Expert Systems With Applications 69 (2017) 62–73

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Data classification with binary response through the Boosting algorithm and logistic regression

Fortunato S. de Menezes[a,b,*], Gilberto R. Liska[b], Marcelo A. Cirillo[b], Mário J.F. Vivanco[b]

[a] Departament of Physics (DFI), Federal University of Lavras (UFLA), P.O.Box 3037, ZIP:37200-000, Lavras, MG, Brazil
[b] Departament of Statistics (DES), Federal University of Lavras (UFLA), P.O.Box 3037, ZIP:37200-000, Lavras, MG, Brazil

ARTICLE INFO

Article history:
Received 14 September 2015
Revised 11 April 2016
Accepted 2 August 2016
Available online 13 September 2016

Keywords:
Boosting algorithm
Data classification
Logistic regression
Information criteria
AIC
BIC
Selection of models
Monte Carlo Simulation

ABSTRACT

The task of classifying is natural to humans, but there are situations in which a person is not best suited to perform this function, which creates the need for automatic methods of classification. Traditional methods, such as logistic regression, are commonly used in this type of situation, but they lack robustness and accuracy. These methods do not not work very well when the data or when there is noise in the data, situations that are common in expert and intelligent systems. Due to the importance and the increasing complexity of problems of this type, there is a need for methods that provide greater accuracy and interpretability of the results. Among these methods, is Boosting, which operates sequentially by applying a classification algorithm to reweighted versions of the training data set. It was recently shown that Boosting may also be viewed as a method for functional estimation. The purpose of the present study was to compare the logistic regressions estimated by the maximum likelihood model (LRMML) and the logistic regression model estimated using the Boosting algorithm, specifically the Binomial Boosting algorithm (LRMBB), and to select the model with the better fit and discrimination capacity in the situation of presence(absence) of a given property (in this case, binary classification). To illustrate this situation, the example used was to classify the presence (absence) of coronary heart disease (CHD) as a function of various biological variables collected from patients. It is shown in the simulations results based on the strength of the indications that the LRMBB model is more appropriate than the LRMML model for the adjustment of data sets with several covariables and noisy data. The following sections report lower values of the information criteria AIC and BIC for the LRMBB model and that the Hosmer–Lemeshow test exhibits no evidence of a bad fit for the LRMBB model. The LRMBB model also presented a higher AUC, sensitivity, specificity and accuracy and lower values of false positives rates and false negatives rates, making it a model with better discrimination power compared to the LRMML model. Based on these results, the logistic model adjusted via the Binomial Boosting algorithm (LRMBB model) is better suited to describe the problem of binary response, because it provides more accurate information regarding the problem considered.

1. Introduction

In many situations a researcher is faced with the need to perform a data classification. This is especially the case when the sample size under consideration present some type of disturbance, so that conventional statistical methods may present unacceptable error classification rates.

Bearing this in mind, a plausible alternative can be achieved by a combination of computational methods and statistical techniques. This problem can be resolved by constructing an automatic classifier, which uses data from the problem at hand to create a rule to classify other data (independent from the previously shown data) in the future. The way this rule is created directly influences aspects such as the performance and interpretability of the classifier.

It is worth noting that when using the statistical technique of logistic regression in situations involving classification, the response to a particular phenomenon does not constitute a continuing situation, i.e., it admits the existence of categories, which may take two or more values. In these cases, logistic regression, whose parameter estimation is performed through maximum likelihood, has been applied frequently, and it returns the probability of a particular event occuring, as estimated using a logistic model. Logistic

* Corresponding author. Fax: +55(35)3829 1961.
E-mail addresses: fmenezes@dfi.ufla.br, fsdmenezes@gmail.com (F.S. de Menezes), gilbertoliska@unipampa.edu.br (G.R. Liska), macufla@dex.ufla.br (M.A. Cirillo), ferrua@dex.ufla.br (M.J.F. Vivanco).

http://dx.doi.org/10.1016/j.eswa.2016.08.014
0957-4174/

regression assumes quite interpretable rules, but with restrictive forms for the relationship between the predictor variables and responses.

Recently Skurichina and Duin (2002), bagging, boosting and the random subspace method have become popular combination techniques for improving weak classifiers. These techniques are designed for, and usually applied to, decision trees. It is shown that the performance of these combination techniques is strongly affected by the small sample size of the base classifier: boosting is useful for large training sample sizes, while bagging and the random subspace method are useful for critical sample sizes. Other results (Dietterich, 2000; Gey & Poggi, 2006; Kalai, 2005) show an experimental comparison of the three ensemble methods of bagging, boosting and randomization. It is show that in situations of little or no classification noise, randomization is competitive with bagging but not as accurate as boosting. In situations with substantial classification noise, bagging is much better than boosting, and sometimes better than randomization.

To improve the interpretability and performance of classification methods applied to a variety of problems, the Boosting algorithms, inspired by statistical physics and computer science, operate by sequentially applying a classification algorithm to a set of versions reweighted training data, providing greater weight to observations misclassified in the previous step. The Boosting algorithm were introduced by Schapire (1990) and since then, several variants have been created. Recently, Friedman (2001) showed that boosting may also be viewed as a method for functional estimation and can be used to estimate a logistic regression model. The purpose of this paper is to analyze the performance of the Boosting algorithm, specifically the Binomial Boosting algorithm (LRMBB) in classification problems involving binary responses compared to the logistic regression model estimated by the maximum likelihood method (LRMML). In addition, the main issues of the statistical approach of the Boosting algorithm will be presented. Sections 1.1 and 1.2 present logistic regression and the Gradient Boosting algorithm. In Section 1.3 the quality criteria for adjustment are presented. In the Section 2 , an example (CHD data) and the methodology are presented. In Section 3.1 simulation results show the strengh of the LRMBB algorithm in comparison to the LRMML algorithm, and Sections 3.2 and 3.3 show the results based on the trained and test data sets. Section 3.4 shows the odds ratio results applied in the models studied, showing that superior discrimination is obtained using the LRMBB model (Boosting algorithm). In the Discussion (Section 3.5) summarizes and compares the results, and the Conclusions (Section 4) states that the problem of binary classification is resolved in a more reliable fashion with superior discrimination using the Binomial Boosting (LRMBB) algorithm rather than the logistic regression (LRMML) algorithm.

1.1. Logistic regression

In linear regression models with single or multiple independent variables X, the dependent variable Y is a continuous random variable in nature. However, in some situations, the dependent variable is qualitative and expressed by two or more categories, in other words, it admits two or more values. In this case, the method of least squares does not provide plausible estimators. A good approximation is obtained by logistic regression, which allows the use of a regression model to calculate or predict the likelihood of a specific event ($\pi(\boldsymbol{x})$) (Atkinson, 1985).

The following presents the binary logistic regression model, which is a particular case of a generalized linear model, more specifically, the *logit* models .

To analyze $\pi(\boldsymbol{x})$, the independent observations $x_1, x_2, \ldots, x_n$ are made. In this context, it is reasonable to assume, as an initial assumption, that $\pi(\boldsymbol{x})$ is a monotonic function with values $0 < \pi(x) < 1$, i.e., $\pi(\boldsymbol{x})$ is a probability distribution function.

Because $\pi(\ \cdot\ )$ ranges between zero and one, a simple linear representation for π over all possible values of $\mathbf{x}$ is not adequate, because its values are linear in the range $(-\infty, +\infty)$. In this case, a transformation must be used to allow for any value of $\boldsymbol{x}$ to have a corresponding value in the range [0, 1]. Considering the logistic transformation, also called *logit*, then

$$logit = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

The ratio $\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$, called chance ($odds$) ranges from $(0; +\infty)$. Then, ($\log_e(odds)$) ranges from $(-\infty; +\infty)$.

Naturally, from Eq. 1, we have

$$e^{\log it} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

The inverse of the logit function (Eq. 1) is the logistic function, given by

$$\pi(\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \tag{2}$$

where $\pi(\boldsymbol{x})$ varies in [0; 1].

In the case where we have an explicative variable in the model, x_1, if $\beta_1 > 0$, π is increasing, and if $\beta_1 < 0$, π is decresing. The case where $\beta_0 = 0$ and $\beta_1 = 0$ corresponds to $\pi(x) = 0,5$.

The estimation vector $\boldsymbol{\beta}$ of the parameters is obtained through the method of maximum likelihood (Hosmer & Lemeshow, 1989).

1.2. Gradient Boosting Friedman Algorithm

Friedman and Hastie (2000) and Friedman (2001) developed more generally, a structure that leads to the direct statistical interpretation of boosting as a method for functional estimation.

In the context of boosting, the objective function is to estimate an optimal prediction of $f^*(\ \cdot\ )$, also called the minimizer population, which is defined by

$$f^*(\cdot) = \arg\min_f E_{Y,X}[\rho(Y, f(\boldsymbol{X}))] \tag{3}$$

where $\rho(\ \cdot, \cdot\ )$ is the loss function which is assumed as differentiable and convex with respect to f. In practice, we work with realizations $(y_i, \boldsymbol{x}_i^T)$, $i = 1, \ldots, n$, of $(\boldsymbol{y}, \boldsymbol{x}^T)$, and the expectation on Eq. 3 is therefore not known. For this reason, instead to minimize the expected value of Eq. 3, the Boosting algorithms instead minimize the observed average loss, which is given by $n^{-1}\sum_{i=1}^{n}\rho(Y_i, f(X_i))$, following interactively the functional space of the parameters of f. The following algorithm was presented by Friedman (2001), and is also called Gradient Boosting Friedman Algorithm.

1. Initialize $\hat{f}^{(0)}(\cdot)$ with a initial guess. Usual choices are

$$\hat{f}^{(0)}(\cdot) = \arg\min_c \frac{1}{n}\sum_{i=1}^{N}\rho(Y_i, c)$$

or $\hat{f}^{(0)}(\cdot) = 0$. Set $m = 0$.

2. Increase m by 1. Calculate the negative gradient $-\frac{\partial}{\partial f}\rho(Y, f)$ and calculated in $\hat{f}^{(m-1)}(\boldsymbol{X}_i)$:

$$z_i = -\frac{\partial}{\partial f(\boldsymbol{x}_i)}\rho(Y_i, f(\boldsymbol{x}_i))\,|_{f(\boldsymbol{x}_i)=\hat{f}^{(m-1)}(\boldsymbol{x}_i)} \qquad i = 1, \ldots, n$$