Beauty Contest

7th July 2014

Abstract

Pithy, concise and informative. May bring the reader to tears due to the beauty of it.

1 Introduction

2 Methods

The availability of large data sets for building regression models to predict the bacterial counts in beach water is both an opportunity and a challenge.

2.1 Data Sources

Possibly move this to the end of the section

Which sites

Where are they

What specific sources sources of data (plug EnDDAT)

Will include a map and tables

2.2 Definitions

At any site, denote the predictor variables by X, which is an $n \times p$ matrix where n is the number of observations and p is the number of predictors. The vector of n observations of bacterial concentration is denoted y. The mathematical model relating y to X is the function $\mu(X,\mathrm{y})$.

2.3 Goals

With input from the US Environmental Protection Agency, the state of Wisconsin has established regulatory standards for beach water quality, which states that a warning is to be posted when the concentration of E.

coli exceeds 235 CFU / 100 mL. The goal of modeling the bacterial concentration is to predict in advance when the concentration will exceed the limit. For the discussion to come, denote the regulatory standard by δ. Each model has two essential components: the mathematical model itself, $\mu(X, y)$, and a decision threshold, $\hat{\delta}$, that is used to interpret the model's predictions.

There is no reason that δ and $\hat{\delta}$ must be equal. Rather, each $\hat{\delta}$ should be chosen so that the number of false positives and false negatives are balanced to the satisfaction of the beach manager. Using cross validation allows us to set $\hat{\delta}$ to expect that the performance over future data will resemble what was observed over the testing data.

2.4 Listing of specific statistical techniques

Fourteen different regression modeling techniques were considered. Each technique uses one of five modeling algorithms: GBM, the adaptive lasso, the genetic algorithm, PLS, or sparse PLS. Each technique is aplied to either continuous or binary regression and to either modeling, or variable selection only.

Continuous or binary regression

The goal of predicting exceednaces of the water quality standard is approached in two ways: one is to predict the bacterial concentration and then compare the prediction to a threshold, which is referred to as continuous modeling. The other is referred to as binary modeling, in which we predict the state of the binary indicator z_i:

$$z_i = I(y_i > \delta) \tag{2.1}$$

The indicator is coded as zero when the concetration is below the regulatory standard and one when the concentration exceeds the standard. All of the binary modeling techniques herein use logistic regression, which uses the logistic link function g to translate $p_i = E(z_i)$ - the probability that the ith observation is an exceedance - into an unbounded quantity.

$$g(p_i) = \log \frac{p_i}{1 - p_i} \tag{2.2}$$

Weighting of observations in binary regression

A weighting scheme was implemented for some of the binary regression techniques. In the weighting scheme, observations were given weights w_i where:

$$w_i = (y_i - \delta)/\hat{sd}(y)$$
$$\hat{sd}(y) = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}$$
$$\bar{y} = \sum_{i=1}^{n} y_i/n$$

That is, the weights are equal to the number of standard deviations that the observed concentration lies from the regulatory threshold δ. Any technique that was implemented with this weighting scheme was separately implemented without any weighting of the observations. The techniques are thus labeled weighted and unweighted, respectively.

Modeling or selection only

Some methods are labeled "select", which means that they are used for variable selection only. In these cases, once the predictor variables are selected, the regression model using those predictors is estimated using ordinary least squares for the continuous regression techniques, or ordinary logistic regression for the binary regression techniques.

2.4.1 GBM

GBM refers to the gradient boosting machine (GBM) of Friedman [2001]. A GBM model is a so-called random forest model - a collection of many regression trees. Prediction is done by averaging the outputs of the trees. Two GBM-based techniques are explored - we refer to them as GBM and GBMCV. The difference is in how the optimal number of trees is determined - GBMCV selects the number of trees in a model using leave-one-out CV, while GBM uses the so-called out-of-bag (OOB) error estimate. The CV method is much slower (it has to construct as many random forests as there are observations, while the OOB method only requires computing a single random forest) but GBMCV should more accurately estimate the prediction error. All the GBM and GBMCV models share the following settings:

Number of trees: 10000

Shrinkage parameter: 0.0005

Minimum observations per node: 5

Depth of each tree: 5

Bagging fraction: 0.5

2.4.2 Adaptive Lasso

The adaptive lasso Zou [2006] is a regression method that simultaneously selects relevant predictors and estimates their coefficients by adding a penalty to the sum of the squared residuals. For continuous modeling techniques the adaptive lasso selects the predictors for linear regression, estimating $\hat{\boldsymbol{\beta}}$ minimize the criterion $\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}$, where λ is a tuning parameter and $\tilde{\boldsymbol{\beta}}$ is a consistent estimate of the regression coefficients.

For binary modeling,

In this work, γ is set to one, $\tilde{\boldsymbol{\beta}}$ are estimated individually by a univariate linear or logistic regression (it is necessary to estimate the coefficients individually because there are usually more covariates than observations) and the adaptive lasso tuning parameter λ is selected to minimize the AICc [Hurvich et al., 1998].

Three of the modeling techniques were based on the adaptive lasso - one fo

2.4.3 Genetic algorithm

The genetic algorithm [Fogel, 1998] is a variable-selection method that works by analogy to natural selection, where so-called chromosomes represent regression models. A variable is included in the model if the corresponding element of the chromosome is one, but not otherwise. Chromosomes are produced in successive generations, where the first generation is produced randomly and subsequent generations are produced by combining chromosomes from the current generation, with additional random drift. The chance that a chromosome in the current generation will produce offspring in the next generation is an increasing function of its fitness. The fitness of each chromosome is calculated by the corrected Akaike Information Criterion (AICc) Akaike [1973], Hurvich and Tsai [1989].

The implementations in this study used 100 generations, with each generation consisting of 200 chromosomes. The genetic algorithm method GALM is the default for linear regression modeling in Virtual Beach [Cyterski et al., 2013]. The study also investigates two genetic algorithm methods for logistic regression: one weighted (GALogistic-weighted) and one unweighted (GALogistic-unweighted).

2.4.4 PLS

Partial least squares (PLS) regression is a tool for building regression models with many covariates [Wold et al., 2001]. PLS works by decomposing the covariates into mutually orthogonal components, with the components then used as the variables in a regression model. This is similar to principal components regression (PCR), but the way PLS components are chosen ensures that they are aligned with the model output. On the other hand, PCR is sometimes criticised for decomposing the covariates into components that are unrelated to the model's output.

To use PLS, one must decide how many components to use in the model. The technique used in this study follows the method described in Brooks et al. [2013], using the PRESS statistic to select the number of components.

2.4.5 SPLS

Sparse PLS (SPLS) is introduced in Chun and Keles [2007]. The SPLS method combines the orthogonal decompositions of PLS with the sparsity of lasso-type variable selection. To do so, SPLS uses two tuning parameters: one that controls the number of orthogonal components and one that controls the lasso-type penalty. The optimal parameters are those that minimize the mean squared prediction error (MSEP) over a two-dimensional grid search. The MSEP is calculated by 10-fold cross-validation. Two techniques utilizing SPLS were

2.5 Implementation for beach regression

The response variable for our continuous regression models is the natural logarithm of the E. coli concentration. For the binary regression models, the response variable is

Include a table with pre/post processing discussion

This includes tuning of parameters

Some specific data issues because we are estimating a threshold exceedence

2.6 Cross Validation

Assessment of the modeling techniques is based on their performance in predicting exceedances of the regulatory standard. Two types of cross validation was used to measure the performance in prediction: leave-one-out (LOO) and leave-one-year-out (LOYO). In LOO CV, one observation is held out for validation while the rest of the data (the model data) is used to train a model. The model is used to predict the concentration of that held out observation, and the process is repeated for each observation. Each cycle of LOYO CV holds out one year's worth of data for validation instead of a single observation. It is intended to approximate the performance of the modeling technique under a typical use case: a new model is estimated before the start of each annual beach season and then used for predicting exceedances during the season. That year's data is then added to the dataset to estimate a model for the next beach season. The LOYO models in this study were estimated using all the available data, even that from future years - so for instance the 2012 models were estimated using the 2010-2011 and 2013 data

Some methods also used cross-validation internally to select tuning parameters. In those cases the internal CV was done by partitioning the model data, leaving out one partition at a time. This process is separate from - and does not affect - the CV to assess predictive performance.

2.7 Performance Metrics

How did we evaluate the performance of each technique on all the different data sets

- for all cases —> AUC (ROC curve)
- continuous variables using PRESS (skill –> Like Nash-Sutcliffe/R^2 over the fitted data)
- True/False Positives/Negatives (needs a threshold)
- Which variables are selected for models where variable reduction takes place
  - challenge regarding the fact that different variables are selected in each fold. Maybe use frequencies?
  - also the number of variables selected (metric of complexity)

OPTIONAL

- AIC/BIC? –> not the same model in each fold so maybe not possible
- Maybe some form of confusion matrices – perhaps a grid of them with or without companion variance plots or other estimates of the range of results

3 Results

The area under the ROC (AUROC) curve assesses how accurately the predictions of the left-out observations are sorted. The methods are ranked at each site by AUROC and a mean rank (across sites) is computed for each method. The mean LOO and LOYO ranks are plotted in Figure 3.1. The GBM and GBMCV techniques did best under both kinds of cross validation, and in both cases the adaptive lasso was the third-best technique. The PLS and GALM techniques both appeared well down the list of the best techniques.

The performance of GBM was about equivalent to that of GBMCV at a much lower computational cost, so GBMCV will hereafter be set aside for discussion in favor of GBM. Further analysis focuses on GBM and the adaptive lasso.

4 Discussion

In general, the GBM, and GBMCV, and AL techniques produced comparable results that were superior to the other techniques in terms of predictive performance. Since the GBMCV models take much longer to compute than the others, we will not include them in our more detailed analysis of the modeling results.

Which type of model is generally the best?

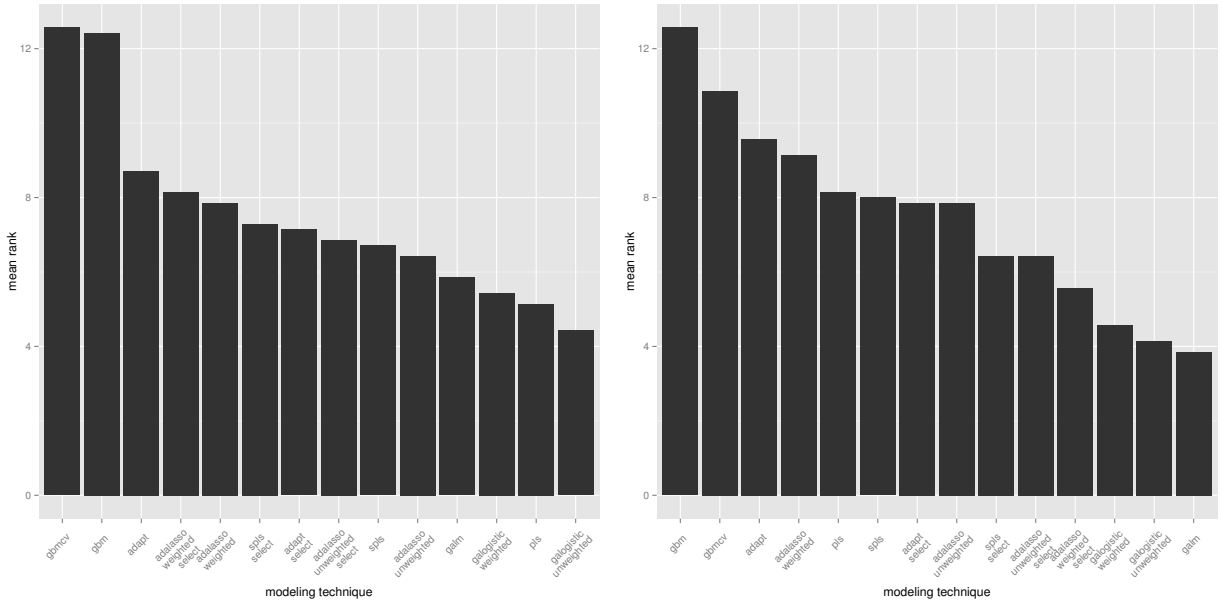

Figure 3.1: Mean ranks of the modeling techniques across the seven sites. At left are the mean ranks under leave-one-out cross validation, at the right are the mean ranks from leave-one-year-out cross validation.

Under what conditions do some outperform others?

Relative value of overall best model versus methods that help trim variables? e.g. how valuable is it to reduce number of predictors? Further, which variables get cut? Expensive ones? Cheap ones?

How important is computational expense? Only an issue for model fitting — not prediction, but worth quantifying. E.g. if GBM with cross validation takes hours, how much better?

Model tuning for GBM versus GBM-CV –> notes on how GBM is faster with similar performance (e.g. CV is overkill maybe)

5 Acknowledgments

6 References

References

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pages 267–281. Akademinai Kiado, 1973.

Wesley R. Brooks, Michael N. Fienen, and Steven R. Corsi. Partial least squares for efficient models of fecal indicator bacteria on great lakes beaches. *Journal of Environmental Mana*, 114:470–475, 2013.

Hyonho Chun and Sunduz Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. 2007. URL `http://www.stat.wisc.edu/ keles/Papers/SPLS`$_N ov$07*.pdf*.

Mike Cyterski, Wesley Brooks, Mike Galvin, Kirt Wolfe, Rebecca Carvin, Tonia Roddick, Michael N. Fienen, and Steven R Corsi. *Virtual Beach 3: user's guide*. United States Environmental Protection Agency, 2013.

David B. Fogel. *Evolutionary computation: the fossil record*. Wiley-IEEE Press, 1998.

Jerome Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001.

Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

Clifford M. Hurvich, Jeffrey S. Simonoff, and Chih-Ling Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Methodology)*, 60:271–293, 1998.

Svante Wold, Michael Sjostrom, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58:109–130, 2001.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101: 1418–1429, 2006.