IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 13, Issue 1 Ver. IV (Jan. - Feb. 2017), PP 104-109
www.iosrjournals.org

Monte Carlo Evaluation of Classification Algorithms Based on Fisher's Linear Function in Classification of Patients With CHD

G.R. Liska[1], G.G. Humada-Gonzalez[2] , G.J.R. Liska[3], C.J Reis[4], J. Bortolini[5]
[1](Exact Sciences Department/ Federal University of Pampa, Itaqui, RS, Brazil)
[2](Mathematic Department/ Sao Carlos University, Asuncion, Paraguay)
[3](Human Sciences Department/ Federal University of Minas Gerais, Belo Horizonte, MG, Brazil)
[4](Statistic Department/ Federal University of Lavras, Lavras, MG, Brazil)
[5](Statistic Department/ Federal University of Mato Grosso, Cuiabá, MT, Brazil)

Abstract: *Classification comprises a variety of problems, which are solved in several ways. The need for automatic classification methods arises in a number of areas, from voice recognition, to modern automobiles, to the recognition of tumors through x-rays to assist doctors, by classifying emails as legitimate or spam. Due to the importance and complexity of such problems, there is a need for methods that provide greater accuracy and interpretability of the results. Among them the Boosting methods, which have emerged in the field of computation, work by sequentially applying a classification algorithm to reweighted versions of the training data set, giving greater weight to erroneous observations. The aim of this study was to study the Fisher Linear Discriminant Analysis (LDA) model and the same one using Boosting algorithm (AdaBoost) in the presence / absence of coronary heart disease (CHD) problem in patients. The criteria used to make the comparisons were sensitivity, specificity, false positive rate and false negative rate. In addition, Monte Carlo simulation was performed to calculate these rates in different partitions of the training set. The Boosting method was successfully applied in LDA and provided a higher sensitivity than the conventional LDA.*
Keywords: *Coronary Heart Disease, Classifiers, AdaBoost, Sensibility, Machine Learning*

I. Introduction

Constructing an automatic classifier consists of using data about the problem at hand to try to create a rule that can be used to classify other data in the future [1]. The manner in which this rule is created directly influences aspects such as the performance and interpretability of the classifier. The oldest methods created in Statistics, such as Discriminant Analysis and Logistic Regression, are characterized by the creation of rules that are quite interpretable, but with restrictive forms for the relationship between the answers and the predictor variables. Some modern methods, such as neural networks, have been noted for being "black boxes" with some precision, but excessive computational cost and poor interpretability in some cases [2].

Innovative methods have emerged, both in the area of Statistics and in others, with very great predictive power. Among them, the Boosting methods, according to Friedman, Hastie and Tibshirani [3], which will be studied / applied in this work. These methods, which have emerged in the area of pattern recognition / computational vision, work by sequentially applying a classification algorithm to reweighted versions of the training data set, giving more weight to observations erroneously classified in the previous step of the algorithm. They were introduced by Shapire [4], but the most applied Boosting algorithm is AdaBoost, and since then several versions of these algorithms have been created [5].

This work aims to study the of Fisher linear discriminant method and will present its version built via the Boosting algorithm in situations where the values of the response variable is binary, that is, the problem in question admits only two possible answers, for example, the absence or presence of a particular event as a function of the (continuous) explanatory variables, or variables that are related to the existence or not of such event. In addition, a Monte Carlo simulation study will be carried out on different training sets to calculate the true and error rates of the classifiers evaluated.

II. Material and Methods

2.1 Description of the data

The Data from the present study were provided by the UCI Machine Learning Repository [6]. The data refer to 270 patients with presence or absence of coronary heart disease (CHD) and this condition is a function of 6 independent variables. These variables are listed in Table 1, as well as the nature of each variable and the possible values they can assume.

Table 1: Relation of the variables present in the problem of the diagnosis of coronary heart disease (CHD).

Variable	Type	Nature	Unit
AGE	covariate	continuous	yeras
PRESS	covariate	continuous	mm/Hg
COL	covariate	continuous	mg/dL
HEART	covariate	continuous	bpm
ST	covariate	continuous	milimeters
VES	covariate	discrete	0, 1, 2 or 3
DIS	response	binary	1: presence of CHD 2: absence of CHD

The response that is intended to be modeled is the presence / absence condition of coronary heart disease (CHD), whose representation is given by DIS. If DIS = 1 corresponds to the presence of CHD in the patient and if DIS = 2 the patient does not have CHD. This condition will be in function of some attributes (covariables), namely: the number of large color vessels by fluoroscopy (VES) (0, 1, 2 or 3); The age of the patient (AGE) in years; The blood pressure at rest (PRESS) in mm / Hg, the serum cholesterol level (COL) in mg / dL, the maximum heart rate reached (HEART) in beats per minute (bpm), and the ST segment length Electrocardiogram in millimeters (ST).

2.2 Description of the Classifiers used and Evaluation Procedure

Among the numerous algorithms used in the boosting method, the one most found for many applications is the binary classification algorithm, known as AdaBoost [7, 8]. To utilize this algorithm, a training set is considered, represented by $L = (x_1,y_1), \ldots, (x_N, y_N)$ where the classes are labeled as $C = \{1,2\}$. In this study, $\mathbf{x}_i$ represents the multivariate vector of the variables in the Table 1 for the patient *i*, *i=1,…,N=270* and y_i the classification of the patient (absence or presence for CHD), whose codification is that of set C. Through an iterative process, the AdaBoost algorithm fits a classifier at each iteration with weighted versions of the training set. At the end, the final classifier is obtained, defined by $F(\mathbf{x}) = \sum_1^M c_m f_m(\mathbf{x})$, with f_m being a classifier that returns values $\{1,2\}$; the c_m values are constant, and the corresponding prediction is the sign of $F(\mathbf{x})$, i.e., $sign(F(\mathbf{x}))$.

Thus, the AdaBoost algorithm attributes greater weight, or weighted values, to the cases that are wrongly classified; the weights are fitted in an adapted manner in each iteration; and the final classifier is a linear combination of the classifiers f_m. In this study, f_m will be the Fisher's linear discriminant analysis (LDA) [9]. Given the theoretical basis of the AdaBoost algorithm, computational implementation was given in carrying out the following steps:

1. The initial weights, $w_i = 1/N, i = 1,2,\text{K},N$, were obtained, with *N* being equal to the size of the sample.
2. Given *M*, in which *M* is the number of iterations of the algorithm, repeat procedure (a) - (c) for *i=1,..,M*.

(a) Fit the classifier $f_m(\mathbf{x}) \in \{1,2\}$ using the weights w_i in the training data. In this study, the base classifier considered is obtained through Fisher's Linear Discriminant Analysis, according to Mingoti [11].

(b) Calculate $\varepsilon_m = E_w\left[I_{(y \neq f_m(\mathbf{x}))}\right], c_m = 0.5 \times \log\left((1-\varepsilon_m)/\varepsilon_m\right)$; in which E_w corresponds to the weighted average obtained in the training set, with the weights $w = (w_1,\text{K},w_N)$.

(c) Make $w_i \leftarrow w_i \exp\left\{-c_m I_{(y \neq f_m(\mathbf{x}))}\right\}, i = 1,2,\text{K},N$ and renormalize so that $\sum_i w_i = 1$.

3. The classification of the taster *i* is given by $sign(F(\mathbf{x})) = sign\left(F(\mathbf{x}) = \sum_1^M c_m f_m(\mathbf{x})\right)$.

In each iteration, the AdaBoost algorithm leads to an increase in w_i weights of the observations wrongly classified by a factor that depends on the ε_m errors of the observations of the training set (step 2(c)). In regard to estimation of the probabilities of classification, two errors were naturally considered, in the following situations:

- Error 1, called false negative; the occurrence of this error is seen if the patient has CHD; however, the classification rule (discriminant or boosting) classifies him/her as absent of CHD.
- Error 2, called false positive; the occurrence of this error is seen if patient has not CHD and the classification rule (discriminant or boosting) classifies him/her as presence of CHD.

Thus, the probabilities of occurrence of these errors were represented by $P[Error\,1] = P[2|1]$ and $P[Error\,2] = P[1|2]$, respectively, and the lower these probabilities are, the better the discrimination function will be. That way, it became possible to evaluate the performance of the Fisher linear and boosting classifiers through counting, according to the classification given in Table 2, in which n_{kl} referred to the number of patients that belong to the group of origin *k* and which are classified by the classifier in relation to group *l* (*k, l, =1,2*). The situation in which $k = l$ results in the correct number of classifiers, and when $k \neq l$, the number of classifications is incorrect.

Table 2: Counting of false positives and negatives in reference to classification of patients in presence or absence of CHD groups.

Patient	**Classification**		Total
	(Group 1) (*Present of CHD*)	(Group 2) (*Absent of CHD*)	
Group 1 (*Present of CHD*)	n_{11}	n_{12}	n_1
Group 2 (*Absent of CHD*)	n_{21}	n_{22}	n_2

Following the frequencies n_{kl}, represented according to Table 2, the estimates of the probabilities of occurrence of errors 1 and 2 were obtained according to expressions (1) and (2), respectively, interpreted as false negative and false positive rates.

$$\hat{P}[2|1] = n_{12}/n_1 \qquad (1) \qquad\qquad \hat{P}[1|2] = n_{21}/n_2\,. \qquad (2)$$

The estimates of probabilities $\hat{P}[1|1]$ and $\hat{P}[2|2]$, according to expressions (3) and (4), are interpreted as sensitivity and specificity rates obtained by the classification rule.

$$\hat{P}[1|1] = n_{11}/n_1 \qquad (3) \qquad\qquad \hat{P}[2|2] = n_{22}/n_2\,. \qquad (4)$$

From expressions (3) and (4), the accuracy or overall success rate or accuracy *(T)* of the classifier may be obtained, given by $T = (n_{11} + n_{22})/n$, in which $n = n_1 + n_2$.

To evaluate the performance of the classifiers obtained by the two methods, the data set was separated into two parts, one part for training, directed to fitting of the LDA and boosting classifiers, and the test part, directed to validation of the classifiers. The training set will consist of 50%, 60%, 70% and 80% partition of the original sample, and the remainder of the partition will make up the testing set, which will be used for calculation of probabilities (1), (2), (3), and (4). The Monte Carlo method was used so that, in 100 simulations, the mean value and the precision of the previously cited probabilities were obtained. The results were obtained and implementation of the boosting algorithm occurred through creating functions using the R software [10]. The results related to rates (1) - (4) are described in the following section.

III. Results and Discussion

The following are the results for sensitivity, specificity, false positive rate and false negative rate. Table 3 presents these values for the Fisher Linear Discriminant Analysis (LDA) classifier. It is observed that the sensitivity of the classifier is 90.7% and the specificity is 68.42%, being considered, therefore, a classifier with reasonable discrimination power. Consequently, the false positive and false negative rates are 9.3% and 31.58%, respectively, considering a partition of 70% for the training sample. Works that have used Fisher's linear discriminant method, like in consumer credit analysis, Guimarães and Chaves-Neto [12] obtained sensitivity and specificity rates of 92.16% and 92.4%, respectively, bur ignoring the partition training set.

Table 3: Results of the classification of patients in each group using the classifier generated by the Fisher linear discriminant model according each partition *p* of the training set. Values represent probabilities (%).

Classification		**Model**							
		p=0.5		*p=0.6*		*p=0.7*		*p=0.8*	
		1	2	1	2	1	2	1	2
Observed	1	84.93	15.07	79.36	20.63	90.7	9.3	85.19	14.81
	2	37.1	62.9	28.89	71.11	31.58	68.42	25.93	74.07
Accuracy		74.81		75.92		80.25		79.63	

In the case of cross-validation following the Monte Carlo procedure, which consisted of performing 100 simulations of the training and test set partition and for each of these simulations, calculated the sensibility and specificity of the LDA classifier, the prediction power of the Fisher Linear discriminant remained consistent, in the sense that the values are close to those of Table 3 (Fig. 1). It is observed that the accuracy of the LDA classifier is high, since the obtained Monte Carlo error for the false positive and false negative rates are low, as can be observed in the error bars in the barplot Fig. 1.

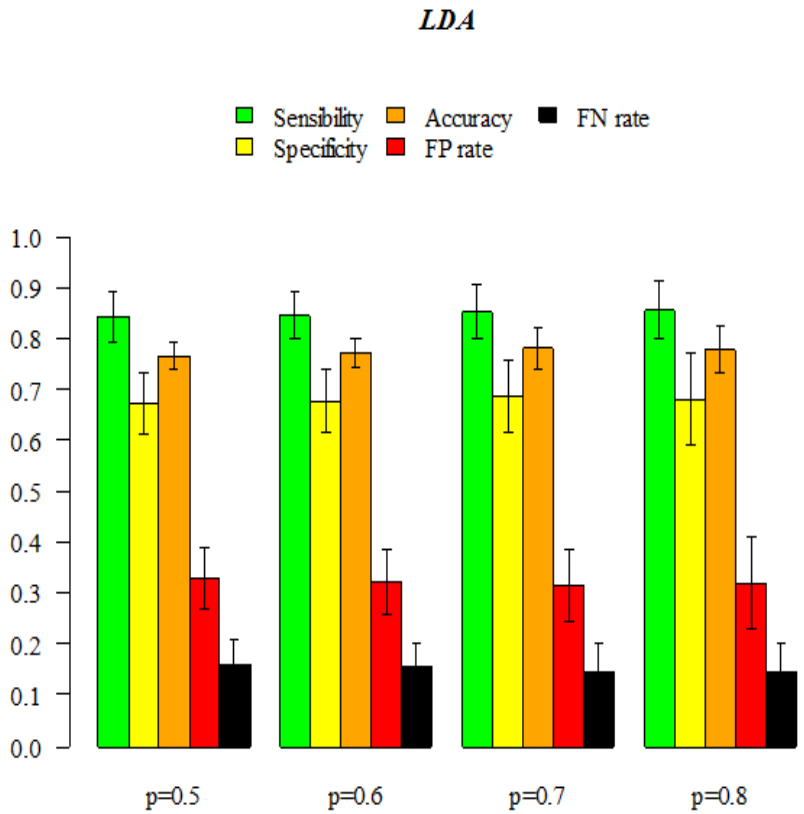

Figure 1: Barplot for the mean and Monte Carlo error of sensitivity, specificity, false positive rate, false negative rate and accuracy of the Fisher linear discriminant in different proportions (*p*) of the training set.

The following are the results for the sensitivity, specificity, false positive rate and false negative rate considering the boosting algorithm in LDA. Table 4 shows these values. It is observed that the sensitivity of the classifier is 91.84%, being considered, therefore, a classifier with very good discrimination power for the patients who actually have CHD. Consequently, the false negative rate is 8.16%. Therefore, the LDA classifier built via boosting improved the classification capacity of CHD patients across all partitions for the training set.

Table 4: Results of the classification of patients in each group using the classifier generated by the LDA boosting method in each partition *p* of the training set. Values represent probabilities (%).

Classification		**Model**							
		p=0.5		*p=0.6*		*p=0.7*		*p=0.8*	
		1	2	1	2	1	2	1	2
Observed	1	95.65	4.35	100.00	0.00	91.84	8.16	100.00	0.00
	2	66.67	33.33	78.05	21.95	56.25	43.75	76.00	24.00
Accuracy		65.18		70.37		72.84		64.81	

In the case of cross-validation following the Monte Carlo procedure, the predictive power of Fisher's linear discriminant analysis was improved when the boosting method was applied, in the sense that the values are close to those of Table 4 (Fig. 2). Similar to the Fig. 1, a practical interpretation of Fig. 2 can be made as follows: the LDA classifier via boosting presented significant improvement for the sensitivity, since its values for the Monte Carlo mean was higher than those of Fig. 1 in all partitions evaluated, consequently, the false negative rate was also lower in all analyzed partitions.

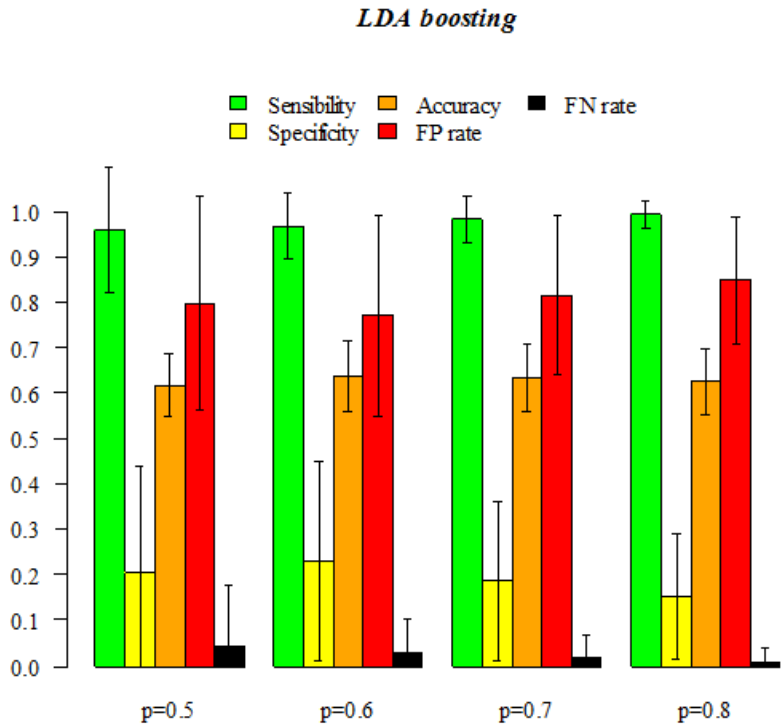

Figure 2: Barplot of the mean and Monte Carlo error of sensitivity, specificity, false positive rate, false negative rate and accuracy of the Fisher linear discriminant through boosting algorithm in different proportions (*p*) of the training set.

A high percentage of correct classification (90.1%) was also found in the study of Estévez et al. [13], who studied the differences between two groups of patients dependent on opium and a control group, considering biochemical and hematological variables of nutritional importance in the blood samples of patients. The literature lacks studies that involve the utility of the boosting method in LDA; nevertheless, Cao et al. [8] compared a boosting algorithm with another classifier well-known in machine learning, the support vector machines, and concluded that the boosting algorithm showed superior accuracy in classification of structural classes of proteins. Boosting was also successfully applied in the studies of Cai et al. [14], Shafik and Tutz [15], and Buhlmann and Hothorn [1].

IV. Conclusions

The classical Discriminant Analysis method can be used to obtain a classification rule for CHD patients, presenting a reasonable success rate. This method proved to be consistent with respect to success-and-error rates. The Boosting method was successfully applied in LDA and provided a higher success rate than the conventional LDA analysis. The Boosting algorithm in LDA, after cross-validation, presented better values of sensitivity and false negative rate in relation to the conventional LDA method. Considering the promising results of the Boosting method, as future work, we intend to evaluate the Boosting method using other types of classifiers, such as Fisher's quadratic discriminant analysis and logistic regression models, in more complex problems.

References

[1]. P. Buhlmann and T. Hothorn, Boosting Algorithms: Regularization, Prediction and Model Fitting, *Statistical Science, 22(4)*, 2007, 477-505.
[2]. B.D. Ripley, *Pattern Recognittion and Neural Networks* (Cambridge University Press, 1996)
[3]. J.H. Friedman, T.J. Hastie and R.J. Tibshirani, *The Elements of Statistical Learning* (Basel: Springer Verlag, 2001).
[4]. R.E. Schapire, The strength of weak learnability, *Machine learning, 5*, 1990, 197-227.
[5]. Y. Freund and R.E. Schapire, Experiments with a new Boosting algorithm, In: International Conference on Machine Learning, 1996, 148-156.
[6]. A. Frank and A. Asuncion, *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml, 2010.
[7]. P. Bartlett and M. Traskin, AdaBoost is consistent. *Journal of Machine Learning Research, 8(1)*, 2007, 2347-2368.
[8]. D.S. Cao, Q.S. Xu, Y.Z. Liang, L.X. Zhang and H.D. Li, The boosting: a new idea of building models. *Chemometrics and Intelligent Laboratory Systems, 100(1)*, 2010, 1-11.
[9]. M. Skurichina and R.P. Duin, Boosting in linear discriminant analysis, *Multiple classifier Systems, Springer Berlin Heidelberg, Lectures Notes in Computer Science, 1857*, 2000, 190-199.

[10]. R Core Team (2016), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, url http://www.R-project.org/.
[11]. S.A. Mingoti, *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada* (Publisher UFMG, 2005)
[12]. I.A Guimarães and A. Chaves-Neto, Reconhecimento de padrões: metodologias estatísticas em crédito ao consumidor, *RAE-Eletrônica, 1(2),* 2002, 1-14.
[13]. J.F.D.F. Estévez, F.D.F Estévez, C.H. Calzadilla, E.M.R. Rodríquez, C.D.Romero and, L. Serra Majem, Application of linear discriminant analysis to the biochemical and haematological differentiation of opiate addicts from healthy subjects: a case-control study. *European Journal of Clinical Nutrition, London, 58(3), 2004,* 449-455.
[14]. Y.D. Cai, K.Y. Feng, W.C. Lu and K.C. Chou, Using logit boost classifier to predict protein structural classes. *Journal of Theoretical Biology, 238(1),* 2006, 172-176.
[15]. N. Shafik and G. Tutz, Boosting nonlinear additive autoregressive time series. *Computational Statistics & Data Analysis, 53(7),* 2009, 2453-2464.