Ecological Modelling 352 (2017) 108–112

Contents lists available at ScienceDirect

Ecological Modelling

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

Automated feature selection for a machine learning approach toward modeling a mosquito distribution

Ralf Wieland[a,*], Antje Kerkow[a,b], Linus Früh[a], Helge Kampen[c], Doreen Walther[a]

[a] Leibniz Centre for Agricultural Landscape Research, Eberswalder Str. 84, 15374 Müncheberg, Germany
[b] Freie Universität Berlin, Department of Biology, Chemistry, Pharmacy, Institute of Biology, Königin-Luise-Str. 1-3, 14195 Berlin, Germany
[c] Friedrich-Loeffler-Institut (FLI), Federal Research Institute for Animal Health, Südufer 10, 17493 Greifswald – Insel Riems, Germany

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 23 February 2017
Accepted 27 February 2017

Keywords:
Data science
Machine learning
Genetic optimization
Cluster computing
Mosquito dispersion modeling

ABSTRACT

This paper introduces a data science method to determine a set of features for training a vector support machine (SVM). The SVM is used to model the relationship between the distribution of one particular invasive mosquito species and climate data. Two biologists selected training data on the basis of their domain expertise. This was compared with the result of the data science simulation. The paper then explores the possible uses of data science to generate new knowledge as well as to identify the weaknesses of this technique.

1. Introduction

This paper uses the following definition of feature selection for model training. The full M dimensional space R^M, where M is the number of datasets, is replaced by a set of features F^N with $N \leq M$. The aim of the feature selection is to estimate the smallest subset of F relevant for a given criterion. The use of a smaller set of features decreases the calculation time for modeling, and reduces the influence of uncertain features on the model, to name only two advantages.

The selection of training parameters (features) for a machine learning approach is often based on domain expertise. The expert decides which features shall be included in the training and which shall not. The quality and quantity of the available features influence whether or not the modeler can select data using a statistical approach, for example to remove correlated data. Here correlation analysis, including PCA (Brumelis et al., 2000) or nonlinear PCA (Vo and Durlofsky, 2016), is used.

Behind this statistical analysis are methods which use supervised or unsupervised approaches. An elaborate method has been demonstrated in Golay and Kanevsk (2016), which is able to detect linear as well as nonlinear dependencies in the dataset. A large set of training methods including feature selection methods is described in Pedregosa et al. (2011). An other interesting stepwise method to remove variables based on an AIC is described in Zeng et al. (2016). A feature selection method based on sensitivity analysis is described in Specka et al. (2011). In contradiction to all of the established methods we want to introduce a method which uses a genetic optimization procedure to find an approximated optimal solution for the NP hard problem of feature selection. The key idea is to use the score of a machine learning approach to control the optimization. This means that the optimization method can be applied for different types of machine learning.

To model the influence of selected weather data on the habitat quality of (invasive) mosquitoes, Kerkow et al. (2017) applied a machine learning approach. In this investigation, researchers used correlation analysis to select the features used; this minimized the number of features by removing dependent features. Nevertheless, a correlation between weather data is always given. For example, it can be assumed that the mean temperature in January is correlated with the mean temperature in February. On the other hand, there should be little correlation between the temperature in January and in August. It is obvious that monthly data and seasonal data which includes that month are correlated, but the use of both datasets enables additional information to be gained. A warm spring is not the same as a warm March. This means that the weather data itself is not sufficient without taking into account its impact on the species to be modeled.

We asked different expert's to make a selection of modeling dependencies (features). The answers differed depending on

[*] Corresponding author.
E-mail address: rwieland@zalf.de (R. Wieland).

http://dx.doi.org/10.1016/j.ecolmodel.2017.02.029
0304-3800/© 2017 Elsevier B.V. All rights reserved.

each experts focus. One expert emphasized the weather during mosquitoes reproductive period in the summer; another expert emphasized the importance of survival during the cold period of the year. On the other hand, according to Wikipedia, data science is “an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms.”[1] Data science can be applied in a variety of fields with the aim of finding patterns and structures in large datasets which can then be interpreted as new knowledge. For example, Nair et al. (2016) show how topology can be used to model socioecological systems. A multi-indicator system is analyzed in Wei et al. (2015), and a dynamic Bayesian non-negative matrix factorization approach is used in Wang et al. (2016) to estimate an optimal solution in an NP hard problem. Another approach by Qi et al. (2016) combines a deep learning method with sample programming using a support vector machine (SVM), which is similar to the method presented in this paper. Other books that cover data science such as Grus (2015) and VanderPlas (2016) supplement the literature presented here. All this leads to the question of whether an automated procedure based on approaches using data science can be developed to find relationships between mosquito distribution and weather data which might explain the mosquito distribution. This paper therefore focuses on the following research questions:

- Can a data science approach determine an appropriate set of features for training?
- Can the automatically selected set outperform data selection by experts?
- Can the automated selection procedure benefit from domain expertise in terms of speed or accuracy?
- Can the automated selection generate new knowledge for biologists?
- What are the limitations of a data science approach?

The aim of this paper is to introduce a data science approach to automating features selection on the basis of a genetic algorithm implemented using a computer cluster.

2. Method

2.1. The CULBASE mosquito database

The species data was obtained from the German national mosquito database, CULBASE, which is maintained at the German Federal Research Institute for Animal Health, Friedrich-Loeffler-Institut (FLI). The CULBASE mosquito database is a relational database using an MS-SQL server. This database contains information about the location (x- and y-coordinates, the name of municipality, the postcode, etc.), the traps (the type of trap, the number of specimens in the trap, date of catch, etc.), the species (name, code, etc.) and the pathogens detected in the mosquitoes (not used here). CULBASE contains data from a monitoring program maintained by the ZALF and the FLI and from a citizen science project ‘Mückenatlas’ (mosquito atlas) (Kampen et al., 2015). Both data was used. We selected four species according to the recommendation of biologists: *Aedes japonicus japonicus*, *Aedes vexans*, *Aedes geniculatus* and *Aedes daciae*. *A. japonicus japonicus*, as an invasive species, is the target species of the modeling task (we suspect *A. japonicus japonicus* as a vector for pathogens). The machine learning algorithm (here a SVM was used) classifies the *A. japonicus japonicus* as class A and all other mosquitoes as class B. The evaluation of a trained SVM with a data set which was not included in the training could be evaluated as a so called confusion matrix (Table 1).

Table 1
The confusion matrix.

$n = a + b + c + d$	Predicted A	Predicted B
Observed A	a	b
Observed B	c	d

The precision p is defined as:

$$p = \frac{a}{a + c} \tag{1}$$

the recall r is defined as:

$$r = \frac{a}{a + b} \tag{2}$$

the $f1$ score which controls the optimization is defined as:

$$f1 = 2 * \frac{p * r}{p + r} \tag{3}$$

the n is the number of observations. It should be underlined that we have used the mosquito data of the years 2011–2014 for training of the SVM. To calculate $f1$ (to control the genetic optimization) the mosquito data of the year 2015 was used.

2.2. Coding of features

The features were freely available weather data from the German Weather Service (DWD). The DWD provides gridded data with a resolution of 1 km × 1 km for Germany. For training, we hypothesized that the following data would be relevant: monthly mean temperature, monthly precipitation sum, seasonal temperature, seasonal precipitation, seasonal drought index, and the annual number of frost days.

The total number of 37 possible features is too large for model training when we take into consideration the size of available mosquito data (for a general discussion see Singer et al., 2016). A subset of 6–12 features should be sufficient. The biologists used 7–8 features. To structure the selection, a code was developed. The first digits represent the index (0, 1, 2, 3, . . .) of the vector [0, 1, 0, . . ., 0], with 1 indicating selection of the features used according to:

T01:	mean temperature in January
. . .	
T12:	mean temperature in December
P01:	sum of precipitation in January
. . .	
P12:	sum of precipitation in December
ST13:	mean temperature from March to May
ST14:	mean temperature from June to August
ST15:	mean temperature from September to November
ST16:	mean temperature from December to February
SP13:	sum of precipitation from March to May
. . .	
SP16:	sum of precipitation from December to February
SD13:	drought index from March to May
. . .	
SD16:	drought index from December to February
FROST:	number of frost days

```
selector={0:’T01’,1:’T02’,2:’T03’,3:’T04’,4:’T05’,5:’T06’,6:’T07’,
          7:’T08’,8:’T09’,9:’T10’,10:’T11’,11:’T12’,
          12:’P01’,13:’P02’,14:’P03’,15:’P04’,16:’P05’,17:’P06’,18:’P07’,
          19:’P08’,20:’P09’,21:’P10’,22:’P11’,23:’P12’,
          24:’ST13’,25:’ST14’,26:’ST15’,27:’ST16’,
          28:’SP13’,29:’SP14’,30:’SP15’,31:’SP16’,
          32:’SD13’,33:’SD14’,34:’SD15’,35:’SD16’,36:’FROST’}
```

[1] https://en.wikipedia.org/wiki/Data_science.