considered adequate since lower standard deviation values indicate more stability of the algorithm regarding different partitions of the dataset. Therefore, to solve the average AUC tie in a given rank position, the lower standard deviation technique is chosen and the other technique is given the next rank position. If the tie remains, an average ranking is computed. For example, if there is a tie between two techniques in position three, they are given the rank 3.5 (the mean of the third and fourth positions). After all rankings have been computed, the averages of the techniques ranks are calculated. If a technique receives, for example, an average rank of 1.3, this means that, in average, it was in this rank in all datasets. The authors also stored the standard-deviation of these ranks to verify the stability of the rank performance of the algorithms.

A statistical test for multiple comparisons is applied to verify at 95% of confidence which technique(s) outperformed their counterparts considering their average ranks. This test is based on the Friedman test (Friedman, 1937; Iman & Davenport, 1980) with the Nemenyi post-test (Nemenyi, 1963). Details of this test can be consulted in Demsar (2006).

5.2. Results

The mean and standard deviation AUC values of each classification technique on the 35 datasets were recorded and are presented in Table 1. They correspond to the mean of the AUC measures obtained following the 10-fold cross-validation methodology. Their standard-deviation rates are indicated in parenthesis. The best AUC obtained for each dataset is highlighted in boldface and the worst in italics. The last row of this table shows the average and standard deviation of the ranks of the techniques performance in all datasets.

The statistical comparison of the techniques is summarized in Fig. 3. This figure presents a scale of the techniques ranking. Best performing techniques lay in the right of the scale, while worst performing techniques are in the left of the scale. CD corresponds to the critical difference interval of the test. If the ranks of two techniques differ by at most CD, their results can be considered statistically similar at 95% of confidence. Those techniques with similar statistical results are joined in Fig. 3 by a thick horizontal line. Therefore, the overall results of SVMs and RFs, for example, can be considered statistically similar, while all other techniques were outperformed by RF.

5.3. Discussion

It is worth noting from Fig. 3 that RF was the best performing technique, statistically outperforming all other ML techniques, except SVMs. SVMs results were also comparable to those of ANNs. kNN, RIPPER, DTs and NB were the worst performing classifiers. ANNs, GARP and LR had an average performance, although nothing can be really said about them from the statistical test performed.

In fact, RF had a remarkable performance in all datasets. Its AUC was the best in 29 out of the 35 datasets employed. Even in datasets where its AUC was not the highest, it was close to the best ones. There were also datasets for which the results of RF were quite superior to those of other techniques, as in datasets 2 and 10, where there was a 10% AUC improvement of RF over the second best performing techniques. The low standard deviation value of RF ranking support the observation that this technique was constantly in the best performing positions in the comparisons made.

On the other hand, the results of other ML techniques were not so stable. This was more remarkable for LR, which was the best performing technique in three datasets, but also the worst performing technique in six datasets. This shows that the LR results were not stable along different datasets.

Table 1
AUC results.

Dataset	SVM	GARP	DT	RIPPER	KNN	Logistic	ANN	NaiveBayes	RandomForest
1	0.94 (0.03)	*0.85* (*0.04*)	0.88 (0.08)	0.89 (0.06)	0.88 (0.06)	0.93 (0.04)	0.92(0.05)	0.90(0.04)	**0.95** (**0.03**)
2	0.86 (0.07)	0.81 (0.05)	0.85 (0.12)	0.86 (0.10)	0.81 (0.11)	*0.79* (*0.08*)	0.86 (0.08)	0.82 (0.06)	**0.96** (**0.06**)
3	0.79 (0.10)	0.80 (0.10)	0.85 (0.05)	0.81 (0.11)	0.85 (0.06)	0.80 (0.10)	*0.78* (*0.09*)	0.79 (0.11)	**0.92** (**0.06**)
4	0.90 (0.08)	0.84 (0.10)	*0.82* (*0.09*)	0.83 (0.08)	0.83 (0.06)	0.91 (0.06)	0.90 (0.08)	0.92 (0.07)	**0.93** (**0.06**)
5	0.88 (0.06)	0.86 (0.07)	*0.79* (*0.15*)	0.87 (0.11)	0.83 (0.07)	0.90 (0.09)	**0.90** (**0.07**)	0.87 (0.10)	0.89 (0.05)
6	0.91 (0.07)	0.89 (0.06)	0.90 (0.08)	*0.84* (*0.10*)	0.88 (0.09)	0.89 (0.09)	0.96 (0.07)	0.86 (0.11)	**0.96** (**0.05**)
7	0.89 (0.06)	0.79 (0.06)	0.81 (0.12)	*0.77* (*0.11*)	0.82 (0.07)	0.89 (0.03)	0.90 (0.06)	0.83 (0.09)	**0.91** (**0.06**)
8	0.90 (0.04)	0.87 (0.07)	0.84 (0.08)	0.84 (0.07)	*0.79* (*0.07*)	0.88 (0.05)	0.85 (0.06)	0.88 (0.07)	**0.91** (**0.05**)
9	0.94 (0.06)	0.94 (0.05)	0.89 (0.08)	0.88 (0.09)	*0.86* (*0.09*)	0.90 (0.06)	0.95(0.05)	0.93(0.06)	**0.96** (**0.03**)
10	0.81 (0.07)	0.82 (0.09)	0.79 (0.11)	0.77 (0.11)	0.70 (0.12)	*0.65* (*0.10*)	0.80 (0.11)	0.76 (0.06)	**0.92** (**0.08**)
11	0.84 (0.09)	0.86 (0.08)	0.81 (0.06)	*0.75* (*0.08*)	0.77 (0.06)	0.79 (0.10)	0.82 (0.07)	0.84 (0.09)	**0.90** (**0.09**)
12	**0.86** (**0.07**)	0.80 (0.06)	0.81 (0.08)	*0.74* (*0.11*)	0.81 (0.09)	0.82 (0.07)	0.82 (0.09)	0.83 (0.08)	0.85 (0.09)
13	0.91 (0.08)	0.93 (0.05)	0.87 (0.08)	*0.77* (*0.16*)	0.86 (0.09)	0.80 (0.07)	**0.95** (**0.05**)	0.87 (0.08)	0.94 (0.06)
14	0.98 (0.03)	0.94 (0.05)	0.96 (0.04)	*0.93* (*0.05*)	0.94 (0.06)	**0.98** (**0.02**)	0.96 (0.05)	0.98 (0.03)	0.98 (0.03)
15	0.79 (0.07)	0.81 (0.05)	0.76 (0.10)	0.78 (0.07)	*0.70* (*0.08*)	0.80 (0.12)	0.79 (0.11)	0.83 (0.09)	**0.86** (**0.09**)
16	0.90 (0.03)	0.85 (0.06)	0.85 (0.07)	*0.81* (*0.09*)	0.81 (0.07)	0.87 (0.06)	0.86 (0.07)	0.90 (0.05)	**0.93** (**0.05**)
17	0.89 (0.12)	0.85 (0.08)	*0.84* (*0.14*)	0.87 (0.10)	0.86 (0.09)	0.87 (0.12)	**0.93** (**0.09**)	0.86 (0.12)	0.93 (0.10)
18	0.85 (0.11)	0.83 (0.07)	0.76 (0.11)	0.82 (0.06)	*0.73* (*0.12*)	0.79 (0.12)	0.84 (0.14)	0.75 (0.12)	**0.91** (**0.08**)
19	0.83 (0.07)	0.83 (0.06)	0.82 (0.11)	*0.76* (*0.09*)	0.76 (0.10)	0.83 (0.08)	0.80 (0.09)	0.83 (0.08)	**0.90** (**0.07**)
20	0.89 (0.05)	0.82 (0.07)	*0.80* (*0.09*)	0.84 (0.05)	0.84 (0.06)	0.84 (0.06)	0.88 (0.08)	0.86 (0.08)	**0.94** (**0.03**)
21	0.90 (0.07)	0.87 (0.10)	0.83 (0.08)	*0.83* (*0.07*)	0.84 (0.07)	**0.91** (**0.07**)	0.90 (0.07)	0.88 (0.10)	0.88 (0.06)
22	0.87 (0.05)	0.83 (0.05)	*0.79* (*0.08*)	0.81 (0.07)	0.85 (0.05)	0.84 (0.04)	0.87 (0.06)	0.85 (0.06)	**0.92** (**0.04**)
23	0.93 (0.06)	0.90 (0.09)	0.87 (0.10)	0.86 (0.08)	0.79 (0.12)	*0.75* (*0.13*)	0.92 (0.05)	0.84 (0.07)	**0.95** (**0.06**)
24	0.86 (0.10)	0.85 (0.14)	0.79 (0.11)	*0.67* (*0.10*)	0.85 (0.06)	0.75 (0.13)	0.86 (0.09)	0.83 (0.12)	**0.92** (**0.04**)
25	0.90 (0.08)	0.82 (0.10)	0.81 (0.10)	*0.78* (*0.11*)	0.81 (0.09)	0.86 (0.10)	0.89 (0.06)	0.84 (0.09)	**0.91** (**0.06**)
26	0.91 (0.05)	0.82 (0.08)	0.85 (0.08)	*0.79* (*0.07*)	0.80 (0.09)	**0.92** (**0.04**)	0.88 (0.05)	0.84 (0.06)	**0.92** (**0.04**)
27	0.93 (0.07)	0.83 (0.07)	*0.82* (*0.12*)	0.83 (0.11)	0.88 (0.08)	0.91 (0.07)	0.90(0.05)	0.90 (0.06)	**0.93** (**0.06**)
28	0.81 (0.05)	0.82 (0.08)	0.80 (0.09)	0.73 (0.08)	0.70 (0.08)	*0.68* (*0.08*)	0.78 (0.09)	0.74 (0.06)	**0.88** (**0.04**)
29	0.81 (0.07)	0.81 (0.08)	0.79 (0.10)	0.81 (0.07)	*0.73* (*0.09*)	0.72 (0.06)	0.79 (0.08)	0.76 (0.08)	**0.88** (**0.07**)
30	0.89 (0.07)	*0.82* (*0.08*)	0.84 (0.11)	0.84 (0.09)	*0.82* (*0.08*)	0.90 (0.07)	0.90 (0.10)	0.88 (0.07)	**0.94** (**0.04**)
31	0.88 (0.07)	0.78 (0.07)	0.79 (0.06)	0.82 (0.08)	0.75 (0.08)	0.76 (0.09)	0.90 (0.06)	*0.74* (*0.10*)	**0.92** (**0.07**)
32	0.82 (0.06)	0.78 (0.09)	0.81 (0.08)	0.77 (0.09)	*0.77* (*0.05*)	0.86 (0.06)	0.85 (0.06)	0.80 (0.08)	**0.91** (**0.05**)
33	0.82 (0.11)	0.81 (0.08)	0.77 (0.12)	0.75 (0.08)	0.76 (0.13)	*0.64* (*0.17*)	0.82 (0.10)	0.79 (0.11)	**0.83** (**0.11**)
34	0.89 (0.11)	0.92 (0.07)	0.87 (0.16)	0.76 (0.12)	0.81 (0.09)	*0.74* (*0.14*)	0.92 (0.08)	0.90 (0.10)	**0.94** (**0.09**)
35	0.75 (0.15)	0.72 (0.14)	0.73 (0.09)	0.69 (0.12)	0.72 (0.09)	0.68 (0.13)	*0.67* (*0.10*)	0.75 (0.15)	**0.82** (**0.10**)
Rank	3.1 (1.2)	5.4 (2.1)	6.4 (1.8)	7.2 (1.9)	*7.3* (*1.5*)	5.3 (2.7)	3.8 (2.0)	5.8 (3.7)	**1.3** (**0.7**)

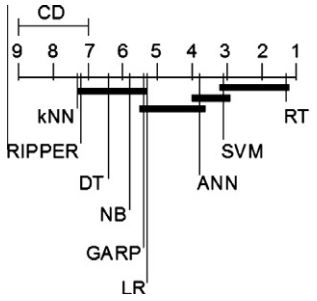


Fig. 3. Results of statistical comparison of techniques.

GARP is one of the most well-known techniques in potential distribution modelling, being specially developed for this particular task. Nevertheless, RF clearly outperformed GARP in all datasets analyzed.

The low performance of DTs and RIPPER can be attributed to the difficulty of symbolic techniques in dealing with continuous-valued attributes. However, the comprehensiveness of the models generated by these techniques can be a good argument toward their consideration in ecological analysis. Symbolic models can be examined by ecologists and support known or new knowledge about data, as performed in Lorena, Siqueira, Giovanni, Carvalho, and Prati (2008).

NB disappointing results may be attributed to its independence assumptions, clearly not real for the datasets investigated. For example, the attributes elevation and slope are clearly related to flow water direction and accumulation. Other similar relationships can be easily drawn from the other attributes employed in this work.

kNN showed the worst overall results. It would be interesting to verify if the use of other distance measures between data items would alter theses results, although a clear disadvantage of kNN in relation to the other techniques is its lack of an explicit model, which also results in a high computational cost for making predictions.

From the conducted studies, the authors were able to detect the effectiveness of three ML techniques in potential distribution modelling: RF, SVMs and ANNs, with a special emphasis to RF outstanding performance. For ANNs and SVMs, in particular, a careful parameter tuning could improve the results achieved, since these techniques are more influenced by a proper parameter adjustment, and default parameter values were used in the experiments.

Since SVMs are already implemented in the *openModeller* tool, the authors will work in the inclusion of RF and ANNs to this tool. DTs will also be included, due to their model comprehensiveness advantage, which shall be more explored in future works of the authors.

6. Conclusion

This work presented an experimental study comparing the use of nine ML techniques to model the potential distribution of 35 Latin American plant species: Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Genetic Algorithm for Rule Set

Production (GARP), Decision Trees (DTs), Random Forests (RF), *k*-Nearest Neighbors (kNN), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs).

Although GARP is one of the most used algorithms in biogeographical studies, results indicate RF as a promising modelling technique, due to its high performance in all datasets. SVMs and ANNs, with proper parameter settings, are also good candidates for potential distribution modelling. Although not explored in this paper, the comprehensiveness of symbolic ML models as DTs also includes them as a good modelling techniques.

In the present study, a boolean presence/absence prediction was considered. As future work, it would be worth considering the probability of predictions. Related to that, the use of fuzzy-based prediction techniques could be an interesting direction.

Besides, although the use of pseudo-absence data is quite usual in potential distribution modelling, it would be interesting to investigate the use of one-class classification approaches in this domain (Tax, 2001).

Numerous environmental layers could be used as attributes in the datasets. The choice adopted in this work was based on literature and ecological expertise. Since the same datasets were used in the induction of all classifiers, this fact does not invalidate the comparisons performed. Nevertheless, automatic feature-selection techniques (Siedlecki & Sklansky, 1993) can help the proper choice of environmental layers for prediction.

A wider comparative study should also include other techniques commonly used in potential distribution modelling, in addition to ML techniques, exploring the advantages/disadvantages of each approach. Other comparison measures can also be included in future works, such as time spent in model generation and prediction calculation.

Acknowledgments

To the financial support of the Brazilian Research Agencies FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References

Anderson, R. P., Laverde, M., & Peterson, A. T. (2002). Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos, 93*, 3–16.

Araújo, M. B., Williams, P. H., & Reginster, I. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation, 96*, 331–345.

Berry, P. M. (2002). Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. *Global Ecology & Biogeography, 11*, 453–462.

Bourg, N. A., McShea, W. J., & Gill, D. E. (2005). Putting a CART before the search: Successful habitat prediction for a rare forest herb. *Ecology, 86*, 2793–2804.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brotons, L. (2004). Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography, 27*, 437–448.

Chang, C.-C., & Lin, C.-J. (2004). LIBSVM: A library for support vector machines. <http://www.csie.ntu.edu.tw/cjlin/libsvm/>.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning: Proceedings of the 12th international conference* (pp. 115–123).

Demsar, J. (2006). Statistical comparisons of classifiers over multiple datasets. *Journal of Machine Learning Research, 7*, 1–30.

Elith, J., Graham, C. H., Anderson, R. P., Dudk, M., Ferrier, S., Guisan, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*, 129–151.

Ferrier, S. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity: The north-east New South Walles experience. I. Species-level modelling. *Biodiversity and Conservation, 11*, 2275–2307.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*, 675–701.

Furnkranz, J., & Widmer, G. (1994). Incremental reduced error pruning. In *Proceedings of the 11th international conference on machine learning* (pp. 70– 77).

Graham, C. H., Moritz, C., & Williams, S. E. (2006). Habitat history improves prediction of biodiversity in rainforest fauna. In *PNAS (Proceedings of the national academy of sciences of the united states of America)* (Vol. 103, pp. 632–636).

Graham, C. H., Ron, S. R., Santos, J. C., Schneider, C. J., & Moritz, C. (2004). Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *EVOLUTION, 58*, 1781–1793.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C., Aspinall, R., et al. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology, 43*(L1).

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters, 8*, 993–1009.

Hannah, L., Midgley, G. F., Andelman, S., Arajo, M. B., Hughes, G. O., Martinez-Meyer, E., et al. (2007). Protected area needs in a changing climate. *Frontiers in Ecology and the Environment, 5*, 131–138.

Hannah, L., Midgley, G. F., Hughes, G., & Bomhard, B. (2005). The view from the Cape: Extinction risk, protected areas, and climate change. *BioScience, 55*, 231–242.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.

Higgins, S. I., Richardson, D. M., Cowling, R. M., & Trinder-Smith, T. H. (1999). Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. *Conservation Biology, 13*, 303–313.

Hijmans, R. J. (2005). *Very high resolution interpolated climate surfaces for global land areas, 25*, 1965–1978.

Hugall, A. (2002). Reconciling paleodistribution models and comparative phylogeographic in the Wet Tropics rainforest lan snail Gnarosophia bellendenkerensis. *Brazier 1875, 99*, 6112–6117.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of Friedman statistic. *Communications in Statistics*, 571–595.

Kleinbaum, D. G., & Klein, M. (2005). *Logistic regression* (2nd ed.). Springer.

Kozak, J. H., & Wiens, J. J. (2006). Does niche conservatism promote speciation? A case study in North American salamanders. *Evolution, 60*, 2604–2621.

Leathwick, J. R. (2005). *Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish, 50*, 2034–2052.

Loiselle, B. A. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology, 17*, 1591–1600.

Lorena, A. C., Siqueira, M. F., Giovanni, R., Carvalho, A. C. P. L. F., & Prati, R. C. (2008). Potential distribution modelling using machine learning. In *The twenty first international conference on industrial, engineering & other applications of applied intelligent systems (IEA/AIE), Wroclaw, Poland. Lecture notes in artificial intelligence* (Vol. 5027, pp. 255–264). Springer-Verlag.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, (pp. 41–48).

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Mitchell, M. (1999). *An introduction to genetic algorithms*. MIT Press.

Nelson, T. A., & Boots, B. (2008). Detecting spatial hot spots in landscape ecology. *Ecography, 31*, 556–566.

Nemenyi, P. B. (1963). Distribution-free multiple comparisons, Ph.D. Thesis, Princeton University.

Nix, H. A. (1986). A biogeographic analysis of Australian elapid snakes. In R. Longmore (Ed.), *Atlas of elapid snakes of Australia. Australian Flora and Fauna Series No. 7* (pp. 5–15). Canberra: Australian Government Publishing Service.

Ortega-Huerta, M. A., & Peterson, A. T. (2004). Modelling spatial patterns of biodiversity for conservation prioritization in north-eastern Mexico. *Diversity and Distributions, 10*, 39–54.

Pearce, J., & Lindenmayer, D. B. (1998). Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (Linchenostomus melanops cassidix) in southeastern Australia. *Restoration Ecology, 6*, 238–243.

Pearson, R. G. (2007). Species' distribution modeling conservation educators and practitioners. *Synthesis*. New York: American Museum of Natural History.

Pearson, R. G., Dawson, T. P., Berry, P. M., & Harrison, P. A. (2002). Species: A spatial evaluation of climate impact on the envelope of species. *Ecological Modelling, 154*, 289–300.

Pearson, R. G., Thuiller, W., Arajo, M. B., Martinez, E., Brotons, L., McClean, C., et al. (2006). Model-based uncertainty in species' range prediction. *Journal of Biogeography, 33*, 1704–1711.

Peterson, A. T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *Quarterly Review of Biology, 78*, 419–433.

Peterson, A. T., Benz, B. W., & Papes, M. (2007). Highly pathogenic H5N1 avian influenza: Entry pathways into North America via bird migration. *PLoSONE, 2*, e261.

Peterson, A. T., Lash, R. R., Carroll, C. R., & Johnson, K. M. (2006). Geographic potential for outbreaks of Marburg hemorrhagic fever. *American Journal of Tropical Medicine & Hygeine, 75*, 9–15.

Peterson, A. T., Ortega-Huerta, M. A., Bartley, J., Sanchez-Cordero, V., Buddemeier, R. H., & Stockwell, D. R. B. (2002). Future projections for mexican faunas under global climate change scenarios. *Nature, 416*, 626–629.

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*(6), 863–868.

Peterson, A. T., Sobern, J., & Sanchez-Cordero, V. (1999). Conservatism of ecological niches in evolutionary time. *Science, 285*, 1265–1267.

Quilan, J. R. (1988). *C4.5 programs for machine learning*. CA: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Raxworthy, C. J., Ingram, C., Rabibosa, N., & Pearson, R. G. (2007). Species delimitation applications for ecological niche modeling: A review and empirical evaluation using Phelsuma day gecko groups from Madagascar. *Systematic Biology, 56*, 907–923.

Segurado, P., & Araújo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography, 31*, 1555–1568.

Siedlecki, W., & Sklansky, J. (1993). *On automatic feature selection. Handbook of pattern recognition & computer vision*. World Scientific Publishing Co.

Sobern, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters, 10*, 1115–1123.

Stockwell, D. R. B., & Peters, D. P. (1999). The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems, 13*, 143–158.

Tax, D. M. J. (2001). *One-class classification; Concept-learning in the absence of counter-examples*, Ph.D. Thesis, Delft University of Technology.

Thuiller, W. (2003). BIOMOD – Optimising predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology, 9*, 1353–1362.

Thuiller, W., Arajo, M. B., Pearson, R. G., Whittaker, R. J., Brotons, L., & Lavorel, S. (2004). Biodiversity conservation – uncertainty in predictions of extinction risk. *Nature, 430*, 33 [discussion following].

Thuiller, W., Richardson, D. M., Pysek, P., Whittaker, R. J., Brotons, L., & Lavorel, S. (2005). Niche-based modeling as a tool for predicting the global risk of alien plant invasions. *Global Change Biology, 11*, 2234–2250.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Verdin, K. L., & Greenlee, S. K. (1996). Development of continental scale digital elevation models and extraction of hydrographic features, In *3rd International Conference/Workshop on Integrating GIS and Environmental Modeling, National Center for Geographic Information and Analysis, Santa Barbara, California, Santa Fe, New Mexico*.

Williams, N. S. G., Hahs, A. K., & Morgan, J. W. (2008). A dispersal-constrained habitat suitability model for predicting invasion of alpine vegetation. *Ecological Applications, 18*, 347–359.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufman.