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/© 2016 Published by Elsevier Ltd.

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

$$\begin{aligned} e^{\log it} &= e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} \\ \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} &= e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} \end{aligned}$$

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 $\hat{f}^{(m-1)}(\boldsymbol{X}_i)$:

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

3. Adjust the negative gradient vector $z_1, \ldots, z_n$ for $X_1, \ldots, X_n$ by a base procedure $\hat{g}^{(m)}(\cdot)$ of real value (for instance, regression).
4. Actualize
   $\hat{f}^{(m)}(\cdot) = \hat{f}^{(m-1)}(\cdot) + \nu \cdot \hat{g}^{(m)}(\cdot)$
   where $0 < \nu \leq 1$ is the factor *step length*.
5. Continue the iteration process between the steps 2 to 4 until $m = M$, for the stop iteration M.

The stop iteration (M), which is the main factor of control, may be determined through cross-validation or by some information criterion. The choice of *step lenght* (ν) in step 4 is of minor importance, but it is recommended that it should be small, as $\nu = 0, 1$. A small value of ν typically requires a larger number of boosting iterations and therefore a longer computation time. For values of ν “sufficiently small”, empirical results show that the predictive accuracy of the model increases in comparison to other values of ν (Buhlmann & Hothorn, 2007).

Several Boosting algorithms may be defined by specifying different loss functions $\rho(\ \cdot, \cdot\ )$. The present work is applied to the situation where the response is binary and, as it will be viewed in the following, the Boosting algorithm suitable for this case is that of *Binomial Boosting*.

For the binary classification case, the response variable $Y \in \{0, 1\}$ with $p(\boldsymbol{x}) = P[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$. Sometimes, for the question of computational efficiency it is more convenient to code the response as $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$. Consider the negative of the log-likelihood of the binomial as the loss function:

$$\rho(y, p(\boldsymbol{x})) = -[y \ln p(\boldsymbol{x}) + (1 - y) \ln (1 - p(\boldsymbol{x}))] \tag{4}$$

$$p(\boldsymbol{x}) = \frac{e^{f(\boldsymbol{x})}}{e^{f(\boldsymbol{x})} + e^{-f(\boldsymbol{x})}} \tag{5}$$

such that

$$f(\boldsymbol{x}) = \frac{1}{2} \ln \left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} \right) \tag{6}$$

is equal to the half of log of the chance (*log-odds*). Then, the binomial loss is given by

$$\rho(y, f(\boldsymbol{x})) = \ln \left(1 + e^{-2\tilde{y}f}\right) \tag{7}$$

which is the upper limit of the error for misclassification, also known as the *step function*. The difference between the losses (4) and (7) is that (4) depends on $p(\boldsymbol{x})$ and by replacing the $p(\boldsymbol{x})$ (given by Eq. (5)) in Eq. (4) and recoding Y by $\tilde{Y}$, we obtain that the loss in Eq. (7), which depends on *f*.

It can be showed that the populational minimizer of the binomial loss in Eq. (7) is given by

$$f^{*}(\boldsymbol{x}) = \frac{1}{2} \ln \left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} \right) \tag{8}$$

where the boosting estimation $p(\boldsymbol{x})$ is defined as above (Friedman & Hastie, 2001b).

We interpret the boosting estimative $\hat{f}^{(m)}(\cdot)$ as a estimation of the populational minimizer $f^{*}(\ \cdot\ )$. Then, the result of Binomial Boosting is estimates of the half of the likelihood. In particular, we define the estimates of the probabilities through the probability

$$\hat{p}(\boldsymbol{x}) = \frac{e^{\hat{f}^{(m)}(\boldsymbol{x})}}{e^{\hat{f}^{(m)}(\boldsymbol{x})} + e^{-\hat{f}^{(m)}(\boldsymbol{x})}} \tag{9}$$

The reason to construct these estimates of the probabilities is based on the fact that boosting with a reasonable stop iteration is consistent (Bartlett & Traskin, 2007).

Boosting can be used to fit generalized linear models in higher dimensions. Consider the basic procedure

$$\hat{g}(x) = \hat{\beta}^{\left(\hat{\lambda}\right)} x^{\left(\hat{\lambda}\right)} \tag{10}$$

where

$$\hat{\beta}^{(j)} = \frac{\sum_{i=1}^{n} X_i^{(j)} z_i}{\sum_{i=1}^{n} \left(X_i^{(j)}\right)^2} \tag{11}$$

and

$$\hat{\lambda} = \underset{1 \leq j \leq p}{\arg\min} \sum_{i=1}^{n} \left(z_i - \hat{\beta}^{(j)} X_i^{(j)}\right)^2 \tag{12}$$

Performing this procedure accomplishes automatically the process of selection of the variables in a regression model with multiple variables. For this reason and using the procedure based in Eq. (10), it can be said that the procedure of the selection of variables is embedded in this algorithm Gradient Boosting Friedman algorithm (Friedman, 2001).

When using Binomial Boosting with this base procedure, we selected in each iteration a variable predictor, not necessarily different for each iteration, and update the function linearly,

$$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \nu \cdot \hat{\beta}^{\left(\hat{\lambda}_m\right)} x^{\left(\hat{\lambda}_m\right)} \tag{13}$$

where $\hat{\lambda}_m$ denotes the index of the predictor variable in iteration *m*.

When using Binomial Boosting with the least squares linear component by component (see Eq. (11) and the loss function (Eq. 7), we obtain an adjustment, including variable selection, of a linear logistic regression model (Buhlmann & Hothorn, 2007).

Table 1
Truth table.

Observed	Prediction of the model	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

1.3. *Criteria of adequability of adjust*

1.3.1. *The ROC curve*

One way to evaluate the performance of models with a binary response is to determine the number of correct answers the model provides. The success of the model can be evaluated using the ROC (*Receiver Operating Characteristic*) curve. The ROC curve is a graphical plot of the *sensitivity* (proportion of true positives) of the predictive model versus the complement of the *specificity* (i.e., proportion of false positives), in a series of thresholds for a positive result (Hanley, 1989). The logistic model returns as a result the probability of a specific event, in our case, the likelihood of a person having a coronary heart disease (CHD). This probability can be converted to a binary result according to the choice of a threshold. We can summarize the results of this procedure in a table, known as a truth table (Table 1).

The *sensitivity* is defined as the ability of a model to find positive answers, in other words, people who actually have CHD, so $sensitivity = \frac{TP}{TP+FN}$, where *TP* is the number of true positives and *FN*, the number of false negative predicted by the model. The *specificity* of model is defined as the proportion of true negatives predicted by the model. The *specificity* of a model is defined as the proportion of true negatives predicted by the model, in other words, the proportion of people who actually do not have CHD but the model predicts as positive, so $specificity = \frac{TN}{TN+FP}$, where *TN* is the quantity of true negatives and *FP* the quantitu of false positives predicted by the model. Thus, we can obtain the *accuracy* of the model, which measures the capacity of the predictive model to correctly classify people who have and do not have CHD, is given by *accuracy* $= \frac{TP+TN}{TP+FN+TN+FP}$. The complement of the *specificity* is the *false positives ratio*, defined as the proportion of incorrect predictions of positives results (event of interest) in relation to the total number of negatives (the complement of the event of interest) observed. Similarly, the complement of the *sensitivity* (i.e., 1 - *sensitivity*) is the *false negative ratios*, i.e., the proportion of incorrect negative predictions in relation to the total number of positives. Note that the sum of the *sensitivity* and the *false negatives ratio* must be 1. The same applies to the sum of the *specificity* and the *false positive ratios*. It is immediately clear that the sum of the *(sensitivity)* and *(1-specificity)* must be 1.

The area under the ROC curve (AUC - *area under curve*) is calculated by the trapezoid rule, namely,

$$AUC = \sum_{i=1}^{n} (x_{i+1} - x_i)\left(\frac{y_{i+1} + y_i}{2}\right) \qquad (14)$$

where i is the threshold of the curve from which the pair of points $(x_i,\ y_i)$ are taken. The AUC measures the discriminatory power of the predictive model, i.e., the success of the model in correctly classifing TP and TN. As a general rule, an acceptable discrimination occurs when the AUC $>$ 0, 7 u.a., and if AUC $\geq$ 0, 8 u.a., the discrimination is said to be excellent (Zhou & Mcclish, 2008).

1.3.2. Hosmer–Lemeshow test

This test is used to verify the quality of adjustment of logistic regression models. It was proposed by Hosmer and Lemeshow (1989), where the evaluation of the statistics of the test can be expressed as supposing that $n = J$ when we have n estimated probabilities. To perform the thes, we first order the n estimated probabilities based on the deciles of their estimated probabilities. We use g $=$ 10 groups in which the first $n'_{1=N/10}$ are those containing the smallest estimated probabilities and $n'_{10} = N/10$ are those with the highest estimated probabilities. We then estimate the expected frequency for $Y = 1,$ which is given by the sum of the estimated probabilities of all individuals within that group. For $Y = 0,$ the expected frequency estimate is given by the sum of the complement of the estimated probabilities of all individuals within that group. Thus, the Hosmer and Lemeshow statistic, $\hat{C},$ is obtained as

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k\bar{\pi}_k)^2}{n'_k\bar{\pi}_k(1 - \bar{\pi}_k)}, \qquad (15)$$

where n'_k is the number of individuals in the k-th group, $\bar{\pi}_k = \sum_{j=1}^{C_k} \frac{m_j\bar{\pi}_j}{n'_k}$, C_k is the total number of combinations of levels inside the $k-$th decile and $O_k = \sum_{j=1}^{C_k} y_j$ is the total number of responses inside the kth group.

The statistical test of Hosmer and Lemeshow has chi-square distribution with (g-2) degrees of freedom (Hosmer & Lemeshow, 1989). The null hypothesis of the test corresponds to a quite satisfatory fit of the model quite satisfactory.

1.3.3. Information criteria: Akaike and Bayesian

The Akaike information criterion (AIC) proposed in Akaike (1974), is a relative measure of the quality of fit of a statistical model. It is defined as $AIC = -2\,l(\,\theta\,|y\,) + 2p,$ where $l(\,\theta\,|y.)$ is the natural logarithm of the likelihood function of the model in θ, and p is the number of parameters of the model. Schwarz (1978) proposed a criterion known as the Bayesian information criterion (BIC), which corresponds to the exchange factor 2, which is the weight of the number of parameters in AIC by $\ln(n)$. Then the BIC is given by $BIC = -2\,l(\,\theta\,|y\,) + p\,\ln(n),$ where n is the number of observations in the sample. Given a set of models for the data, the preferred model is the one that exhibits the lowest AIC or BIC value, namely the smaller the AIC or BIC value, the better the fit of the model to the data (Akaike, 1974), (Emiliano & Vivanco, 2009), (Emiliano & Vivanco, 2014).

2. Methodology

2.1. Data

The study used data provided by the *UCI Machine Learning Repository* (Frank & Asuncion, 2010). The data are related to 270 patients with the presence (absence) of coronary heart disease (Coronary Heart Disease - CHD) and this condition as a function of 13 independent variables. On Table 2 presents these variables, including the nature of each of them, and the possible values they can take.

The response that it is intended to model is the condition of the presence(absence) of coronary heart disease (CHD), where the representation is given by *DIS*. The value $DIS = 1(DIS = 0)$ corresponds to the presence (absence) of CHD in patients. In addition to the answer, there exists three variables of a binary nature, which are the independent variables SEX, SUG and EXE. The variable SEX defines about the sex of a patient (0: female, 1:male). The SUG variable is related to the glycemic level in the blood of a person (0: $\leq$ 120 mg/dL; 1: 120 mg/dL), and the variable EXE is related to the situation of induced angina, which is a condition in which a person may feel chest pain even in a resting situation (0: no pain; 1: pain). There also exists three explicative variables of a nominal nature, namely: PAIN refers to the type of chest pain, which can be classified in four different forms (1: typical angina; 2: atypical angina; 3: no angina pain; 4: asymptomatic. The variable ELE is related to the behavior of the ST segment on electrocardiogram, where their levels 2 and 3 are indicative of CHD (1:normal; 2: with abnormal ST-T wave ST-T; 3: showing probable hypertrophy of the left ventricle). The variable THAL represents the presence of thalassemia, which is an hereditary disease which affects the blood (3: normal; 6: defect; 7: reversible defect). The variable SLOPE is related to the slope of the ST segment, which is the segment of the electrocardiogram used to diagnose acute ischemic events, and as it is an ordinary variable, the three levels describe conditions that express the probabability of ischemia (1: ascending slope; 2: horizontal slope; 3: descending slope). The variable VES, whose nature is discrete, represents the number of large veins colored by fluoroscopy (0,1,2 or 3). The other variables are of continuous nature and represent the age of the patient (AGE) in years, the arterial resting pressure (PRESS) in mm/Hg, the serum cholesterol levels of the blood (COL) in mg/dL, the maximum cardiac pressure achieved (HEART) in bpm units and the length of the ST segment of the electrocardiogram in milimeters (ST).

2.2. Adjustment of logistic regression models

To estimate the parameters of the logistic regression model using the Boosting algorithm, the Binomial Boosting algorithm (LRMBB) will be used. To execute the Binomial Boosting algorithm, it is necessary to define two components: (i) the loss function and (ii) the base procedure (defined in Section 1.2). The Binomial Boosting algorithm uses the binomial loss function and the base procedure of least squares component by component, as the DIS response configures a binary situation, and we are interested in adjusting a generalized linear model.

The Binomial Boosting algorithm, during the process of parametric estimation, already use the selection of variables, returning, therefore, those independent variables which minimize the loss function used, leading to a model with independent variables which contributes significantly to the model.

Table 2
Relation of variables presented in the problem of diagnostic of coronary heart disease (CHD).

Variable	Nature	Description
AGE	Continue	in years
SEX	Binary	0: female 1: male
PAIN	Nominal; 4 levels	1: tipical angina 2: atipical angina 3: painless angina 4: asymptomatic
PRESS	Continue	em mm/Hg
COL	Continue	em mg/dl
SUG	Binary	0: $\leq$ 120mg/dL 1: $>$ 120mg/dL
ELE	Nominal; 3 levels	1: normal 2: with ST-T anormal 3: showing probable hypertrophy of left ventricle
HEART	Continue	in bpm
EXE	Binary	0: no 1: yes
ST	Continue	in milimetros
SLOPE	Ordinary; 3 levels	1: Ascendent slope 2: horizontal slope 3: Descendent slope
VES	Discrete	0, 1, 2, ou 3
THAL	Nominal; 3 levels	3: normal 6: defect 7: reversible defect
DIS	Binary	0: absent for CHD 1: present for CHD

To estimate the parameters of the logistic regression model through maximum likelihood (LRMML) the method described in Section 1.1 was used. In the following, the *stepwise* selection method of variables through AIC was used, with the purpose of eliminating the independent variables that do not contribute significantly to the probability of occurrence of coronary cardiac disease (CHD).

2.3. Comparison of LRMBB and LRMML

To evaluate the performance of the models obtained with the two methods, the data were split into two parts: (i) the training part, used to estimate of the parameters of the LRMML and LRMBB models, and (ii) the test part, which is aimed at the validation of LRMML and LRMBB models. The training set is constituted by partitions of 30, 40, 50, 60, 70, 80 and 90% of the original sample, which contains 270 patients. The complement of the partitions will constitute the test set. The validation is performed by comparing the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) on the models obtained after the process of the selection of the variables, and the model that is preferred is the one whose information criteria are smaller.

The Hosmer–Lemeshow test (Hosmer & Lemeshow, 1989) will be used to verify the existence of problems in the LRMML and LRMBB model adjustment. The choice of the optimal partition (training set and test set) will be the one where the result of the Hosmer–Lemeshow test is not significant for the LRMML and LRMBB models.

To determine the threshold appropriate to classify a patient according the presence(absence) of CHD, the ROC curve will be used on both models, LRMML and LRMBB.

For each of the models LRMML and LRMBB estimated with the optimal partition, the sensitivity, specificity, accuracy, false negative rate, false positive rate and AUC will be calculated, anf the model that presents the best values for these quantities wiil be selected.

The odds ratio of the occurrence of CHD for all independent variables, adjusted by the LRMML and LRMBB, will be calculated.

The results obtained using the proposed methodology above, were performed on the R statistical computing system (R core team, 2011).

3. Results and discussion

3.1. Monte Carlo simulation study of accuracy and precision: LRMBB and LRMML models

To study the accuracy and precision performance of the LRMBB and LRMML models, we simulate a logistic model where each covariable is considered a mixture given by

$$X_j = (1-\delta)\,N(0,1) + \delta\, log\,Normal(0,1) \tag{16}$$

where δ was fixed with values $\delta = 0.05, 0.10$ and 0.20

Figs. 1–3 show that the estimates of the logistic regression model obtained by the Boosting method (LRMBB model) are more accurate, and more precise (lower error) with increasing sample size (n). This is apparent for all parametric values evaluated. For the same scenarios, however, considering the maximum likelihood estimation method (LRMML model), the results illustrated in Figs. 4–6, indicate that the estimates of the logistic model remain accurate but are less precise (i.e., errors remain high even for a large sample size n). Given this fact, considering that the distribution of the covariable shows outliers resulting from the lognormal distribution characterized by excess kurtosis, contradictory results with respect to the estimation method were observed. Thus, we note that the Boosting method (LRMBB model) presented itself robust to extreme observations in the covariable, whereas the maximum likelihood method (LRMML model) was sensitive to those observations (i.e., this model is sensitive to the fraction of outliers (δ) present in the distribution as show in Figs. 4–6).

Based on the robust results of the estimates of the the logistic model parameters, the implementation of the logistic model was performed, considering the Boosting method for the purpose of

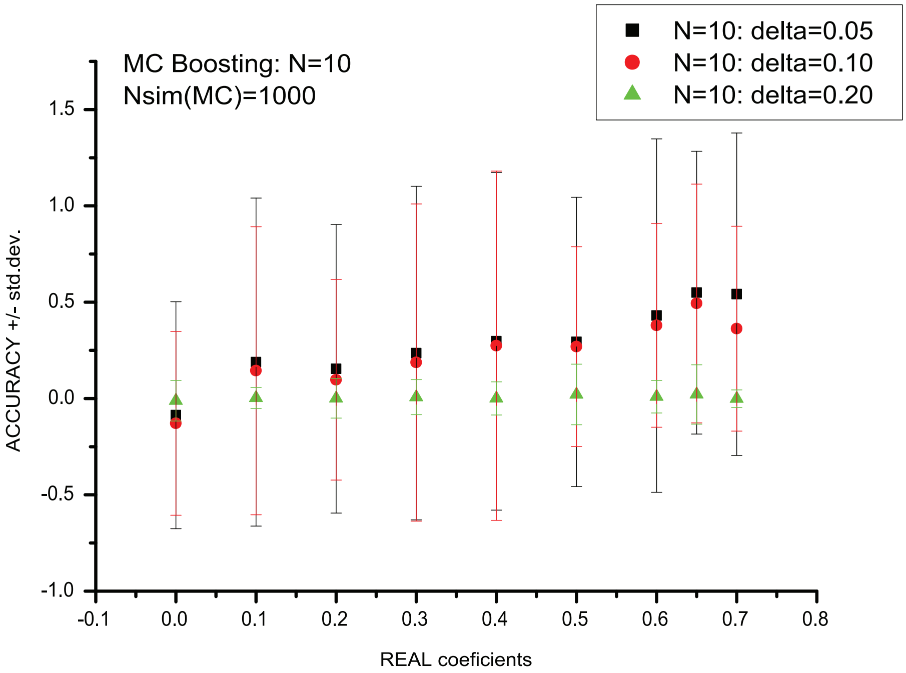

Fig. 1. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates parameters of the logistic regression model obtained by boosting method considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 10$.

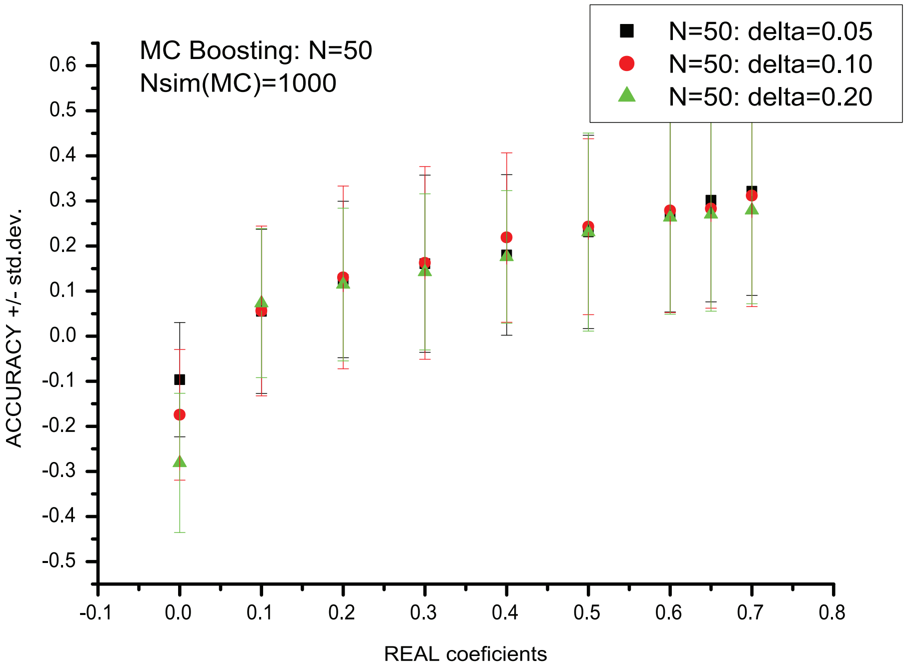

Fig. 2. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates parameters of the logistic regression model obtained by boosting method considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 50$.

classifying the presence (absence) of CHD based of the estimation of the parameters of the linear regression model for the application described in Section 3.2.

3.2. Training and test set

The variable *DIS* represents the situation of success or failure of an event, which can be associated with a Bernoulli random variable. The complete model for this situation (see Eq. 2) is given by

$$P(DIS = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}} \qquad (17)$$

where

$$\begin{aligned} g(\boldsymbol{x}) =& \beta_0 + \beta_1 AGE + \beta_{21} SEX_1 + \beta_{22} SEX_2 \\ &+ \beta_{31} PAIN_1 + \beta_{32} PAIN_2 + \beta_{33} PAIN_3 \\ &+ \beta_{34} PAIN_4 + \beta_4 PRESS + \beta_5 COL \\ &+ \beta_{61} SUG_1 + \beta_{62} SUG_2 + \beta_{71} ELE_1 \\ &+ \beta_{72} ELE_2 + \beta_{73} ELE_3 + \beta_8 HEART \end{aligned}$$

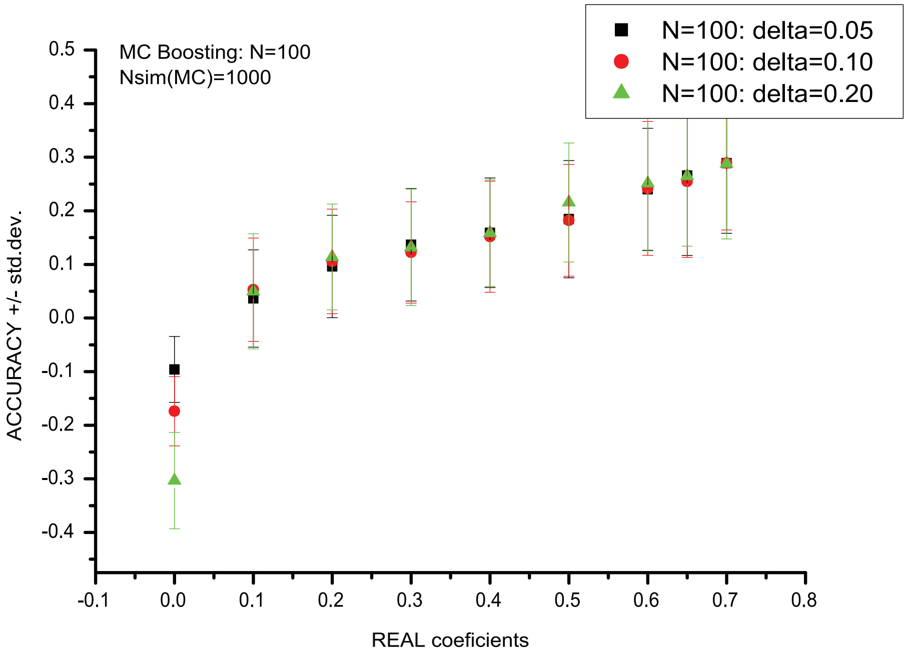

Fig. 3. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates parameters of the logistic regression model obtained by boosting method considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 100$.

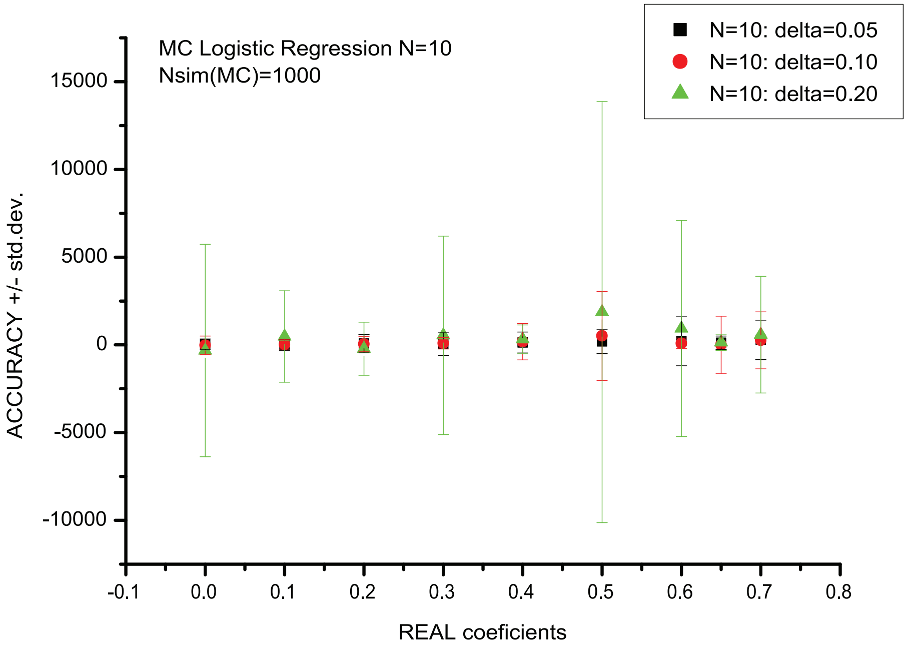

Fig. 4. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates of parameters by logistic regression model obtained by Maximum likelihood estimation considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 10$.

$$
\begin{aligned}
&+\beta_{91}EXE_1 + \beta_{92}EXE_2 + \beta_{10}ST \\
&+\beta_{111}SLOPE_1 + \beta_{112}SLOPE_2 \\
&+\beta_{113}SLOPE_3 + \beta_{12}VES + \beta_{131}THAL_1 \\
&+\beta_{132}THAL_2 + \beta_{133}THAL_3
\end{aligned}
$$

where the categorical independent variables *SEX, PAIN, SUG, ELE, EXE, SLOPE* and *THAL* are of type dummy (assume levels of factors), and it is assumed that the first levels of each of these independent variables is zero, reporting, therefore, to the model in agreement with the first level of each factor.

Table 3 presents the results of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for several cuts in

Table 3
Results of Akaike information criterion (AIC) and Bayesian information criterion (BIC) on several training and test set.

Slice		AIC		BIC	
Train. (%)	Test (%)	LRMBB	LRMML	LRMBB	LRMML
30,00	70,00	62,4032	76,7070	73,6052	96,1570
40,00	60,00	77,8526	90,2510	91,2616	118,0397
50,00	50,00	98,1087	107,7300	113,2539	139,6903
60,00	40,00	123,3942	130,7300	137,8766	160,2329
70,00	30,00	144,1786	154,4300	160,4283	180,7688
80,00	20,00	160,2569	171,7100	177,5306	195,5802
90,00	10,00	174,1674	189,1100	193,1337	203,3674

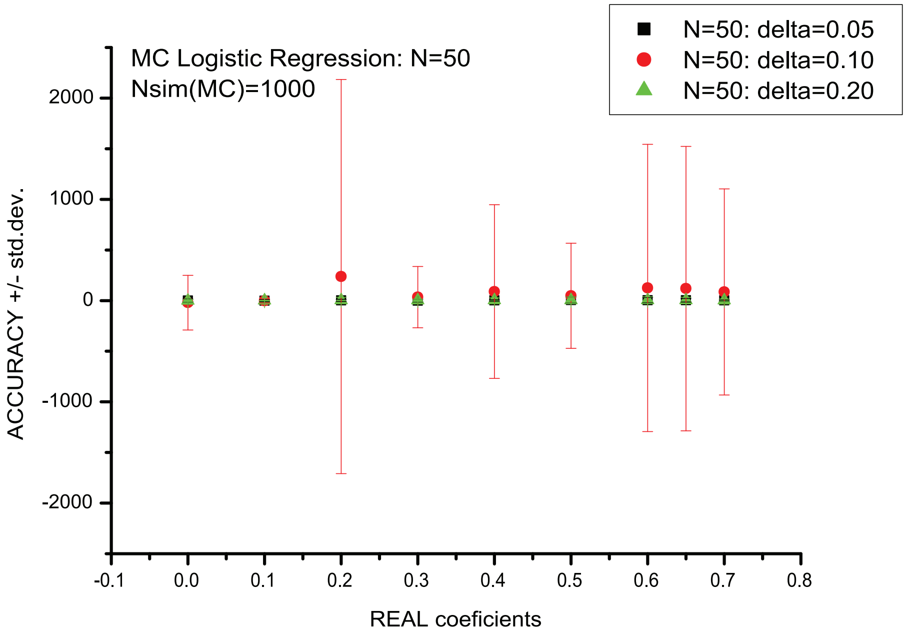

Fig. 5. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates of parameters by logistic regression model obtained by Maximum likelihood estimation considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 50$.

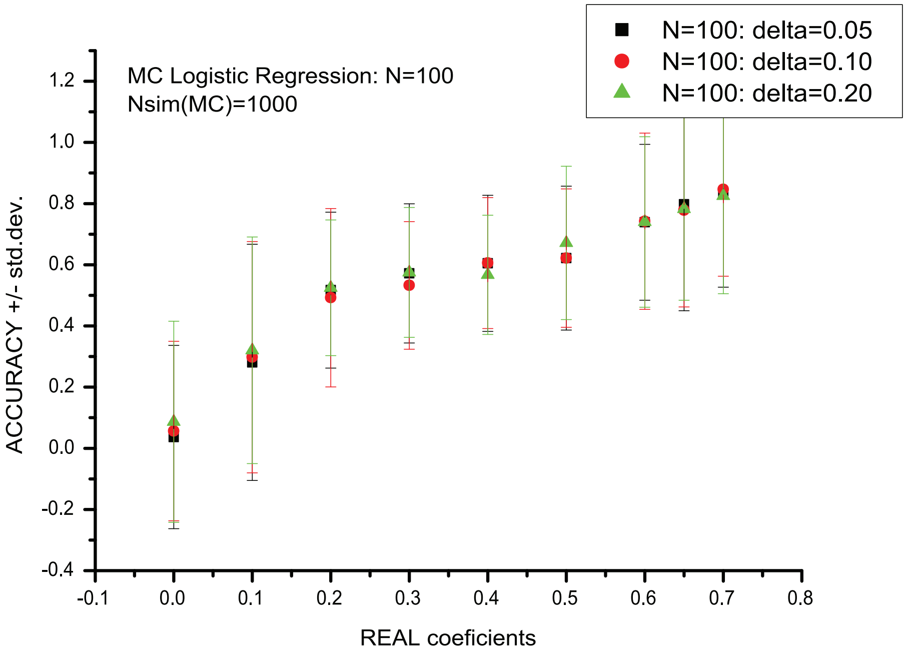

Fig. 6. Average (μ_{coef}) of coeficients plus one standard error (σ) of the MC estimates of parameters by logistic regression model obtained by Maximum likelihood estimation considering different fraction of mixtures rates ($\delta = 0.05$, $\delta = 0.1$ and $\delta = 0.2$) and sample size $n = 100$.

the dataset as well as the amount of data resulting from each cut for training and testing. It is observed that for all cuts, the LRMBB model had lower values of AIC and BIC,and is therefore more suitable.

It is observed that, considering a 5% of significance level (see Table 4), the LRMBB model performed for each cut indicates that the model fit is adequatefor the Hosmer–Lemeshow test. The same does not occur for the LRMML model on cuts of 30, 40, 50 and 60% for the training set, in which the null hypothesis of adequacy of fit is rejected at the level of 5% of significance.

Table 4
Results of the Hosmer–Lemeshow test (*valor-p*) in several training and test sets.

Slice		N		Hosmer–Lemeshow	
Train. (%)	Test (%)	train.	test	LRMBB	LRMML
30,00	70,00	81	189	0,4326	0,0030
40,00	60,00	108	162	0,5758	0,0001
50,00	50,00	135	135	0,5101	0,0001
60,00	40,00	162	108	0,5574	0,0265
70,00	30,00	189	81	0,1596	0,0506
80,00	20,00	216	54	0,2341	0,6549
90,00	10,00	243	27	0,7017	0,3996

Table 5
Estimation of parameter ($\hat{\beta}$) referring to logistic model fitted to the data on coronary heart disease.

Variable	β	LRMBB	LRMML	
		$\hat{\beta}$	$\hat{\beta}$	Standard error
Constant	β_0	−4,6268	−10,7509	2,5400
AGE	β_1	NA	NA	NA
SEX	β_{21}	0	0	–
SEX	β_{22}	0,7979	1,5066	0,5991
PAIN	β_{31}	0	0	–
PAIN	β_{32}	NA	NA	–
PAIN	β_{33}	NA	2,0030	1,0524
PAIN	β_{34}	1,5284	3,9293	1,0390
PRESS	β_4	0,0107	0,0336	0,0131
COL	β_5	0,0028	NA	–
SUG	β_{61}	0	0	–
SUG	β_{62}	NA	NA	–
ELE	β_{71}	0	0	–
ELE	β_{72}	NA	NA	–
ELE	β_{73}	0,2730	NA	–
HEART	β_8	−0,0053	NA	–
EXE	β_{91}	0	0	–
EXE	β_{92}	0,5502	NA	–
ST	β_{10}	0,3762	NA	–
SLOPE	β_{111}	0	0	–
SLOPE	β_{112}	0,6459	1,7693	0,4991
SLOPE	β_{113}	NA	1,9773	1,0305
VES	β_{12}	0,6967	1,0290	0,3270
THAL	β_{131}	NA	NA	-
THAL	β_{132}	NA	NA	-
THAL	β_{133}	1,1746	1,4205	0,4984

NA: Not Adjusted.

3.3. Proposed model

Taking as a reference the results in Tables 3 and 4, we now specify the proposed LRMBB and LRMML models, whose estimates of parameters are presented in Table 5. As seen in Table 3 and 4, a study was made of the behavior of the model estimated by both methods for different cuts in the data set. The literature recommends that the test set have enough observations to represent the training set. Thus, a partition of training (test) of 70% (30%) satisfies the level of significance for both models and will be applied hereafter.

The LRMBB model is the one that minimizes the loss function as shown in Sections 1.2. As this is an iterative method, each iteration of the algorithm estimates a model, and for this model is calculated their Akaike information criterion, so that the model that minimizes the loss function in this case is also the one that gives the lowest value of AIC (just as the BIC). Fig. 7 shows the evolution of the AIC as it increases with the number of iterations of the algorithm.

Thus, Fig. 7 illustrates that the optimal number of iterations of the Binomial Boosting algorithm is 146, which provides a value of 144.1786 for the AIC (see Table 3). It is also observed in Fig. 7 that the algorithm should not run indefinitely, as this, besides increasing the AIC value would force the inclusion of unimportant variables in the model. Therefore, the probability of an individual via LRMBB $\boldsymbol{x}_i$ having a coronary disease is estimated by the expression $\pi_{Boost}(\boldsymbol{x}_i)$ in Eq. (18).

$$P(DIS_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \pi_{Boost}(\boldsymbol{x}_i) = \frac{e^{g_{Boost}(\boldsymbol{x}_i)}}{1 + e^{g_{Boost}(\boldsymbol{x}_i)}} \tag{18}$$

where

$$\begin{aligned} g_{Boost}(\boldsymbol{x}_i) = &- 4,6268 + 0,7979\, SEX_{2i} \\ &+ 0,0107\, PRESS_i + 0,0028\, COL_i \\ &+ 0,2730\, ELE_{3i} - 0,0053\, HEART_i \\ &+ 0,5502\, EXE_{2i} + 0,3762\, ST_i \\ &+ 0,6459\, SLOPE_{2i} + 0,922\, VES_i \\ &+ 0,325\, THAL_{3i} + 1,5284\, PAIN_{4i} \end{aligned}$$

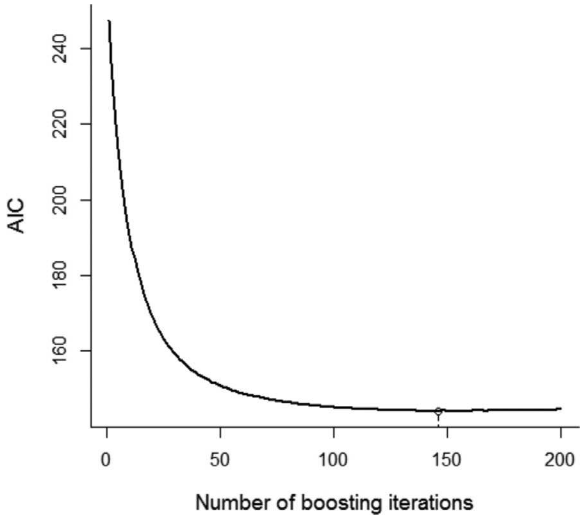

Fig. 7. Plot of the evolution of the Akaike information criterion over the number of iterations of the algorithm Binomial Boosting

Note that the Binomial Boosting algorithm selected 11 of the 13 independent variables in the final model. Therefore, the probability of the occurrence of CHD is neither influenced by the age (AGE) of the people nor by the level of their blood glucose (SUG). This model also explains that if the person is male, the probability of CHD cardiac is increased which is best explained by the increase in the odds ratio, the results of which are presented in the following section. The model also explains that the likelihood of a person having CHD increases if the person presents asymptomatic chest pain ($PAIN_4$), the result of the electrocardiogram at rest is classified as high (ELE_3), if the patient has a positive result for induced angina (EXE), if the slope of the ST segment during peak exercise exhibits horizontal inclination ($SLOPE_2$) and if there is a reversible defect for thalassemia ($THAL_3$). This likelihood is further increased with the continuous variables related to blood pressure (*PRESS*), cholesterol level (*COL*), cardiac frequency (*HEART*), long-segment elevation (*ST*) and number of large vessels colored by fluoroscopy (*VES*) (Table 5).

The probability of an individual $\boldsymbol{x}_i$ having a coronary disease via the LRMML model is estimated by the expression $\pi_{LR}(\boldsymbol{x}_i)$ in Eq. (19)

$$P(DIS_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \pi_{LR}(\boldsymbol{x}_i) = \frac{e^{g_{LR}(\boldsymbol{x}_i)}}{1 + e^{g_{LR}(\boldsymbol{x}_i)}} \tag{19}$$

where

$$\begin{aligned} g_{LR}(\boldsymbol{x}_i) = &- 10,7509 + 1,5066\, SEX_{2i} + 2,0030\, PAIN_{3i} \\ &+ 3,9293\, PAIN_{4i} + 0,0336\, PRESS_i \\ &+ 1,7693\, SLOPE_{2i} + 1,9773\, SLOPE_{3i} \\ &+ 1,0290\, VES_i + 1,4205\, THAL_{3i} \end{aligned}$$

Note that the method *stepwise* selected 6 of the 13 independent variables of the LRMML model (see Table 5). Thus, the probability of occurrence of CHD is not influenced by the following independent variables: age (*AGE*), your blood sugar level (*SUG*), cholesterol level, a result of the electrocardiogram (*ELE*), cardiac frequency (*HEART*), occurrence of induced angina (*SUG*) and long-segment elevation (*ST*), as these independent variables were not selected for the final model (Eq. 19) after the *stepwise* procedure. This model also explains that, if the person is male, the probability of coronary heart disease is increased as well as if the person has chest pain

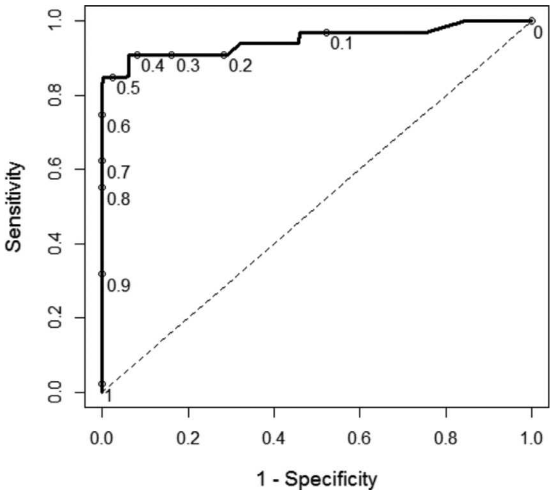

Fig. 8. ROC curve for LRMBB.

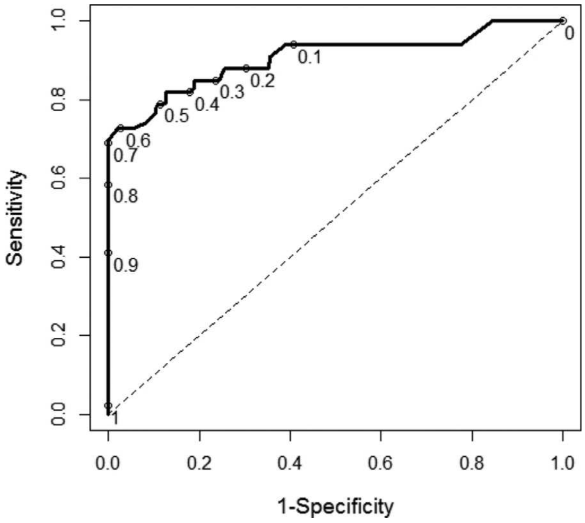

Fig. 9. ROC curve for LRMML.

Table 6
Confusion table for the LRMBB model adjusted to the CHD data.

Observed	Model	
	Presence	Absence
Presence	27	6
Absence	0	48

Table 7
Confusion table for the LRMML model adjusted to the CHD data.

Observed	Model	
	Presence	Absence
Presence	26	7
Absence	5	43

Table 8
Odds ratio (OR) estimates for the independent variables selected by the models and LRMBB LRMML, concerning to data on coronary heart disease.

Variavel	Parametro	MRLBB OR	MRLMV OR
SEX	β_{22}	2,2210	4,5112
PAIN	β_{33}	NA	7,4111
	β_{34}	4,6109	50,8701
PRESS	β_4	1,1133[a]	1,3986[a]
COL	β_5	1,0280[a]	NA
ELE	β_{73}	1,3139	NA
HEART	β_8	0,9480 [a]	NA
EXE	β_{92}	1,7336	NA
ST	β_{10}	1,4568	NA
SLOPE	β_{112}	1,9076	5,8667
	β_{113}	1,9692	7,2230
VES	β_{12}	2,0071	2,7982
THAL	β_{22}	3,2369	4,1393

[a] Odds ratio correpondent to increment of 10 unitis. NA: Not Adjusted.

that is probably not angina ($PAIN_3$) or is asymptomatic ($PAIN_4$), is the slope of the ST segment during peak exercise exhibits horizontal inclination ($SLOPE_2$) or descending ($SLOPE_3$) inclination or if the person exhibits a reversible defect for thalassemia ($THAL_3$). This likelihood is further increased with the continuous variables related to blood pressure (*PRESS*) and the number of major vessels colored by fluoroscopy (*VES*).

Once obtained the model to determine the occurrence of CHD is obtained, one can see the discrimination power of this model, i.e., its ability of the model to correctly classify individuals who do have CHD and those who do not have CHD. Figs. 8 and 9 shows the ROC curve of the LRMBB and LRMML, respectively. It is observed that both models have a high discriminative power, as the areas under each the ROC curve of models is 0.947 a.u. and 0.905 a.u., respectively.

Given the above in Section 1.3.1, another advantage of the ROC curve is the possibility of choosing a suitable threshold for the classification of patients according to the presence (absence) of CHD. In addition, Figs. 8 and 9 suggest that an appropriate threshold would be 0.5 for both models. Therefore, to evaluate the *sensitivity* and *specificity* of the model, the following criterion will be used to classify a patient as positive for the presence of CHD ($Y = 1$): when the probability of the occurrence of CHD is greater than 0.5 (50%). Otherwise will be classified as being negative for the presence of CHD ($Y = 0$).

Tables 6 and 7 and the definition of sensitivity (see Section. 1.3.1) show that the *sensitivity* of the LRMBB model is 82%,i.e., 82% of the CHD patients who actually have CHD are classified by the model as positive for this feature. The false negative rate of the model was 18%,i.e., 18% of the people who have CHD were falsely determined by the model to be negative. The false positive rate was 0%, so the model does not classified any patient who does not have CHD as positive and, as a consequence, the *specificity* was 100%. The accuracy of the model was 92.59%.

Similarly, it is observed that the *sensitivity* of LRMML was 79%. The rates of false negatives and false positives of the model were 21 and 10%, respectively. The *specificity* was 90%, i.e., of the set of patients who do not have CHD, 90% were classified as being in that condition. The accuracy of the model was 85.18%. (Table 7).

3.4. Odds ratio

One of the advantages of using a logistic regression model is to obtain the relationship between the probability of occurrence of CHD and a particular independent variable. This relationship is called the odds ratio, and Table 8 summarizes these values for each of the estimated parameters of the LRMBB and LRMML models (read as the final models the ones containing only the selected independent variables shown in Table 5).

Denote by $OR_{SEX,\ Boost}$ and $OR_{SEX,\ RL}$ the odds ratio of disease with relation to the coronary cardiac patient sex that are obtained via the LRMBB and LRMML models, respectively. The odds of a CHD positive patient among male and female patients is estimated by $\hat{OR}_{SEX,Boost} = \exp\{\beta_{22,Boost}\} = \exp\{0,7979\} = 2,2210$ and $\hat{OR}_{SEX,RL} = \exp\{\beta_{22,RL}\} = \exp\{1,5066\} = 4,5112$.

Then, $\hat{OR}_{SEX,Boost}$ indicates that a patient of the male sex has a 122,1% greater chance of having coronary heart disease than a female patient via the LRMBB model, while the same event occurs with a chance of 351.12% via the LRMML model.

Denote by $OR_{PAIN4,\ Boost}$ and $OR_{PAIN4,\ RL}$ the ratio of chance of a CHD positive patient and chest pain of type 4 in relation to a patient to a CHD positive patient and chest pain of type 1 as obtained via the LRMBB and LRMML models, respectively, i.e., $\hat{OR}_{PAIN4,Boost} = \exp\{\beta_{34,\ \mathrm{Boost}}\} = \exp\{1,5284\} = 4,6109$. Then, using Binomial Boosting, the chance of a patient with the chest pain of type 4 having CHD is 361,09% (almost 5 times higher) than of a patient who presents pains of type 1. In the same way, via LRMML, this probability is 4987,01% (almost 50 times higher!) than of the other, i.e., $\hat{OR}_{PAIN4,RL} = \exp\{\beta_{34,\ RL}\} = \exp\{3,9293\} = 50,8701$.

In the case of continuous variables, there is a slight difference in the interpretation of the odds ratio. In this case, every increase of one unit in this type of variable, results in a corresponding increase in the chance of a patient being diagnosed with CHD. In the case of the variable length associated with the ST segment, while keeping the other independent variables fixed, an increase of 1 mm in that segment implies an increase of 45.68% in the chance of a patient being classified with coronary heart disease.

However, the increase of one unit in some independent variables does not have much practical sense, as is the case with the covariate associated with the patient's blood pressure (*PRESS*). Thus, via the LRMBB model, an increment of 10 mm/Hg on this covariate this implies an increase of 11.33% chance of a patient being diagnosed with CHD. Likewise, via LRMML, the chance increases to 39.86%.

For the covariate associated with the maximum heart rate achieved (HEART), an increase of 10 bpm entails a decline of 5.2% in the chance of a patient having CHD via LRMBB.

3.5. Discussion

This paper presented a comparison of logistic regression model estimated via the Binomial Boosting Algorithm (LRMBB) and by the method of maximum likelihood (LRMML). In the literature, no studies were found that made this sort of comparison; therefor, this session will present a discussion of several studies that a Boosting type of algorithm and compared its performance with that of other classifiers.

Cai (2006) used the LogitBoost algorithm to classify various structures of proteins in molecular biology. The authors compared the efficiency of the LogitBoost algorithm with another method well known in the machine learning community, the method of Support Vector Machines method (*Support Vector Machines*), observing a superior performance by almost 9% over the Boosting Algorithm in the prediction of structural classes for a given dataset.

Cao (2010) compared the stochastic Gradient Boosting Friedman algorithm, which is a version of Boosting Algorithm (FGD) with decision trees and *bagging* with two methods commonly used in chemotherapy, the discriminant analysis method of partial least squares (PLS-DA) and Bagging. The data set of CHD (the same as used in this work) obtained by the UCI Machine Learning group was used. The error rate obtained by the stochastic gradient Boosting methods, *bagging* and PLS-DA were 14.7, 18.6 and 16.2%, respectively, showing the superiority of the Boosting algorithm.

In a study with simulated data of gene expression, Dettling and Buhlmann (2003) showed that the algorithm LogitBoost presented more accurate results when compared with the nearest neighbor and Classification Tree methods, on the orders of 12.37 and 10.21%, respectively. Furthermore, comparing the results obtained via the LogitBoost and AdaBoost algorithms on six public data sets related to types of cancer, the results of the former showed a slight improvement.

Using a data set similar to the one used in this work, but that contained 303 patients, with six samples with missing data, and then using the remaining set of 297 patients, Rodrigues and Macrini (2008) applied a neural network to this data set and obtained a hit rate of 91%. They also compared this result with the methods of discriminant analysis and the C4.5 algorithm, which presented hit rates of 87.1 and 82.3%, respectively, albeit of the same nature, the hit rate (accuracy) obtained by the Binomial Boosting algorithm was 92.59%.

Schonlau (2005) presents the implementation of the boosting software *Stata* and applies Boosting in two situations in the context of regression, one with the simulated data of a normal model and another with the simulated data of a logistic model. In the first situation the adjusted model obtained $R^2 = 21.3\%$ while applying Boosting gave $R^2 = 93.8\%$. The hit rate of the adjusted logistic model was 54.1%, and that of Boosting was 76.0%.

4. Conclusions

Considering the case studied, i.e., the classification of the presence (absence) of CHD disease, the Boosting method, more specifically the Binomial Boosting algorithm, produced a model with better fitness for the classification of the presence (absence) of CHD, as the accuracy, sensitivity, specificity, false positive rate and false negative rate of this model were better. This presents an example in expert systems where, given some variables, we need to classify with high accuracy the class in which the system is presented.

The model estimated via the Binomial Boosting Algorithm presented a more appropriate relationship with the estimated odds ratios, i.e., their value are lower when compared with the odds ratio obtained via the method of maximum likelihood.

The Binomial Boosting algorithm is therefore a powerful alternative for the analysis of situations whose response is binary in expert on intelligent systems.

In future research, we plan to study cases with multiclass classification using boosting and other classification algorithms, and we compare their results in the cases of noise and unnoisy data. This situation is particularly important when there is some noise in the collected of some of the variables used in the classification.

Acknowledgments

The authors wish to thank the Brazilian Agencies CNPq and Capes for supporting this work.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Atkinson, A. C. (1985). Plots, Transformations and Regression. *Technical Report*. Oxford: Oxford University Press.

Bartlett, M., & Traskin, P. (2007). Adaboost is consistent. *Journal of Machine Learning Resources, 8*, 2347–2368.

Buhlmann, T., & Hothorn, P. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*(4), 477–505.

Cai, Y. D. (2006). Using logitboost classifier to predict protein structural classes. *Journal of Theoretical Biology, 238*, 172–176.

Cao, D. S., et al. (2010). The boosting: A new idea of building models. *Chemometrics and Intelligent Laboratory Systems, 100*, 1–11.

Dettling, P., & Buhlmann, M. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics, 19*(9), 1061–1069.

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*, 139–157.

Emiliano, M. J. M. F. S., & Vivanco, P. C. (2014). Information criteria: How do they behave in different models? *Computational Statistics and Data Analysis, 69*, 141–153.

Emiliano, M. J. M. F. S. A. F. G., & Vivanco, P. C. (2009). Foundations and comparison of information criteria: Akaike and bayesian. *Biometric Brazilian Journal, 27*, 394–411. Eletronic ISSN 1983-0823

Frank, A., & Asuncion, A. (2010). *Machine learning repository* http://archive.ics.uci.edu/ml.

Friedman, J. (2001). Greedy function aproximation: A gradient boosting machine. *The Annals of Statistics, 29*, 1189–1232.

Friedman, T. J. T. R. J., & Hastie, J. H. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist, 28*, 337–407.

Friedman, T. J. T. R. J., & Hastie, J. H. (2001). *The Elements of Statistical Learning*. Basel: Springer Verlag.

Gey, J., & Poggi, S. (2006). Boosting and instability for regression trees. *Computational Statistics and Data Analysis, 50*, 533–560.

Hanley, J. A. (1989). Receiver operating characteristic (roc) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging, 29*, 307–335.

Hosmer, S., & Lemeshow, D. W. (1989). *Applied Logistic Regression* (2nd). John Wiley.

Kalai, A. T. S. R. A. (2005). Boosting in the presence of noise. *Journal of Computation and System Sciences, 71*, 266–290.

R core team (2011). *R: A language and environment for statistical computing*. Vienna, Austria.: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rodrigues, J. L. R. M. E. C., & Macrini, T. B. (2008). Seleção de variáveis e classificação de padrões por redes neurais como auxílio ao diagnóstico de cardiopatia isquêmica. *Pesquisa Operacional, 28*(2), 285–302.

Schapire, R. E. (1990). The stength of weak learnability. *Machine learning, 5*, 197–227.

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a stata plugin. *The Stata Journal, 5*(3), 330–354.

Schwarz, G. (1978). Estimating the dimensional of a model. *Annals of Statistics, 6*(2), 461–464.

Skruichina, R., & Duin, M. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications, 5*, 121–135.

Zhou, D. K. O. N. A., & Mcclish, X. H. (2008). *Statistical methods in diagnostic medicine*. Wiley Series in Probability and Statistics. ISBN 9780471347729.