International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 4, Issue 1, 2013, pp124-130
http://bipublication.com

A PERSPECTIVE STUDY ON THE ROBUSTNESS OF STATISTICAL CLASSIFICATION AND DISCRIMINATION METHODS

R. Vishnu Vardhan and Sameera G

Dept of Statistics, Pondicherry University,
Puducherry – 605014. rvvcrr@gmail.com, samskutti@gmail.com

[Received-25/08/2012, Accepted-15/01/2013]

ABSTRACT

In recent years, many researchers have paid serious attention to the technique namely Receiver Operating Characteristic (ROC) curve, is an important tool for classification in medical diagnosis. The extent of correct classification is obtained using an intrinsic measure of ROC curve namely Area under the curve (AUC). Another intrinsic measure of ROC curve that measures the distance of the curve from chance diagonal is Maximum Vertical Distance (MVD). In this paper, we have made an attempt to develop some algorithms for the computation of the above said measures. Further, various combinations of cofactors were considered and Linear Discriminant Analysis was carried out. The sensitiveness of these models was explored by incorporating and eliminating the outliers in the data sets. One more additional feature of this work is that a Discriminant score cutoff was proposed, which will initiate to classify the new individual into one of the two groups, instead of the clinical cutoff.

Keywords: ROC curve, AUC, MVD, Outliers, Discriminant Analysis, Discriminant Score cutoff

INTRODUCTION

Discrimination and Classification are two multivariate techniques that deal with separating distinct set of objects and allocating new objects to predefined groups. Discrimination is exploratory in nature. It is often used to investigate the observed differences when casual relationships are not well understood. Classification is less exploratory in nature in the sense that it leads to well defined rules that help in assigning new objects. Classification problem arises when a researcher proposes new therapies or diagnostic tools. The ability of the tool to discriminate between the various groups in the study is often studied with the help of tools like Logistic Regression, Discriminant Analysis etc., Another interesting tool used in medical research is Receiver Operating Characteristic curve analysis. With the advent of software, ROC analysis is widely used in Medicine and radiology apart from non-medical applications as Signal Detection Theory.

ROC curve and its Measures

The ROC curve has its mathematical formulation which helps in fitting and estimating the parameters of the curve. The entire classification is carried out on the basis of a threshold value often referred to as *Gold Standard*, it determines the true condition status and also provides a source of information completely different from

the tests under evaluation. The true condition status indicates the presence of disease otherwise absence. Two basic measures of ROC curves are sensitivity (S_n) and specificity (S_p).

Sensitivity refers to the ability of a test to detect the condition when it is present and Specificity refers to the ability of test to exclude the condition in patients without the condition.

An ROC curve is a plot of 1-S_p versus S_n [7]. The construction of ROC curve primarily depends on the four possible states which are obtained on the basis of a threshold value i.e., TP, TN, FN and FP. The resulting curve is called empirical ROC [7]. In probability notation these four possibilities are given by

Status	Event
TP	[S > c / D]
FP	[S > c / $\bar{D}$]
TN	[S < c / $\bar{D}$]
FN	[S < c / D]

Here S is the test value of a diagnostic test.

Conventionally, it is assumed that diseased (X) and the healthy (Y) individuals follow Normal distribution and hence the name *Binormal ROC curve*, with unknown monotonic transformation [3]. Another important measure in ROC curve analysis is the Area Under the Curve (AUC), it is referred as the best index of discrimination (Hanley and Mc Neil, 1982, 1983). Basing on the AUC value the performance of a diagnostic test can be interpreted and it takes its range from 0.5 to 1.0. Higher the AUC values, the better will be the performance of a diagnostic test. If, for a test AUC<0.5, one should not consider that test for further classification.

Parametric form of ROC curve

Let X~N (μ_x, σ_x^2) and Y~N (μ_y, σ_y^2), where X represents data from patients with disease and Y represents data from those without disease. We can define false-positive and true-positive rates as

$$FP(c) = 1\text{-}P(S>c) = 1\text{-}\Phi\left[\frac{c-\mu_y}{\sigma_y}\right] = \Phi\left[\frac{\mu_y-c}{\sigma_y}\right] \quad (1)$$

and

$$TP(c) = P(S>c) = \Phi\left[\frac{(\mu_x-c)}{\sigma_x}\right] \quad (2)$$

where c is the cutoff decision point and Φ is the cumulative normal distribution function.

The ROC curve can be traced out by the functions FP and TP as (Mc Clish, 1989)

$$[FP(c), TP(c)] = \left\{\Phi\left[\frac{\mu_y-c}{\sigma_y}\right], \Phi\left[\frac{\mu_x-c}{\sigma_x}\right]\right\} \quad -\infty<c<\infty \quad (3)$$

The expression of area under the ROC curve is given as (Farragi and Reiser, 2002)

$$AUC = \Phi\left[\frac{a}{\sqrt{1+b^2}}\right] \quad (4)$$

here $a = \frac{\mu_x-\mu_y}{\sigma_x}$ and $b = \frac{\sigma_y}{\sigma_x}$.

Non Parametric Form of ROC Curves

In this approach, no theoretical model of the underlying populations is assumed and the ROC curve is constructed by using the empirical values of {S_n, 1-S_p} for each threshold value 'c'. The empirical ROC is of the form

$$R(p) = 1 - G\{F^{-1}(1-p)\}, \quad 0 \le p \le 1 \quad (5)$$

Here F and G are distribution functions of two populations. The next step in the analysis is the calculation of the AUC. In the nonparametric approach the AUC can also be estimated using *trapezoidal rule*. The empirical method has the advantage that it would not impose any assumption on structure of the data. It is shown that $\widehat{AUC}$ obtained by this method has equivalence with the well known Mann-Whitney U-statistic, two sample Wilcoxon rank sum statistic and the c-index [7].

The area under the ROC curve is given by [8]

$$AUC = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Optimal cutoff value: Among several cutoff values, we wish to choose the best one as optimal. In clinical applications, the cutoff is chosen in such a way that the test will have a low FPR (S_p < 0.10) or high TPR (S_n > 0.80). It

means that the optimal cutoff corresponds to the pair {min $(1-S_p)$, max (S_n)} among different cutoff values. Standard statistical software like SPSS produces the optimal cutoff as a part of the output. Medcalc is another software that supports ROC analysis.

Confidence Interval for AUC: Hanley and Mc Neil [5] have derived the standard error of the estimated AUC given by

$$S.E_A = \sqrt{\frac{A(1-A)+(n_D-1)(Q_1-A^2)+(n_H-1)(Q_2-A^2)}{n_D\, n_H}} \quad (7)$$

where $Q_1 = \left(\frac{A}{2-A}\right)$ and $Q_2 = \left(\frac{2A^2}{1+A}\right)$ and A denotes AUC.

The confidence interval is given by $A \pm Z_{\alpha/2} S.E_A$

Maximum Vertical Distance (MVD)

This is the simple index used to measure the difference of the ROC curve from the chance diagonal. The distance ranges between 0 and 1 where 0 indicates that the obtained ROC curve coincides with chance diagonal and 1 indicates a perfect classification. The conventional formula of MVD [7] is $MVD = \max |y(x)-x|$ (8)

The above equation can be represented using probabilistic definitions of ROC curve parameters and the expression for MVD can be rewritten as

$$MVD = \max_c |y(c)-x(c)| = \max_c |p(S>c|D)-p(S>c|\bar{D})| \quad (9)$$

$$\widehat{MVD} = \sup_{c\in(-\infty,\infty)} |\hat{F}(c)-\hat{G}(c)|$$

Here $\hat{F}$ and $\hat{G}$ are the empirical distribution functions of the two population distributions.

The above expression of MVD is just the Kolmogorov Smirnov statistic for two population distribution functions.

In the following section, we will provide the developed algorithms which are used to estimate the AUC, MVD and also helps in constructing ROC curve.

DISCRIMINANT ANALYSIS

Discriminant analysis is a technique where the discrimination is done by linear composites. A function that separates objects may sometimes serve as an allocator and one that allocates objects may sometimes act as discriminator. Using this technique, the discrimination function can be obtained which helps in classification of new individuals or objects into one of the groups. This technique helps in identifying the group centroids of various groups considered using which a cutoff can be obtained.

Let X_i (i = 1, 2, …..m) be the variable from $N_p(\mu_i,\Sigma_i)$ with n_i samples each. Assume that $\Sigma_1 = \Sigma_2 = \Sigma_3 = \ldots\ldots. \Sigma_m$. The optimality of classification depends on the assumptions of data which are

1. The predictor variables follow multivariate normal distribution.
2. The covariance matrices of different groups of data are homogeneous.

The Wilk's Lambda value explains the relationship between within-group variance and total variance. Wilk's [9] proposed Wilk's Lambda as

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{|\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'|}{|\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})'|} \quad (10)$$

If Box's M test statistic is significant, we can say that the covariance matrices are heterogeneous. Box [1] proposed M-statistic, where

$$M = (n-k)\ln|S_u| - \sum_{i=1}^{m}(n_i-1)\ln|S_{iu}| \quad (11)$$

$$S_u = \frac{nS}{n-k};\ S_{iu} = \frac{n_i S_i}{n_i-1}$$

According to Box MC^{-1} has a Chi square distribution with $\frac{1}{2}p(p+1)(m-1)$ degrees of freedom, where

$$C^{-1}=1-\frac{(2p^2+3p-1)}{6(p+1)(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1}-\frac{1}{n-k}\right] \quad (12)$$

The above MC^{-1} statistic is known as Box's M-test statistic.

The discriminant function is nothing but a linear combination of the variables in the study. It is given as

$$D=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_mx_m \quad (13)$$

where D is the discriminant score.

Let the populations under study be denoted as Π_1 and Π_2. The classification rule can be made functional and the object is placed under Π_1 [2] if

$$\left[(\bar{X}_1-\bar{X}_2)'S_{pooled}^{-1}\right]x-\frac{1}{2}(\bar{X}_1+\bar{X}_2)'S_{pooled}^{-1}(\bar{X}_1+\bar{X}_2)\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (14)$$

otherwise the object is assigned to Π_2. Here $\bar{X}_1$ and $\bar{X}_2$ are unbiased estimates of μ_1 and μ_2, respectively and S_{pooled} is the unbiased estimate of Σ. c(1|2) and c(2|1) are the costs of misclassification. p_1 and p_2 are prior probabilities of Π_1 and Π_2 respectively. Here the x value is the new observation which has to be classified into one of the groups.

EXCEL ALGORITHMS

Algorithm 1

Step-1: Enter the data in column B, preceded by a status variable in column A indicating H group by 0 and D group by 1.

Step-2: Put heading *Status* in the cell A2 and X in B2.

Step-3: Enter the data from B3 onwards with corresponding flag in column A.

Step-4: Sort the data in ascending order of X and rearrange them into two columns depending on the status value 0 or 1.

Step-5: Designate the two new columns with headings H group, D group in C2 and D2 respectively.

Step-6: Create the columns TP, TN, FP, FN, 1-SP and SN as headings in the cells E2, F2, G2, H2, I2 and J2.

Step-7: Count the TP, TN states using the excel paste function COUNTIFS ( )

TP = IF (B3< > "", COUNTIFS ($C$4:$C$22,">"&B3), COUNTIFS ($C$4:$C$22,">"&C3))

and

TN = IF (B3<>"", COUNTIFS ($B$4:$B$22,"<="&B3), COUNTIFS ($B$4:$B$22,"<="&C3))

Step-8: Compute FP = IF (B3< > "", COUNTIFS ($B$4:$B$22,">"&B3), COUNTIFS ($B$4:$B$22,">"&C3))

and FN = IF (B2<>"", COUNTIFS ($C$4:$C$22,"<="&B3), COUNTIFS ($C$4:$C$22,"<="&C3))

Step-9: Calculate 1-SP and SN as

1 – SP = 1 – (TN/ (TN+FP)) and SN = TP/(TP+FN)

Step-10: Plot (1 – SP) versus SN to obtain the ROC curve.

Step-11: The area under the curve can be computed using

$$AUC=\frac{TP+TN}{TP+TN+FP+FN}$$

Step-12 : Calculate Q_1 and Q_2 as

Q_1 = K5/(2-K5); Q_2 = (2*K5^2)/(1+K5)

Step-13 : The $S.E_A$ is computed as

$S.E_A$ = SQRT((K5*(1-K5)+(B1-1)*(K1-K5^2)+(D1-1)*(K2-K5^2))/(B1*D1))

Step-14 : The C.I is computed as

C.I = K5±K3*K6

Algorithm 2

Step-1: Enter the data in column B, preceded by a status variable in column A indicating H group by 0 and D group by 1.

Step-2: Put heading *Status* in the cell A2 and X in B2.

Step-3: Enter the data from B3 onwards with corresponding flag in column A.

Step-4: Sort the data in ascending order of X and rearrange them into two columns depending on the status value 0 or 1.

Step-5: Enter the different values of X in column C with a variable name t. This indicates the cutoff values. Sort them in Descending order.

Step-6: Compute $\hat{fp}$ and $\hat{tp}$ as

$\hat{fp}$= ((COUNTIFS($B$3:$B$82,">"&C3,$A$3:$A$82,0))/number of H's

$\hat{tp}$= ((COUNTIFS($B$3:$B$82,">"&C3,$A$3:$A$82,1))/number of D's

$$\hat{F}(c) = 1\text{-}fp \text{ and } \hat{G}(c)=1\text{-}tp$$

Step-7: The maximum vertical distance is obtained as

$$MVD=MAX\left(ABS\left(\hat{F}(c)\text{-}\hat{G}(c)\right)\right)$$

RESULTS AND DISCUSSIONS

The two different data sets used are, Tuberculosis (TB) and Nephrology – CVS (N – CVS). Three variables such as Pleural fluid Adenosine Deaminase (ADA) of TB data, Calcium and Albumin of N – CVS data were considered to calculate Area under the curve (AUC) and Maximum Vertical Distance (MVD) measures. In Tuberculosis data set, we have a Status variable which indicates the presence (1) or absence (0) of the disease, along with this the additional variables are age, sex and ADA. In N - CVS data set, we have a Status variable and fifteen other variables are also present. Using Binary Logistic regression, five important variables namely Calcium, Albumin, age, sex and Diabetic Mellitus (DM) have been identified.

The entire statistical analysis was performed using various multivariate techniques such as Linear Discriminant Analysis, Binary Logistic Regression and Receiver Operating Characteristic (ROC) curve. The entire work has been segregated into two segments; one is computation of the AUC and MVD along with confidence intervals (C.I) of the AUC. In the same segment, it has been executed and interpreted that the measure MVD is an analogy for the measure AUC. Importance of both the measures was also highlighted. The second segment constitutes of the exploratory way of understanding and analyzing the Linear Discriminant Functions by involving cofactors in the model. The structural change in the Linear Discriminant Function was observed by using various combinations of cofactors.

To compute the AUC, MVD and C.I of AUC spreadsheet algorithms were developed and the same have been used in identifying the threshold values. An experiment has been done to classify the individuals using Discriminant scores instead of the clinical cutoffs. Basing on the classification status the average discriminant score is used to classify the individual as healthy or diseased.

Variables	MVD	Threshold	AUC	C.I	
				L.L	U.L
ADA	1	36	1	1	1
CAL	0.3742	9.6	0.719	0.6484	0.7904
ALB	0.17566	3.5	0.585	0.5156	0.6543

Using algorithm 1 AUC's of the ROC curves for the variables considered are obtained along with their confidence intervals. The Maximum vertical distance is computed using algorithm 2 and it points directly to the cutoff (threshold) value. It is also observed that the maximum vertical distance and AUC are directly proportional to each other. In the case of ADA, the MVD value is 1 and the AUC is also 1 indicating a 100% correct classification. Calcium shows a moderate case with an AUC 0.719 indicating that 71.9% of

cases are classified correctly and albumin, a worst case with an AUC of 0.585.

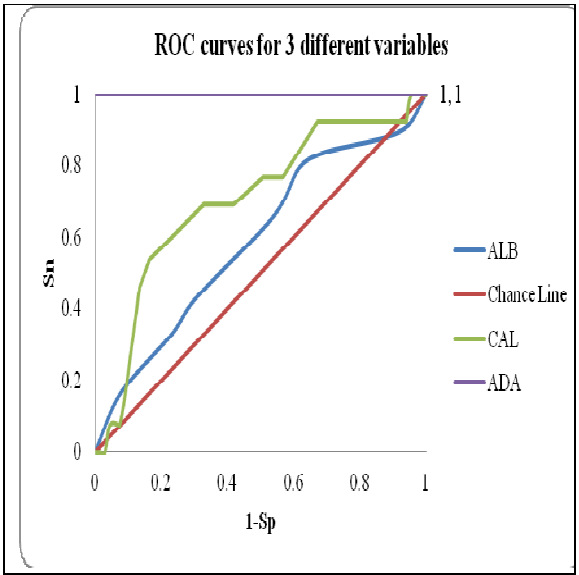

Figure 1: Different forms of ROC curves.

In figure 1, we can see the ROC curves for three different variables. The plot of ADA touches the left corner of the graph indicating a perfect classification. The curves of Calcium and Albumin represent the moderate and worst cases. Further, Discriminant Analysis has been carried out with various combinations of cofactors in the model. This was carried out for both the data sets. The main purpose of handling the model with various combinations of cofactors is to observe the significant impact of those cofactors on percentage of correct classification. The statistics were reported with a brief explanation for each model which was considered. The exploratory analysis was performed for both the data sets with outliers and without outliers in order to observe sensitiveness of the models which are under study. In fact, the outliers were identified only in TB data set and not in N – CVS data set. The models sensitivity was observed for the TB data set and their implications were focused. The linear discriminant functions have been obtained to classify the new observations into either diseased or healthy groups. Variables such as age and sex have an important role in determining the health condition. Hence, they have been used to obtain the discriminant function.

Tuberculosis data set:

	Discriminant Equation	% of Correct Classification	d_i
With Outliers	D = -2.642 + 0.046*ADA	90.0	-0.433
	D = -1.356 + 0.044*ADA - 0.031*Age	97.0	-0.485
	D = -1.373 + 0.044*ADA - 0.031*Age +0.071*Sex	98.0	-0.485
Without Outliers	D = -3.733 + 0.064*ADA	96.3	-0.722
	D = -2.935 + 0.063*ADA – 0.020*Age	98.8	-0.753
	D = -3.081 + 0.062*ADA – 0.023*Age + 0.433*Sex	97.5	-0.768

d_i : Discriminant Score Cutoff

The above table briefly summarizes the variables used in the model and percentages of classification with and without outliers when TB data set was used. When outliers are present in the data, the percentage of classification is highest (98%) when the variables ADA, Age and Sex were included. After removing the outliers, the percentage of classification is highest (98.8%) when the variables ADA and Age were included in the model.

Nephrology – CVS data set

Discriminant Equation	% of Correct classification	d_i
D = -21.361 + 2.281*CAL	60	0.203
D = -19.324 +1.639*CAL + 0.076*Age	66.3	0.312
D = -18.660 + 1.444*CAL +0.069*Age +1.213*Sex	66.3	0.369
D = -10.006 + 2.651*ALB	62.5	-0.129
D = -1.517 - 0.793*ALB +0.086*Age	63.8	0.227
D = -4.775 – 0.281*ALB + 0.077*Age + 1.419*Sex	68.8	0.288
D = -17.061 + 0.064*Age + 1.147*Sex – 0.362*ALB + 1.456*CAL	68.8	0.373

d_i: Discriminant Score Cutoff

When N – CVS data set was taken into consideration the highest percentage of classification was found to be 68.8%. This percentage is obtained in two cases i.e., when Albumin, Age and Sex were included and Calcium, Albumin, Age and Sex were included.

This indicates that though calcium was found to be important it need not be included in the model for classification using discriminant analysis.

REFERENCES

1. Box, G.E.P (1949), A General Distribution Theory for a class of Likelihood Criteria, Biometrika, 36, 317-346.
2. Carl J Huberty (1994), Applied Discriminant Analysis, John Wiley and Sons.
3. David Farragi and Benjamin Reiser (2002), Estimation of the area under the ROC curve , Statistic in Medicine; 21; 3090-3106.
4. Donna Katzman Mc Clish (1989), Analyzing a Portion of the ROC curve , Medical Decision Making 1989; 9; 190-195.
5. James A Hanley, Barbara J Mc Neil (1982), A Meaning and use of the area under a receiver Operating Characteristics (ROC) curves , Radiology; 143; 29-36.
6. James A Hanley, Barbara J Mc Neil, (1983), A method of Comparing the Areas Under Receiver Operating Characteristics Analysis derived from the same cases, Radiology; 148; 839-843.
7. Krazonowski W and Hand D J (2009), ROC curves for Continuous Data , Chapman and Hall.
8. Thompson ML (2003), Assessing the diagnostic accuracy of a sequence of tests , Biostatistics; 4 (3): 341-351.
9. Wilks, S.S. (1932), Certain Generalizations in the Analysis of Variance, Biometrica, 24, 471-494.