A prediction algorithm for niche modeling on big environmental data

David R.B. Stockwell*

April 3, 2015

Abstract

This package is an R implemention of the **WhyWhere** data mining algorithm [5]. Developed for biodiversity modelling on big data, the approach is simple and rigorous, efficient to compute, and provides accuracy equal to the best alternative approaches. Thus, it represents the way forward in utilize large data sets of information about species and their environmental relationships.

1 Introduction

This package is concerned with the development of models of species distributions on arbitrarily large environmental data-sets, although the approach could potentially be used in the other fields that require geospatial prediction. Large 2D geospatial data arises from global coverage at fine resolution of temperature, rainfall and other surfaces that might potentially be correlated with some entity of interest.

The operation of the package is similar to the **dismo** package in which implements a number of species distribution models (MaxEnt, Bioclim, Domain, GLM, GAM, and RandomForest). It also follows the pattern of such sampling background points, and the program flow as developed in [4] and commonly used for species distribution modeling.

The novel contribution of WhyWhere is to implement an algorithm that addressed the big data problem, as the size of data sets expand there is a need to reexamine traditional approaches to data analysis with expensive memory or computational demands and if necessary redesign the algorithms that in order to achieve the same purposes. A number of features were proposed in the WhyWhere algorithm in [5] to remedy issues that often block or limit application of other algorithms in some way:

*All correspondence to be addressed to the author at david.r.stockwell@cqu.edu.au, Adjunct Researcher at Central Queensland University.

1. Memory limitations: No more than one environmental data set in memory at one time. In most the **dismo** package all methods take prepared data in a 'wide' file format, i.e. similar to an excel spreadsheet. For this the data is extracted from a raster data brick. All environmental data must be in memory for the data brick which can exceed the memory limitations of the machine.

2. Incompatible resolutions in environmental data sets: In the **raster** package data coarse scale data-sets are enlarged to have the same resolution in order to form a data "brick" of coordinated grids of the same resolution. This is approach is memory expensive and is avoided in **WhyWhere**.

3. Mixed types and distributions: Robust to distributional properties. Some methods such as generalized linear regression (GLMs) are very sensitive to distributional assumptions. The method in **WhyWhere** can also handle factor variables. Few algorithms can handle both continuous and categorical variables. **WhyWhere** is similar to MaxEnt in this respect.

A number of advances have been implemented since the original **WhyWhere** implementation. We make use of the **raster** package for handling gridded spatial data [6] and also the **data.table** package for fast aggregation of large data [1]. The model is developed to optimize the Area Under Curve statistic rather than a Chi-squared statistic. In the original WhyWhere [5] a pcolor quantization algorithm was employed to segment the environmental space into groups in order to assign predicted probabilities for each group. Higher dimension models are now developed by combining vectors using fuzzy set conjunction (minimum) or disjunction (maximum). The approach is quite natural to predicting distributions on large numbers of large files, and the modifications are even more similar to machine learning approach.

The guide is arranged as follows. Section 2 illustrates the basic usage. Section 3 describes the theoretical basis. Section 4 compares the performance with the popular MaxEnt method on well known data-sets.

2 Example

The species distribution models are typically built from two data sources: the coordinates of the species of interest, and the environmental values, obtained as 2D map layers. The function whywhere is a wrapper around a number of more basic functions to perform the entire analysis in one step. Below illustrates a complete analysis on the Bradypus file of the **dismo** package [3] with output in Figure 1.

The inputs to **WhyWhere** are as follows: a **data.table** with the longitude and latitude of known locations, and a list of environmental data files that may be read into the raster package. A couple of other parameters are available: *multi* for conducting the search in multiple dimensions, *limit* for the minimum for entry into the results, *beam* to specify the number of entries to keep in the

```
> source("../R/ww2.R")
> #data(wrld_simpl)
> files <- list.files(path=paste(system.file(package="dismo"),'/ex',sep=''), pattern='grd',
> file <- paste(system.file(package="dismo"), '/ex/bradypus.csv',sep='')
> Pres <- read.table(file,  header=T,sep=",")
> Pres$species=NULL
> result=ww(Pres,files)
> plot.ww(result)
```

Figure 1: **WhyWhere** predicted distribution of the Bradypus data set.

beam search table, *e* the size of border around presence points, and *plot* if you want to plot as you go. The result is a list of components: results, lookup, etc. The main algorithm implements a classic beam search as follows:

1. label input coordinates with 1 and find range
2. generate random points within this range and label with 0
3. for all environmental files do
   (a) develop membership function for file
   (b) bind variable name and AUC to result
   (c) test conjunct of this variable with best
   (d) if result is better then bind to list
   (e) delete last row from result if too many
4. output

The two analytic steps of interest are the membership function and the combination of variables into expressions. The membership function takes the following inputs and outputs model data: *file* is a single environmental file, *ext* is the geographic extent, presence and absence locations, *pa* a vector of zeros and ones for presence and absence locations, and returns a model object *membership*. The model consists of a table of ranges of the variable, or factors for a categorical variable, with the counts and posterior probability in each category.

Figure 2 shows a table output for the highest rated variable in the Bradypus dataset. The categorical ranges are listed on the *x* axis, the prior distribution (background or 0s) is the black solid line, and the distribution of presences (1s) in each category is the dashed line. The posterior probability is the product of these (gray bars). Note that the almost uniform distribution of background classes due to the quantile cuts.

The procedure in membership is as follows:

1. get the file in raster format

```
> plot.dseg(result)
```

Figure 2: Response function bio6 plotted.

2. crop according to extent
3. extract environmental values from the file
4. determine number of breaks
5. cut into variable into quantiles (or factors)
6. construct table of factors with number of counts in each
7. calculate posterior probability
8. label row with probability
9. calculate AUC
10. return vector of items labeled with probability

Returning to the main algorithm, calculated before calling membership: the geographic extent, and the random sample of background data. This is necessary as the clip and background points must not vary. If the supplied species data does not contain background data, absences, then they are generated.

The membership function is used to produce a vector of probabilities. Vectors of probabilities for each variable are then combined using the fuzzy maximum principle to maximize the AUC. The rationale for this approach and comparison is in the following section.The output is a table of the best d variables as indicated by the AUC.

This scheme allows only one environmental variable to be loaded at a time. The rasters may have different resolutions, as the environmental values are extracted only in the one environmental variable. The variable is also cropped at this time. We also test a combination of the variables with the next best variables by applying the minimum of the item probability vectors and recalculating the AUC. It is possible to monitor the progress of **WhyWhere** with the plot option. This plots out the best model sofar and prints out a list of the best models considered.

3 Theory

Elements of justification of the approach are spread throughout the literature. The basic approach, that of segmentation of the continuous variables into equal portions has been around since early image processing days. The GIF format would compress images by reducing the number of colors in the red, green, blue 3D space. Tests found superiority of the median cut algorithm, for color reduction where each class has an equal number [2] . Here we quantiles for the same reason.

Median cuts have a good statistical basis as uniform sizes minimize the variability of the estimates of the posterior probability. Experiments performed with other schemes such as equal cuts were inferior. The exception is the factor variable which uses only those values in the layer. When using factor variables all factors are labeled as such.

The benefit of the approach is that it does not make assumptions about the distribution. Many make inappropriate assumptions about the distribution, even linear or monotonically increasing. In ecological niche theory a species is more abundant, or be more likely, within a limited range of an environmental variable. The typical variables used are temperature and precipitation ranges, and here a uni-modal or humped distribution is the simplest viable distribution. In addition, many environmental variables are categorical variables such as soil and vegetation type, so the approach is capable of handling the true relationships of species to environment without making assumptions about them.

The use of the quotient of counts needs justification. The locations where the species occurs can be thought of as a sampling of the environmental space. That is, while the counts of the environmental values has a distribution P_0 for over the range of the variable, the distribution of the sample where the species is present is P_1 over the same values. The most significant variable has the greatest difference between these distributions. The membership function is developed from the ration of counts P_1/P_0.

When we look at the change in this distribution between the background counts and the counts of the species sample, the classic approach to evaluating the significance of the the Chi-squared test, and this is what was used in the original **WhyWhere** package. The K-S test can also be used. However Chi2 doesn't work in evaluating the multivariate case.

Turns out we don't need it. If we work out the probability for each of the training data points and apply fuzzy set methodology to these vectors the resulting vector of probabilities can be evaluated using the AUC.

In another derivation the membership function can be viewed as a relaxation of some aspects of strict Bayesian statistics.

$$P(S|E) = P(E|S)P(S)/P(E) \tag{1}$$

But the problem is ratios of counts are not strictly probabilities. In any sense, probability of the occurrence of a species $P(S)$ is not well defined, or not generally of interest. In a typical example the probably of finding a species is dependent on season, search effort, and so is not well controlled. Usually we only want to know the best areas to search for a species, as predicted by the most significant environmental variables. So dropping the $P(S)$ we still have a proportional relationship which sufficient to compare alternative environmental variables.

$$P(S|E) \propto P(E|S)/P(E) \tag{2}$$

Other similar developments are using in MaxEnt where the membership is done using more complicated. The results are very similar as the final comparative section shows.

3.1 Fuzzy Conjunctions

The inferential basis is fuzzy set theory, where instead statements that are either true or false, a membership function describes a fuzzy truth value as a function $f : \mathbb{R} \rightarrow [0,1]$ from a variable V to the real unit interval $[0,1]$. One must consider membership functions taking values from other spaces such as categories, (also known as factors in R) N or on a space of many variables $V_1 \times V_2 \times ... \times V_n$ where each V_i is an interval in N or R.

Experience has shown that a particularly useful way of combining unitary membership functions to resemble the AND, OR and NOT operators of classical logic are Zadeh operators:

AND: $x \wedge y = min(f(x), f(y))$,
OR: $x \vee y = max(f(x), f(y))$, and
NOT: $\neg x = (1 - f(x))$.

There have been many approaches to learning fuzzy rules from from given data, and approaches to representation of the discovered rules. One of the outstanding problems is the trade-off between accuracy and interpretability, or prediction and explanation in the ecological literature. In particular in this package, ecological theory can motivate our approach, thus satisfying both criteria. It is the desire to address the problem of explanatory models that motivated t development of **WhyWhere** – to describe Why is a species Where? – without sacrificing predictive accuracy or computational tractability.

In the one dimensional case the variable with the largest significance is the best to select. This gives a list of variables that the species responds to the strongest.

But when we add more responses into a fuzzy expressions in a multi-variable data-sets the most accurate conjunctive expression may not contain the best single variable. This means we cannot use such methods as greedy search to identify higher dimensional expressions. The problem of expressing environmental relationships of more than one variable has been an important topic in statistical and ecological research, generalized linear modelling machine learning on the other. Here we need to consider the logical relationship between variables that are expected from ecological theory.

3.1.1 Law of the Minimum

The principle of Liebig's Law of the Minimum states that growth is controlled by the scarcest necessary resource. This is logically an AND operation or a fuzzy conjunction of limiting factors. Note that a model composed of an arithmetic sum would represent the concept of growth determined by the overall sum of resources contributed from different sources, and so is not consistent with the Law of the Minimum. Most model used to predict species distributions are of this form, and so their capacity to explain species distributions is questionable.

3.1.2 Competitive exclusion principle

In ecology, Gauss's law of competitive exclusion is a proposition that two species competing for the same resource cannot coexist at constant populations if all other things remain equal, due to effect of slight advantages magnified over generations. This behavioral shift leads to ecological niches. This is logically an OR operation or a fuzzy disjunction:

A possible example of this case is where a set of location actually contains two different species with different different habitats. A disjunction of two habitats may model this case better.

To evaluate both unit variate and multivariate combination of environmental variables, we predict the probability of presence on each location and calculate the area under the curve of the receiver operating statistic (AUC for short).

There is a further statistic that is the area under rank correlation or AUC. This arises where a models outputs a figure in the range of zero to one, but a prediction must be made of 0 or 1. We must select a cutoff value for the prediction to assign to zero or one. The AUC is the probability that would use the optimal accuracy.

A high AUC indicates that sites with high membership are more likely to be areas of presence, and vice versa. An AUC score of 0.5 is no better than random. The AUC can be calculated from.

4 Comparison

We test other models against the response output for the single dimensional case. The response functions for MaxEnt and **WhyWhere** are quite similar as follows - when we can get the rgdal package installed!!!

5 Conclusion and Further Work

The R package **WhyWhere** is a useful packages to fit, plot and test empirical species as a conjunction of response functions. More complex logical expressions are planned, as are improvements in computational efficiency and access to cloud data. The method provides simple and intuitive, efficient to compute, and typical predictive results that are at least equal to the best alternative approaches.

References

[1] M Dowle, T Short, S Lianoglou, and A Srinivasan. data.table: Extension of data.frame. 10 2014.

[2] Paul Heckbert. Color image quantization for frame buffer display. In *SIGGRAPH '82: Proceedings of the 9th annual conference on Computer graphics and interactive techniques*, pages 297–307, New York, NY, USA, 1982. ACM Press.

[3] R.J. Hijmans, S. Phillips, J. Leathwick, and J. Elith. dismo: Species distribution modeling with r. 2011.

[4] D Stockwell and D Peters. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal Of Geographical Information Science*, 13(2):143–158, 1999.

[5] David R B Stockwell. Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 192:188–196, 2006.

[6] Robert J. Hijmans & Jacob van Etten. raster: Geographic analysis and modeling with raster data. 2012. R package version 2.0-12.