Regression Discriminant Analysis (RDA) Variants

Jude C Obi

Department of Statistics,
COOU,[*]
Nigeria
obi@jc.com

August, 2017

Abstract

This study is a follow-up to earlier publications on the relationship of least squares regression to FDA. In particular, we focus on the paper; On the Regression Discriminant Analysis (RDA), and its Identical Relationship to the Fisher's Discriminant Analysis, and argue that since $\hat{\boldsymbol{\beta}}$ is a vector of coefficients for the least squares regression, given that $y \in (+1, -1)$, then a substitution of $\hat{\boldsymbol{\beta}}$ with either $\hat{\boldsymbol{\beta}}_{ridge}$ or $\hat{\boldsymbol{\beta}}_{lasso}$ gives a regression discriminant analysis variant. We note that both $\hat{\boldsymbol{\beta}}_{ridge}$ and $\hat{\boldsymbol{\beta}}_{lasso}$ are respectively vectors of coefficients for ridge regression and lasso (least absolute shrinkage and selection operator), given that $y \in (+1, -1)$. We therefore identify Ridge Regression Discriminant Analysis (RRDA) and Lasso Discriminant Analysis (LaDA) as the two regression discriminant analysis variants. Further empirical investigation follows, mainly to show that RRDA and LaDA compete favourably against a known Fisher's Discriminant Analysis (FDA) variant namely, Regularized Fisher's Discriminant Analysis (RFDA). Since RRDA, LaDA or RFDA can be used in place of FDA in high dimensions, we further determine the most suitable replacement for FDA assuming we are in high dimensions.

Index Terms: *Machine learning, Regression based binary classification, Linear discriminant analysis, Least squares discriminant analysis variants.*

I. Introduction

The use of regression in classification has been in the form of logistic regression [4]. The logistic regression fits a non linear model to a linear combination of explanatory variables, and it is superior to FDA when normality assumptions are violated [13]. If normality is assumed, it is an alternative to FDA [9]. However, our interest is in a regression based classification procedure that fits a linear model for classification based on the multiple regression.

To this end, the work of [17, 5, 12] are of interest to us because they involve fitting a linear model for classification based on the multiple regression. In particular, we focus on the work of [12].

The authors proved that the least squares vector of coefficients $\hat{\boldsymbol{\beta}}$ is proportional to γ, where γ is the weight vector of FDA's classification function given that $y \in (+1, -1)$. A detailed review of their work is as follows:

a. Data and some notations

Let $X_1\,(n_1 \times p)$ and $X_2\,(n_2 \times p)$ be datasets for two populations Π_1 and Π_2, and let $n = n_1 + n_2$. Let

$$X = \left[\begin{array}{c} X_1 \\ X_2 \end{array}\right] (n \times p) \tag{1}$$

denote the whole dataset, and $H = I_n - (1/n)\,1_n 1_n^T$ denote the $n \times n$ centring matrix. In a similar way, let H_1 and H_2 denote the $n_1 \times n_1$, and $n_2 \times n_2$ centring matrices respectively.

Let $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\bar{\mathbf{x}}$ denote the sample means of X_1, X_2 and X respectively. Note that

$$\bar{\mathbf{x}} = (n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2)\,/n$$

[*] Chukwuemeka Odumegwu Ojukwu University

is a weighted average of the two class means. We also need the unweighted average

$$\mathbf{x}_{\text{av}} = \left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)/2,$$

and the difference,

$$\delta = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2. \tag{2}$$

b. Fisher's allocation rule

Several matrices are of interest in discriminant analysis:

$$\begin{aligned} T &= X^T H X, \\ B &= \left(n_1 n_2/n\right)\delta\delta^T, \\ W &= X_1^T H_1 X_1 + X_2^T H_2 X_2. \end{aligned}$$

A classic result [8] states that

$$T = W + B.$$

The Fisher's allocation rule is based on Fisher's linear discriminant function given by:

$$f(\mathbf{x}) = \delta^T W^{-1}\left(\mathbf{x} - \mathbf{x}_{\text{av}}\right).$$

The allocation rule in respect of a new input $\mathbf{x}$ says: allocate $\mathbf{x}$ to Π_1 if $f(\mathbf{x}) \geq 0$, and to Π_2 otherwise.

It is important to note that sometimes $f(\mathbf{x})$ is constructed using $S_{\text{pooled}} = W/(n-2)$ instead of W, but the allocation rule is the same. Since W is symmetrical, write

$$\gamma = W^{-1}\delta; \tag{3}$$

then Fisher's discriminant function simplifies to

$$f(\mathbf{x}) = \gamma^T\left(\mathbf{x} - \mathbf{x}_{\text{av}}\right).$$

c. Multiple Regression

Let $\mathbf{y} = \begin{pmatrix} +\mathbf{1}_{n_1\times 1} \\ -\mathbf{1}_{n_2\times 1} \end{pmatrix}$ denote a response vector of length n, and consider a regression of $\mathbf{y}$ on X. Then, the ordinary least squares regression function can be written as

$$g(\mathbf{x}) = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\mathbf{x},$$

where $\hat{\alpha} = \bar{y} - \hat{\boldsymbol{\beta}}^T\bar{\mathbf{x}}$, and $\hat{\boldsymbol{\beta}} = \left(X^T H X\right)^{-1} X^T H\mathbf{y} = T^{-1}X^T\left(H\mathbf{y}\right)$.

Note that $\hat{\boldsymbol{\beta}}$ is estimated using the centred data matrix HX, we then claim that

$$\hat{\boldsymbol{\beta}} \propto \gamma, \tag{4}$$

where γ is as defined in (3).

d. Proof

First note that the centred vector $H\mathbf{y}$ has entries $+1-\bar{y}$ in the first n_1 places and $-1-\bar{y}$ in the final n_2 places. Since $\bar{y} = \left(n_1 - n_2\right)/n$, $H\mathbf{y}$ simplifies to $2n_1n_2/n$ times a vector with $+1/n_1$ in the first n_1 places and $-1/n_2$ in the final n_2 places.

Hence,

$$\begin{aligned} X^T(H\mathbf{y}) &= \left(1/n_1\right)X_1^T\mathbf{1}_{n_1} - \left(1/n_2\right)X_2^T\mathbf{1}_{n_2} \\ &= \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \delta, \end{aligned} \tag{5}$$

where δ is as defined in (2).

Showing that $\hat{\boldsymbol{\beta}} \propto \gamma$ is equivalent to showing that $T^{-1}\delta \propto \gamma$, which is true if and only if

$$\begin{aligned} \delta &\propto T\gamma \\ &\propto TW^{-1}\delta \\ &\propto \left(W+B\right)W^{-1}\delta \\ &\propto \left(I + \left(n_1n_2/n\right)\delta\delta^TW^{-1}\right)\delta \\ &\propto \delta + \left(n_1n_2/n\right)\delta\left(\delta^TW^{-1}\delta\right) \\ &= \left\{1 + \left(n_1n_2/n\right)\left(\delta^TW^{-1}\delta\right)\right\}\delta \\ &= u\delta, \end{aligned}$$

where $u = \left\{1 + \left(n_1n_2/n\right)\left(\delta^TW^{-1}\delta\right)\right\}$ is a constant. Hence, the result is proved.

e. Regression rule

Set,

$$\begin{aligned} g(\mathbf{x}) &= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\mathbf{x} \\ &= \bar{y} - \hat{\boldsymbol{\beta}}^T\bar{\mathbf{x}} + \hat{\boldsymbol{\beta}}^T\mathbf{x} \\ &= \bar{y} + \hat{\boldsymbol{\beta}}^T\left(\mathbf{x} - \bar{\mathbf{x}}\right), \end{aligned} \tag{6}$$

and allocate to Π_1 if $g(\mathbf{x}) \geq 0$, otherwise to Π_2. If on the other hand we set $\mathbf{x} = \mathbf{x}_{\text{av}}$, then,

$$g\left(\mathbf{x}_{\text{av}}\right) = \bar{y} + \hat{\boldsymbol{\beta}}^T\left(\mathbf{x}_{\text{av}} - \bar{\mathbf{x}}\right) \neq 0,$$

unless $n_1 = n_2$. Hence, the naive regression is different from Fisher's rule. We have used the term naive regression to explain that the function g, specified in (6), is identical to FDA if and only if $n_1 = n_2$.

f. Alternative rule

Alternatively, we can shift the regression predictor by a constant value to

$$\begin{aligned} g^*\left(\mathbf{x}\right) &= g\left(\mathbf{x}\right) - \left(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\mathbf{x}_{\text{av}}\right) \\ &= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\mathbf{x} - \hat{\alpha} - \hat{\boldsymbol{\beta}}^T\mathbf{x}_{\text{av}} \\ &= \hat{\boldsymbol{\beta}}^T\left(\mathbf{x} - \mathbf{x}_{\text{av}}\right), \end{aligned} \tag{7}$$

and define another rule: allocate $\mathbf{x}$ to Π_1 if $g^*\left(\mathbf{x}\right) \geq 0$ and to Π_2 otherwise.

The authors noted that the allocation rule given by f and g^* are identical, hence they called g^* a regression based discriminant function instead of g.

We now argue that since $\hat{\boldsymbol{\beta}}$ in (7) is based on the least squares given that $y \in (+1, -1)$, a substitution thereof with either $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ or $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ under the same condition, gives rise to a classification function called RRDA or LaDA.

In the section that will follow, we shall carry out empirical investigation to show that both RRDA and LaDA competes favourably against RFDA.

II. Empirical Investigation

We shall analyse the error rates of RRDA, LaDA and RFDA on different datasets. Since all the classifiers can be used in place of FDA when $p >> n$, we would like to know which of them, if any, significantly differs in its error rates given the different datasets. Here, p refers to the number of explanatory variables, and n is the number of training instances. As p becomes larger than n, the investigation will also determine which classifier is a most suitable replacement for FDA. In other words, we shall determine a classifier that will be recommended in place of FDA in high dimensions.

A good number of the datasets used in the investigation will be sourced from the UCI Machine Learning Repository [10], and KEEL dataset repository [1]. We shall pre-process all the datasets to ensure that each class label is identified with the name “class", and consists of a vector of $+1$ and -1 discrete variables. This way, the problem of rewriting the program we use each time a different dataset is involved is avoided. The datasets include:

Appendicitis

The data represents 7 medical measures taken over 106 patients on which the class label represents whether the patient has appendicitis (class label $+1$) or not (class label -1). We have a total of 21 samples in class $+1$ whereas class -1 consists of 85 samples. The dataset was sourced from KEEL dataset repository.

Australia Dataset

The Australia dataset concerns credit card applications, and all attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. It has dimensions 690 × 14, with two classes representing approved and not approved. The data source is [1], and website; http://sci2s.ugr.es/keel/dataset.php? cod=53.

CoIL 2000

The dataset was used in CoIL 2000 challenge, and contains information on customers of an insurance company. It is a binary classification dataset, and consists of 85 variables including product usage data, and socio-demographic data. The number of samples involved is 9822, with a total of 9236 in class $+1$ and 586 in class -1. It was sourced from the UCI Machine Learning Repository.

Colon

Colon is a gene expression dataset from the microarray experiments of colon tissue samples

[2]. The dataset consists of 62 samples and 2000 genes (features). It has two classes, namely tumour tissue with 40 samples, and normal tissue with 22 samples. It is contained in the plsgenomics package in R. The names of the genes were not given and we represented them conveniently.

Gisette
The dataset is one of five datasets used in the NIPS 2003 feature selection challenge, and it was put together by [7]. The sample size is 7000, with 5000 features and each of the two classes has 3500 samples. The dataset is also contained in the UCI Machine Learning repository.

Handheight
The Handheight dataset is two dimensional, and consists of heights and stretched hand span of 167 male and female college students. Each student decided which of their hands to measure. Class $+1$ has 89 samples whereas class -1 consists of 78 samples. The source of the data is [16].

Heart
This is a real world binary classification heart disease dataset, and the task is to detect the absence (-1) or presence (1) of heart disease. It contains 270 samples and 13 features, with 120 samples in class $+1$ and 150 samples in class -1. The data was sourced from the UCI Machine Learning Repository.

Heberman
This dataset contains cases from a study that was conducted between 1958 and 1970, at the University of Chicago's Billings Hospital, on the survival of patients who had undergone surgery for breast cancer. The task is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 year (negative). The sample size is 306 with 3 features, and class $+1$ has 225 samples whereas class -1 contains 81 samples. The dataset was sourced from the KEEL dataset repository.

Hepatitis
Hepatitis is a real world dataset; it contains a mixture of integer and real valued attributes, with information about patients affected by the hepatitis disease. It consists of 80 samples and 19 features. Class $+1$ has 67 samples whereas class -1 has 13 samples, and the task is to predict if these patients will die (-1) or survive (1). It was sourced from the UCI Machine Learning Repository.

Hill valley with noise (HVWN)
The hill valley with noise dataset consists of 606 instances, and 100 features for both training and test sets. Noise contamination of the dataset is retained, thereby differentiating it from the hill valley without noise dataset. The data was sourced from the UCI Machine Learning Repository.

Ionosphere
Ionosphere is a radar dataset collected by a system in Goose Bay, Labrador. The system consists of a phased array of 16 high-frequency antennas with a total transmitted power of the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. “Good" radar returns are those showing evidence of some type of structure in the ionosphere. “Bad" returns are those that do not; their signals pass through the ionosphere.

Leukemia
The leukemia dataset is a gene expression data consisting of 3051 genes, with 38 tumour MRNA samples from the leukemia microarray study [6]. The tumour MRNA samples are of two cancerous classes, here denoted as -1 and $+1$. Since the number of training samples is small, we used the same dataset for training to also test the classifiers. Our interest here is merely in the performances of the classifiers given such scenarios. The dataset is contained in R package plsgenomics. Although the gene names were given, we have represented them conveniently because they are lengthy, and we have no intrinsic interest in the gene names.

Magic
This dataset was used to simulate registration of high energy gamma particles, in a ground-based atmospheric Cherenkov gamma telescope, using the imaging technique. The

dataset was generated by a Monte Carlo program [3], and the task is to discriminate statistically images generated by primary gammas, from the images of hadronic showers initiated by cosmic rays in the upper atmosphere. It contains 19020 samples and 10 features; the source is the UCI Machine Learning Repository.

Mammographic
This dataset was used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field, which is the target attribute). The dataset was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. It has dimensions 830×5, and the source is the KEEL dataset repository.

Parkinsons
The Parkinsons dataset is of dimension 195×23, and involves a range of biomedical voice measurements of some people with and without Parkinson's disease (PD). It was sourced from the UCI Machine Learning Repository. Documentation on the dataset shows that each column is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals.

Prostate
The prostate dataset is a gene expression dataset [15]. The dataset is contained in R package spls and consists of two classes, namely 52 prostate tumour and 50 normal classes. The number of genes involved is 6033. The names of the genes were not given and as a result, we represented them conveniently.

Ringnorm
Ringnorm is a 20 dimensional, 2 class classification dataset. Each class is drawn from a multivariate normal distribution, and class 1 has mean 0 and covariance 4 times the identity. Class 2 has mean $(a, a, \cdots, a)$ and unit covariance ($a = 2/\sqrt{20}$). The number of instances is 7400, and like most simulated datasets, the dataset is useful for testing performances of binary classifiers. The source is the KEEL dataset repository.

Saheart
The Saheart dataset pertains to a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. The saheart data were taken from a larger dataset described in [14]. The class label indicates whether the person has a coronary heart disease: negative (-1) or positive ($+1$). The dataset has dimensions 462×9, and is contained in the ElemStatLearn package in R.

Sonar
This dataset contains how many signals obtained from a variety of different aspect angles, spanning 90 degrees for mines and 180 degrees for rocks. Each pattern is a set of 60 numbers in the range 0.0 to 1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time. The output attribute contains the letter $+1$ if the object is a rock and -1 if it is a mine (metal cylinder). The source is the UCI Machine Learning Repository.

Spambase
The spambase dataset contains information about 4597 e-mail messages. The task is to determine whether a given email is spam (class $+1$) or not (class -1), depending on its contents (4 duplicated instances have been removed from the original data set). Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail. The dataset was sourced from the UCI Machine Learning Repository.

SPECTF Heart
The SPECTF Heart dataset is of dimension 267×44, and consists of diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient involved

in the study is classified into one of two categories: normal and abnormal. Altogether, 44 continuous feature patterns were created for each patient. The source of the dataset is UCI Machine Learning Repository.

Twonorm
This dataset is 20 dimensional, and consists of 2 classes. Each class is drawn from a multivariate normal distribution. Class $+1$ has mean $(a, a, ..a)$ while Class -1 has mean $(-a, -a, .. - a)..a = 2/sqrt(20)$. The dataset has dimensions 7400 × 20, and is contained in the KEEL dataset repository.

Wisconsin Diagnostic Breast Cancer (WDBC) Dataset
WDBC is a real world dataset, and contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The number of instances is 569 and the task is to determine if a tumour found is benign or malignant (-1 = malignant, and 1 = benign). It was sourced from the UCI Machine Learning Repository.

i. Discussion

Table 1 consists of the error rates of the three classifiers and the system time it took each classifier to finish execution. The LaDA variables refers to the number of variables used by LaDA to compute the error rate given each dataset. On examination of the Table, it appears that differences in the error rates of the classifiers on most datasets are not significant. Exceptions may include six datasets, namely the Colon, Hepatitis, HVWN, HVWON, SPECT Heart and Sona. The error rates on these datasets are comparatively higher. It seems that the RRDA error rates are marginally smaller on Hepatitis, HVWN, SPECTF Heart and Sona datasets. LaDA seems to have marginally smaller error rates on Colon dataset, whereas RFDA recorded a marginally smaller error rates on HVWON dataset. Altogether, it appears that RRDA performed relatively better on the six datasets, but it can still be argued that the different error rates between RRDA and others are not significant.

The box plots of Figure 1(a) show that LaDA has a marginally smaller median error rate, in comparison with RRDA and RFDA. The median error rates of RRDA and RFDA appear to be similar. However, based on the size of the LaDA box, it seems that we have more variations in error rates of LaDA than in the error rates of RRDA and RFDA.

On the other hand, differences in the time it took each classifier to finish execution may be seen as insignificant with some datasets, and very significant with other datasets. For instance, with datasets Australia, Handheight, Heart, Mammographic, Parkinsons, Saheart, Sonar, Twonorm and WDBC, differences in system time appear insignificant. The same may not be true with datasets like Gisette, Prostate, and possibly, Colon and Leukaemia. It seems that in the instances where $n > p$, the system time is relatively smaller vis-a-vis when $p > n$. An exception here is the dataset Gisette, because despite the fact that $n > p$ the system time for the classifier is still very high. The RFDA particularly recorded the highest system time on datasets Gisette and Prostate. They constitute the two most prominent outliers in Figure 1(b). Comparatively, it seems we have more outliers with the system time for RFDA. The implication is that RFDA may not be a preferred classifier in high dimensions.

Test for Normality, Homogeneity of Variances, and the Repeated Measures ANOVA

A Shapiro Wilks normality test on the error rates of LaDA, RRDA and RFDA failed to reject compliance with the normality assumption at p-values of 0.4059, 0.383 and 0.9509 respectively. Similarly, a Bartlett's test also failed to reject compliance with the homogeneity of variances at a p-value of 0.9590. Hence, we confirm that we have complied with both normality and homogeneity of variances assumptions.

Name Dataset	Dimensions	LaDA	LaDA Variables	Sys. Time in Sec.	RRDA	Sys. Time in Sec.	RFDA	Sys. Time in Sec.
Appendicitis	106 × 7	0.125	5	1.96	0.125	1.89	0.1563	5.71
Australia	689 × 14	0.1353	8	1.98	0.1353	2.19	0.1159	1.9
Coil2000	9822 × 85	0.2677	50	18.82	0.2691	21.6	0.2959	15.65
Colon	62 × 2000	0.1579	18	3.96	0.2105	9.01	0.2105	49.66
Gisette	7000 × 5000	0.028	1407	1495.2	0.023	2248.8	0.091	8958.24
Handheight	167 × 2	0.18	2	1.84	0.18	1.85	0.18	1.47
Heart	270 × 13	0.1489	11	3.37	0.1489	3.01	0.1596	1.67
Heberman	306 × 3	0.2637	2	1.82	0.2418	1.95	0.2418	1.56
Hepatitis	80 × 20	0.1944	3	8.09	0.1389	2.07	0.2222	1.65
HVWON	1212 × 100	0.3069	36	16.54	0.33	6.62	0.2822	6.19
HVWN	1212 × 100	0.3465	30	16.96	0.3152	7.18	0.396	7.09
Ionosphere	350 × 32	0.1619	20	2.28	0.1714	2.28	0.1905	11.75
Leukemia	76 × 3051	0	37	3.85	0	14.03	0	180.6
Magic	19020 × 10	0.2059	9	9.31	0.2075	10.72	0.205	8.91
Mammographic	830 × 5	0.2289	2	1.79	0.2088	1.98	0.1968	1.46
Parkinsons	195 × 23	0.1207	20	3.36	0.1897	2.93	0.1897	1.76
Prostate	102 × 6033	0.0968	68	10.92	0.0968	32.42	0.0968	6496.2
Ringnorm	7400 × 20	0.2477	20	6.36	0.2482	5.2	0.2477	4.1
Saheart	462 × 9	0.3525	7	1.74	0.3381	2.04	0.3597	1.87
Simulated dataset	2000 × 40	0.0425	38	7.76	0.0363	4.13	0.04	3.39
Sonar	208 × 60	0.2742	17	2.78	0.1935	2.58	0.2258	1.96
Spambase	4597 × 57	0.1159	52	7.85	0.1145	7.8	0.1319	5.59
SPECTF Heart	267 × 44	0.3209	22	2.44	0.2834	4.04	0.3369	2
Twonorm	7400 × 20	0.0216	20	5.64	0.0216	5.59	0.0216	4.22
WDBC	569 × 30	0.0175	25	2.14	0.0175	2.09	0.0702	1.96

Table 1: *Error rates of RFDA, LaDA and RRDA, and the system time it took each classifier to finish execution given different datasets.*

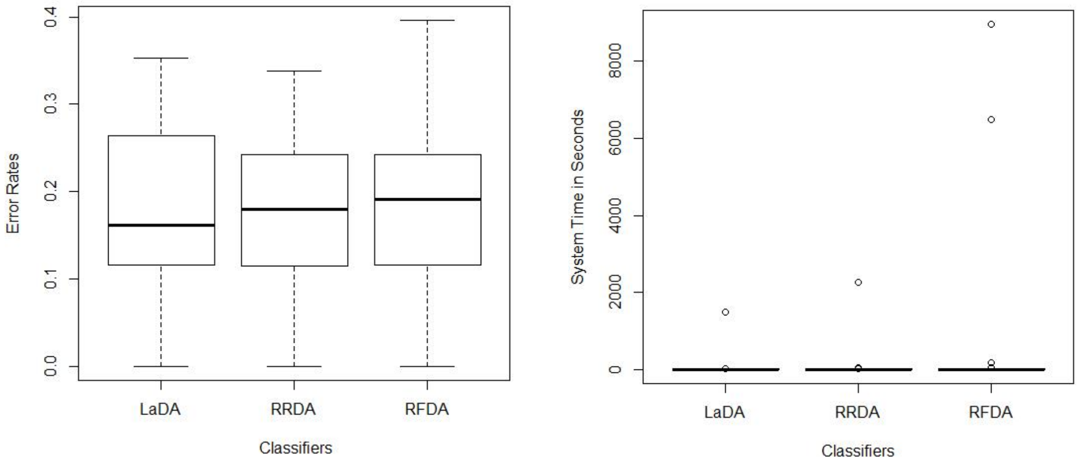
We therefore carried out repeated measures ANOVA test, which rejected the hypothesis that differences in the error rates of the classifiers are non significant at a p-value of 0.0233. We further considered a post hoc test based on paired t-test, and obtained the following output in R:

```
Pairwise comparisons using paired t tests
data:  Values and Classifiers

      LaDA  RFDA
RFDA 0.057   -
RRDA 0.437 0.013

P value adjustment method: none
```

The result shows that differences in the error rates between LaDA and RRDA are non significant at a p-value of 0.437. We narrowly rejected the hypothesis that differences in error rates between RFDA and LaDA are significant at a p-value of 0.057. Lastly, we failed to reject the hypothesis that differences in error rates between RFDA and RRDA are significant at a p-value of 0.013. By considering the p-value at which we rejected the existence of significant differences between the error rates of RFDA



(a) Box plots of the error rates of LaDA, RRDA and RFDA

(b) Box plots of the system time for LaDA, RRDA and RFDA

Figure 1: *Box plots in respect of the error rates and system time for LaDA, RRDA and RFDA.*

and LaDA, one may be cautious to assume that both classifiers are as good as the other. If we further consider the number of variables LaDA used, we are of the view that LaDA performed better than RFDA. In all, based on the error rates, we argue that RRDA relatively performed better, followed by LaDA and lastly, RFDA.

Assuming we considered a paired t-test with a Bonferroni adjustment, we would have obtained the following result:

```
Pairwise comparisons using paired t tests
data:  Values and Classifiers

     LaDA RFDA
RFDA 0.17  -
RRDA 1.00 0.04

P value adjustment method: bonferroni
```

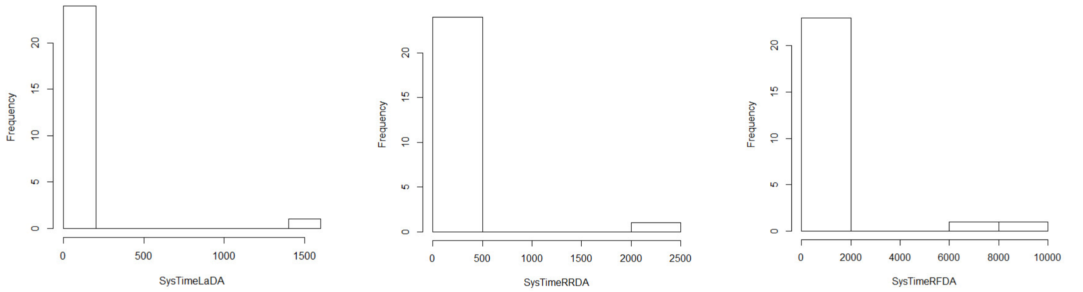
The Bonferoni adjustment gives a result similar to the paired t-test without adjustment. For this reason, our interpretation of the performances of the classifiers remains unaltered.

Regarding the system time, the histograms of Figure 2 strongly suggest that the system time in respect of each classifier does not follow a normal distribution. It is clearly caused by the heavy presence of outliers. Confirming our position, the Shapiro Wilks normality test rejected compliance with the normality assumption in respect of system time of LaDA, RRDA and RFDA at p-values of 1.48*e*-10, 1.429*e*-10 and 1.012*e*-09 respectively. Also, the Bartlett's test on homogeneity of variances rejected compliance with the homogeneity of variance assumption at a p-value of 2.2*e*-16. For these reasons, we shall use a non parametric Friedman's test for comparison of the system time of the three classifiers.

Consequently, the output of the Friedman's test in R gives:

```
Friedman rank sum test
data:  datx
Friedman chi-squared = 8.7475, df = 2,
p-value = 0.0126
```

With a p-value of 0.0126, at 5% level of significance, we reject the null hypothesis of no difference in the system time of the three classifiers. This means that the observed differences are significant, hence, a post hoc analysis for the Friedman's test will follow. In this regard, we used the Nemenyi post hoc test [11] and obtained the following output in R, at a 5% level of significance:



(a) Histogram for LaDA on 25 datasets.

(b) Histogram for RRDA on 25 datasets.

(c) Histogram for RFDA on 25 datasets.

Figure 2: *Histograms in respect of system time for RFDA, LaDA and RRDA on* 25 *datasets.*

```
Pairwise comparisons using Nemenyi
multiple comparison test
data:  datx
             LaDA.SysTime   RRDA.SysTime
RRDA.SysTime       0.989        -
RFDA.SysTime        0.036      0.024
```

The Nemenyi test shows that we have significant differences in system time between LaDA and RFDA, and between RRDA and RFDA. Differences in the system time between LaDA and RRDA are non significant at a p-value of 0.989.

Because we suspected that datasets like Colon, Gisette, Leukaemia and Prostate may be responsible for significant differences in the system time between RFDA and others, we removed them and carried out the Friedman test again. At a p-value of 8.598e-05, at the same level of significance, we equally rejected the null hypothesis of no difference in the system time between RFDA and other two classifiers.

A post-hoc Nemenyi test similarly rejected the null hypothesis of no difference in system time between RFDA and LaDA, and between RFDA and RRDA at p-values of 0.00025 and 0.00150 respectively. Also at a p-value of 0.88862, we failed to reject differences in the system time between LaDA and RRDA.

We equally observed that the dimensions of the datasets, where RFDA recorded increased number of system time are relatively higher. It suggests that increase in the dimensions of a dataset, would mean more system time for RFDA.

On account of the foregoing, we state that based on the system time, either of LaDA or RRDA can be preferred to the use of RFDA. In high dimensions, we recommend the use of LaDA as a preferred classifier in place of FDA, primarily for its additional feature as a variable selector.

III. Conclusions

RRDA and LaDA are valid binary classifiers based on the outcome of the tests carried out in section II. Both classifiers can be used in place of FDA when $p > n$. If $p >> n$, they are still preferred alternative to FDA in comparison with RFDA. On the other hand, RFDA has a weakness of using more computational time vis-a-vis RRDA and LaDA, hence in high dimensions, the use of RRDA or LaDA may be preferred to RFDA.

On the strength of the number of variables used by LaDA to compute its error rates, we infer that if the objective of a binary classification problem is to use a few important variables, LaDA may be a choiced classifier.

References

[1] J Alcalá et al. "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework". *Journal of Multiple-Valued Logic and Soft Computing* 17.2-3 (2010), pp. 255–287.

[2] Uri Alon et al. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays". *Proceedings of the National Academy of Sciences* 96.12 (1999), pp. 6745–6750.

[3] Thomas Bretz et al. "The drive system of the major atmospheric gamma-ray imaging Cherenkov telescope". *Astroparticle Physics* 31.2 (2009), pp. 92–101.

[4] David R Cox. "The Regression Analysis of Binary Sequences". *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), pp. 215–242.

[5] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, 2012.

[6] Todd Golub et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". *Science* 286.5439 (1999), pp. 531–537.

[7] Isabelle Guyon. "Design of Experiments of the Nips 2003 Variable Selection Benchmark". *NIPS 2003 workshop on feature extraction and feature selection*. 2003.

[8] Paul J Hewson. "Multivariate Statistics with R" (2009).

[9] Gareth James et al. *An Introduction to Statistical Learning*. Vol. 6. Springer, 2013.

[10] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: `http://archive.ics.uci.edu/ml`.

[11] Peter Nemenyi. "Distribution-free Multiple Comparisons". *Biometrics*. Vol. 18. 2. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210. 1962, p. 263.

[12] Jude C Obi, Peter Thwaites, and John Kent. "On the Regression Discriminant Analysis (RDA), and its Identical Relationship to the Fisher's Discriminant Analysis". *International Journal of Scientific and Research Publications (IJSRP)* Volume 7 (Issue 7 2017).

[13] Maja Pohar, Mateja Blas, and Sandra Turk. "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study". *Metodoloski zvezki* 1.1 (2004), p. 143.

[14] J Rousseauw et al. "Coronary Risk Factor Screening in three Rural Communities". *South African Medical Journal* 64.430-436 (1983), p. 216.

[15] Dinesh Singh et al. "Gene Expression Correlates of Clinical Prostate Cancer Behaviour". *Cancer cell* 1.2 (2002), pp. 203–209.

[16] Jessica Utts and Robert F Heckard. *Mind on Statistics*. Cengage Learning, 2011.

[17] Jieping Ye. "Least Squares Linear Discriminant Analysis". *Proceedings of the 24th International Conference on Machine Learning*. ACM. 2007, pp. 1087–1093.