COLBY COLLEGE

HONORS THESIS

Species Distribution Model Development and Validation

Author:
Victoria CHISTOLINI

Supervisor:
Dr. Stephanie TAYLOR

A thesis submitted in fulfillment of the requirements for Honors in the degree of Computer Science

in the

Record Lab
Bigelow Laboratory for Ocean Sciences

December 3, 2017

COLBY COLLEGE

Abstract

Dr. Nick Record
Bigelow Laboratory for Ocean Sciences

Computer Science

Species Distribution Model Development and Validation

by Victoria CHISTOLINI

The Thesis Abstract is written here ...

Acknowledgements

Thank you to Stephanie and Nick and Ben.

Contents

Chapter 1

Introduction

1.1 What is Species Distribution Modeling

According to the Center for Disease Control (CDC) in 2015, 95% of reported Lyme Disease cases came from only 14 states; of these 14, 12 were on the east coast and all 6 of the New England States were represented. Since 1996 the annual number of confirmed cases of Lyme Disease per year has increased by over 10,000 additional cases in 2016 [a]. Given the growing magnitude of the problem presented by ticks and Lyme Disease infection, there has been a deep interest to understand where Lyme Disease carrying ticks are located and how we can most effectively reduce human contact. Beginning in 1995, the Maine State Government began a project to create detailed records of the locations of discovered infected ticks. Doctors were encouraged to have their patients bring in any ticks that they found on their bodies, in their homes or on their pets, for free testing to determine if the tick were carrying Lyme Disease [b]. Data about the locations of tick sightings, tick gender and disease status, were recorded as part of the study until 2014.

Species Distribution Modeling (SDM) is a term used to categorize a whole class of models developed for the purpose of understanding the patterns and relationships of an observed species and its environment [c]. Often, the purpose of these models is to predict the range of a species based on where the species has been recorded during surveillance. Another application of SDM models is to predict where or not a species will be found in a certain location based on the environmental variables of the location. It is with this latter focus that we will pursue Species Distribution models throughout this work.

1.1.1 Data used in Species Distribution Models

This database of tick encounters developed by the Maine state government houses an enormous amount of information about the distribution of likely locations for ticks sightings in the state of Maine, however in its raw form it is missing a crucial metric: information about where ticks are not likely to be found. This lack of essential data is the crux of what makes species distribution models difficult to develop.

Classical modeling techniques use a set of predictor variables to classify events under certain conditions as likely or unlikely to happen. A logistic regression model, for example could take predictor variables about patients' heart rate and temperature to determine the response of the likelihood that the patient has the flu. In order to make a good prediction about the patient, the model needs heart rates and temperatures from patients who are healthy and those who are sick in order to be able to

discriminate sick from healthy metrics. Unfortunately, our tick dataset does not provide information about the absence of ticks in locations, which motivates the need to use different models. MAXENT LEADIN

Our dataset looks like...

1.2 The MaxEnt Model

-> how does it work
-> why machine learning: no model assumptions...

1.3 Predictor Selection

->types of dimensionality reduction (PCA, MCMC, correlation in predictors)

-> bias correction

1.3.1 Current MaxEnt model

-> predictors used, model structure

1.3.2 Additional Features of MaxEnt

-> population density,
host density,
weightings versus adding predictors

1.4 Ensemble Models

...

1.4.1 Ensembles with MaxEnt

* model physics
* initial conditions
* training data

1.5 Model Assessment

1.5.1 Assessment of a single model

-> TSS,
Confusion Matrix,
AUC,

AIC,
forecast AUC,
cross validation

1.5.2 Assessment of an ensemble of models

* summarizing ensembles -> a bayesian approach

Appendix A

Code for Experiments

A.1 Experiment 1

The code for this experiment has been written in R