Document downloaded from http://www.elsevier.es, day 26/04/2024. This copy is for personal use. Any transmission of this document by any media or format is strictly prohibited.

MATERIAL SUPLEMENTARIO

Model description

Stochastic Gradient Boosting

Stochastic Gradient Boosting (SGB) is a type of ensemble algorithm. An ensemble algorithms consist of multiple base models, each one of those provides a different solution to the problem; The solutions of all the base models, are finally combined (usually by weighted voting or averaging) into a single final model output, which is usually a more stable and accurate prediction. The intuition behind this is that many different predictors trying to predict the same outcome variable will perform better than any single predictor alone. Two major ensemble techniques (with many variations) have been developed, and the main difference between them is how the predictors are constructed; these techniques are: (1) boosting, where the predictors are not fitted independently, but sequentially, this mean that the subsequent predictors are based on the results of previous predictors, and (2) bagging where the predictors are fitted using random subsets of the original training data for each model [1],[2].

SGB is a hybrid of the boosting and bagging approaches and uses L-terminal node small trees as base model [3]. At each boosting iteration a random sub-sample of the training dataset is selected to fit a tree; besides, SGB is based on a steepest gradient algorithm which places emphasis on misclassified training data that are close to their correct classification; finally, at each iteration, relatively small trees are developed (in this study trees with 3,5,7 and 9 terminal nodes were evaluated) and summed, then each observation is classified according to the most common classification among the trees. SGB is capable of manage qualitative and quantitative variables, and remain robust to missing data and outliers. SGB model has been successfully applied for prediction of mortality in head injury [4], where the Boosted Tree Classifier method achieved both the highest AUROC and accuracy rate.

So, as in all function estimation methodologies, the ensemble algorithms begin with a training sample $\{y_i, x_i\}_1^N$ which are composed of a set of explanatory variables $x = \{x_1, \cdots, x_n\}$ and a response variables y; and the goal is to find a function $F^*(x)$ that maps x to y and minimizes a loss function $\Psi\big(y, F(x)\big)$ [1].

Stochastic Gradient Boosting (SGB) is a type of ensemble algorithm that approximates $F^*(x)$ using an additive expansion:

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m)$$

Where $h(x; a_m)$ is simple base model of x with parameters $a = \{a_1, a_2, \dots\}$, and β_m is an expansion coefficient that must be jointly fit with the parameters a_m [1]. The algorithm stars with guess of F_0 and then the expansion coefficient, the parameters and the function are calculated iteratively for $m = 1,2, \cdots, M$, as follows:

$$(\beta_m, a_m) = \arg\min_{\beta, a} \sum_{i=1}^{N} \Psi\big(y_i, F_{m-1}(x_i) + \beta h(x_i; a)\big)$$

and

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$$

To solve the above equation (β_m, a_m), SGB first fit $h(x; a)$ by least-squares:

$$a_m = \arg\min_{a, \rho} \sum_{i=1}^{N} [\tilde{y}_{im} - \rho h(x_i; a)]^2$$

Where (with an arbitrary differentiable loss function)

Document downloaded from http://www.elsevier.es, day 26/04/2024. This copy is for personal use. Any transmission of this document by any media or format is strictly prohibited.

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

Then, the optimal value of β_m is calculated

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)\right)$$

SGB uses trees as base model and combine the strengths of two algorithms: regression trees and boosting [3]. Specifically, for this algorithm $h(x; a)$ is a L-terminal node small regression tree; at each iteration of m, a regression tree partitions the explanatory variables space into L-disjoint sub regions $\{R_{lm}\}_{l=1}^{L}$ in each of which a constant response value $\bar{y}_{lm}$ is calculated, so the solution to β_m reduces to:

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{x_i \in R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma)$$

and

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_{lm} 1(x \in R_{lm})$$

To improve the performance, SGB incorporates randomness, so at each iteration a random permutation $\{\pi(i)\}_1^N$ is selected (without replacement and with size $\widetilde{N}$) for fit the regression tree. For this study the output variable y is binary, and the loss criterion is the deviance $\Psi\left(y, \hat{F}\right) = 2\log\left(1 + \exp\left(-2y\hat{F}\right)\right)$ [1].

The SGB algorithm involves a parameter-tuning process. the three main parameters are: M, the total number of boosting iterations (the number of trees that are aggregated in the final model); υ, the learning rate or shrinkage coefficient which determines the contribution of each tree to the growing model and helps to control over-fitting by controlling the gradient steps and L, is the iteration depth (the number of splits performed on each tree) [2],[5].

To determine the optimal combination of the mentioned parameters, 10 fold cross-validation procedure was applied for each parameter configuration with these values M (50, 100, 150, 200, 250, …, 650, 700, 750), L (3, 5, 7, 9) and υ (0.01, 0.05, 0.1, 0.5). In this procedure, the elements of the train subset were randomly divided into 10 groups, nine of these groups were selected for fitting a model and the other one was used for testing it, the process was repeated ten times, so each group was used for testing and training. By averaging the results produced in each iteration, an overall quality estimate was obtained. Finally, the combination of parameters that present better performance (highest AUROC in this study) across all 10 folds was selected as the final model. The validation subset was never used in the development of the SGB model, but it was used to evaluate the performance of the final model. Figure 1 presents an example of the parameter tuning process.

The general effect on the model of each predictor was calculated using their relative variable importance. This measures are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged across the entire boosting ensemble [3],[6]. The relative influence of each variable is scaled so that the sum adds to 100, with higher numbers indicating stronger influence on the response [3]. The most influential predictors were selected by developing a model with the predictors that had the greatest relative importance, and that in the end presented an AUROC similar to that of the complete model.

Document downloaded from http://www.elsevier.es, day 26/04/2024. This copy is for personal use. Any transmission of this document by any media or format is strictly prohibited.

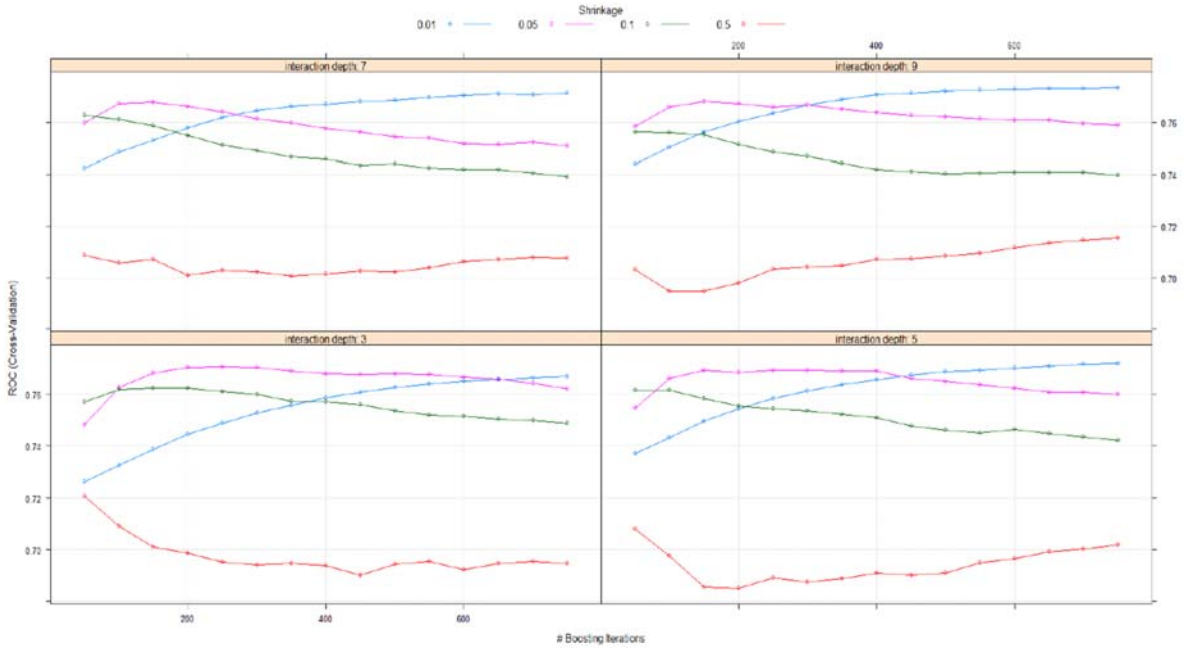

Figure 1. SGB model tuning parameters and AUROC.

Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) [7], is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model. LASSO puts a constraint on the sum of the absolute values of the model parameters, applying a regularization process where it penalizes the coefficients of the regression variables and set some of them exactly to zero. After this process the variables with a coefficient different from zero are selected to be part of the model. In practice there is tuning parameter λ, that controls the controls the amount of shrinkage that is applied to the estimates [7], which is selected using cross-validation in a way that the resulting model minimizes the sample error.

The method for two-class classification seeks the probability of class membership $p(X)$ and is based on a hypothesis function that lies between 0 and 1. For logistic regression [8]:

$$p(X) = \frac{exp^{(\beta_0+\beta_1^T)}}{1 + exp^{(\beta_0+\beta_1^T)}}$$

Which is equivalent to:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1^T$$

Where β_1 is a vector, with as many components as there are predictors, and the objective is to find the values of β_1 that results in a *p(X)* that most accurately classifies all the observed data points. Logistic regression models can be fitted by maximum likelihood. The log-likelihood can be written:

$$l = \sum_{i=1}^{N} \left\{ y_i(\beta_0 + \beta_1^T x_i) - \log\left(1 + e^{(\beta_0+\beta_1^T x_i)}\right)\right\}$$

Document downloaded from http://www.elsevier.es, day 26/04/2024. This copy is for personal use. Any transmission of this document by any media or format is strictly prohibited.

LASSO regularization works by adding a penalty term to the log likelihood function, thus the quantity to be minimized is:

$$l + \lambda \sum_{j=1}^{p} |\beta_{1j}|$$

Where λ is a parameter that is selected using crossvalidation in a way that the resulting model minimizes the sample error. The effect of the LASSO penalty term is to set some of the models coefficients exactly to zero, and thus allowing the variable selection.

Hosmer-Lemeshow test was used on the model to verify its ability to provide a risk estimate that corresponds to the observed mortality (Calibration), The goodness of fit was evaluated using the Pearson's Chi-square Test

References

1. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378
2. Lawrence R, Bunn A, Powell S, Zambon M (2004) Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sens Environ 90:331–336 . doi: 10.1016/j.rse.2004.01.007
3. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813 . doi: 10.1111/j.1365-2656.2008.01390.x
4. Sut N, Simsek O (2011) Comparison of regression tree data mining methods for prediction of mortality in head injury. Expert Syst Appl 38:15534–15539 . doi: 10.1016/j.eswa.2011.06.006
5. Godinho S, Guiomar N, Gil A (2016) Using a stochastic gradient boosting algorithm to analyse the effectiveness of Landsat 8 data for montado land cover mapping: Application in southern Portugal. Int J Appl Earth Obs Geoinf 49:151–162 . doi: 10.1016/j.jag.2016.02.008
6. Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Stat Med 22:1365–1381
7. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso Robert Tibshirani. J R Stat Soc Ser B Stat Methodol 58:267–288 . doi: 10.1111/j.1467-9868.2011.00771.x
8. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Math Intell 27:83–85 . doi: 10.1198/jasa.2004.s339