A Divergence Measure for STROC Curve in Binary Classification

S. Balaswamy¹, R. Vishnu Vardhan¹*, and MB Rao²

Received 3 September 2014; Published online 20 September 2014

© The author(s) 2014. Published with open access at www.uscip.us

Abstract

Receiver Operating Characteristic (ROC) Curve is used to classify the individuals into one of the two populations by assessing the performance of a diagnostic test. In practice, we come across the situation where there is a need to define the lower and upper truncation points. In this paper, an ROC model based on truncated distributions is proposed and its importance in classification problems is highlighted. The accuracy of a diagnostic test depends on two populations and their characteristics as well as their overlapping area. In order to estimate the overlapping area or distance between both populations, the concept of Kullback Leibler Divergence (KLD) is applied for Single Truncated ROC (STROC) curve as well as to study the asymmetry case of STROC Curve. To explain this phenomenon, KLD has been estimated using a real data set.

Keywords: ROC Curve; Truncated Distribution; STROC Curve and Kullback - Leibler Divergence

1. Introduction

Over the past seven decades, the ROC Curve has made its landmark in classification problems which helps in assessing the performance of a test or diagnostic procedure. The ROC Curve is embedded by the two intrinsic measures Sensitivity ($S_n$) and Specificity ($S_p$) along with its accuracy measure: Area under the ROC Curve (AUC). The term ROC analysis was coined during II world war to analyze the radar signals (Peterson, Birdsall and Fox, 1954). The application of ROC Curve technique was promoted in diversified fields such as experimental psychology (Green and Swets, 1966), industrial quality control (Drury and Fox, 1975) and military monitoring (Swets, 1977).

Green and Swets (1966) was first to use the Gaussian model for estimating the ROC Curve. Dorfman and Alf (1968, 1969) gave the maximum likelihood estimates for the ROC Curve parameters considering yes or no type responses and rating data. The importance of ROC Curve in medicine was due to Leo Lusted (1971) to analyze the radiographic images. Metz et al. (1998) proposed a methodology to descrizitze the scores of rating type by embedding the approach of Dorfman and Alf (1969). Hanley (1988) explained the importance and robustness of Binormal ROC Curve. Metz (1989) proposed a unique type of ROC Curve where the decision variable is a strictly increasing function of the likelihood ratio, named it as the “Proper” ROC Curve. Earlier work on this Proper ROC Curve was due to Egan

---

*Corresponding author, e-mail: rvvcr@gmail.com
1* Department of Statistics, Pondicherry University, Puducherry – 605 014, India.
2 Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio 45267 – 0056.
(1975) which takes into the account of chi-square and gamma distributions. England (1988) focused on identifying the optimal thresholds by proposing an exponential type ROC Curve. The ROC models based on Lomax and Gamma distributions was given by Dorfman et al. (1997). In later years, the estimation of parameters of ROC Curves was brought into the setup of regression framework by Pepe (1997). Further different methodologies to estimate the ROC Curve by studying the effect of covariates where the data is of continuous type is given by Pepe (1998). Goodness of fit procedures and inferential aspects to test the accuracy measure and intrinsic measures of ROC Curve was given by K.H. Zou et al. (2005) and Bio Zhang (2008). For more details on the methodological development of ROC Curves can be found in Zou et al. (2002) and Krzanowski and Hand (2009).

In this paper, we have proposed a newer version of ROC, which focuses on some particular situations where the data will have extended tails. In such situations, the conventional Bi-normal ROC will not take into account of true accuracy. Here, we have highlighted the concept of truncation in handling such extended tailed data. Further, it is shown that the expected value of log likelihood ratio of proposed ROC Curve is equivalent to the KLD, along with this the asymmetric properties of proposed ROC Curve are also discussed.

2. Methodology

The ROC Curve is a tradeoff between false positive rate (FPR) and true positive rate (TPR) of the test and is defined as \( \text{ROC}(t) = 1 - G[F^{-1}(1 - t)], \ t \in [0, 1] \); where \( F \) and \( G \) are the cumulative density functions of normal \( (X) \) and abnormal \( (Y) \) populations respectively. Here, \( t \) is a threshold, which plays an important role in classifying the individuals into one of the two groups. Conventionally the fitting and estimation of Binormal ROC Curve and its parameters are done using the full information of a particular scenario. However, we can come across a situation where few samples may not be required for fitting the model. i.e., if the samples are restricted to values which lie above or below a given point or the researcher might be interested in some particular range. In other words, the density curve of a particular scenario with extended tails influences in handling the true information of the data and it’s characteristics. The dataset CA19-9 (Del Villano et al. 1983 & Zhou et al. 2002) considered in this paper meets the above explanation, that is, there are many samples far from the cutoff, where their group status can easily be identified and the tails of the data is observed to be stretched at the end points. In such situations, the concept of truncation can be incorporated to extract the true hidden accuracy of the biomarker.

In this paper, we propose an ROC Curve which is based on the concept of truncation and made an attempt to show that for data with extended tails, how the form of truncated ROC curve performs better rather than the conventional Binormal ROC curve. For normal population, the truncation point is considered on the left \((a)\) and in the case of abnormal population, the truncation point is taken on the right \((b)\) since, the extreme small and large values are usually observed in normal and abnormal populations respectively. Hence, it is named as Single Truncated Receiver Operating Characteristic (STROC) Curve.
Let \( X \) and \( Y \) be two random variables of normal (H) and abnormal (D) populations which follow left and right truncated normal distributions respectively. Let \( f(x) \) and \( g(y) \) be the probability density functions of \( X \) and \( Y \) along with their cumulative distribution functions \( F(x) \) and \( G(y) \). The expressions of probability density function and cumulative distribution function of \( X \) and \( Y \) are given as,

\[
f(x) = \frac{1}{\sigma_H} \phi \left( \frac{x - \mu_H}{\sigma_H} \right); \quad a \leq x \leq \infty \quad \text{and} \quad g(y) = \frac{1}{\sigma_D} \phi \left( \frac{y - \mu_D}{\sigma_D} \right); \quad -\infty \leq y \leq b
\]

\[
F_X(x) = \frac{\phi \left( \frac{x - \mu_H}{\sigma_H} \right) - \phi \left( \frac{a - \mu_H}{\sigma_H} \right)}{1 - \phi \left( \frac{a - \mu_H}{\sigma_H} \right)}; \quad a \leq x \leq \infty
\]

\[
G_Y(y) = \frac{\phi \left( \frac{y - \mu_D}{\sigma_D} \right)}{\phi \left( \frac{b - \mu_D}{\sigma_D} \right)}; \quad -\infty \leq y \leq b
\]

where ‘\( a \)’ and ‘\( b \)’ are the left and right truncation points considered in normal and abnormal populations respectively.

As per the definition of ROC Curve, the STROC Curve can also be defined as the tradeoff between FPR and TPR. Let us consider, the false positive rate at a particular threshold \( t \) as,

\[
x(t) = p(S > t | H) = 1 - \frac{\Phi \left( \frac{t - \mu_H}{\sigma_H} \right) - \Phi \left( \frac{a - \mu_H}{\sigma_H} \right)}{1 - \Phi \left( \frac{a - \mu_H}{\sigma_H} \right)}
\]

on further simplification, threshold ‘\( t \)’ can be obtained as,

\[
t = \mu_H + \sigma_H \Phi^{-1} \left( x(t) \int \Phi \left( \frac{a - \mu_H}{\sigma_H} \right) \right)
\]

Let us define the true positive rate for threshold ‘\( t \)’ is,

\[
y(t) = p(S > t | D) = 1 - \frac{\Phi \left( \frac{t - \mu_D}{\sigma_D} \right)}{\Phi \left( \frac{b - \mu_D}{\sigma_D} \right)}
\]
Now, by substituting (3) in (4) we get, the expression for Single Truncated ROC (STROC) Curve,

$$y(t) = \frac{\phi\left(\frac{\mu_D - \mu_H}{\sigma_D}, \frac{\sigma_D}{\sigma_D} \phi^{-1}(x(t) - x(t)\phi(\alpha_H))\right) - \phi\left(\frac{\mu_D - b}{\sigma_D}\right)}{\phi\left(\frac{b - \mu_D}{\sigma_D}\right)}$$  \hspace{1cm} (5)

STROC Curve has a positive slope and is monotonically increasing with false positive rate and invariant under strictly increasing transformation of the scores. Whenever \(b \to \infty\) and \(a \to -\infty\), then the truncated normal distribution will tend to the normal distribution. Similarly, if \(b \to \infty\) and \(a \to -\infty\), then STROC Curve also reduces to the form of Binormal ROC curve, i.e.,

$$y(t) = \Phi\left(\frac{\mu_D + \mu_H}{\sigma_D} + \frac{\sigma_H}{\sigma_D} \Phi^{-1}[x(t)]\right)$$

### 2.1 Area under the STROC Curve

The Area under the Curve is a useful and popular index for assessing the overall accuracy of a diagnostic test. For a test, if AUC equals to one, it is named as a perfect test and with \(AUC = 0.5\), it is named as a worst test which cannot be used for further classification. AUC can be interpreted as the probability that the test will correctly rank a randomly chosen case or non-case pair with respect to their test values (Bamber 1975).

The Area under the STROC Curve (AUC) can be obtained by integrating the expression (5) over the range \([0, 1]\).

$$AUC = \int_{0}^{1} y(t) \, dx(t)$$

$$AUC = \int_{0}^{1} \frac{1}{\phi\left(\frac{b - \mu_D}{\sigma_D}\right)} \left[\phi\left(\frac{\mu_D - \mu_H}{\sigma_D}, \frac{\sigma_D}{\sigma_D} \phi^{-1}(x(t) - x(t)\phi(\alpha_H))\right) - \phi\left(\frac{\mu_D - b}{\sigma_D}\right)\right] \, dx(t)$$

on integrating and simplifying, the expression for AUC of STROC is

$$AUC = \frac{1}{\phi\left(\frac{b - \mu_D}{\sigma_D}\right)} \left[\frac{1}{2} + \left(\frac{\mu_D - \mu_H}{\sigma_D}\right)^2 \sigma_H e^{-2\phi(\alpha_H)} - 2\pi \sigma_D \left(1 - \phi\left(\frac{\mu_D - b}{\sigma_D}\right)\right)\right]$$  \hspace{1cm} (6)

The major importance of classification lies in estimating overlapping area of both distributions. In order to explain the overlapping area of two populations (in other sense the distance between both distributions), the proximity measure namely Kullback - Leibler Divergence (KLD) is used as an intrinsic measure of STROC Curve.

### 3. A Divergence Measure (KLD) for STROC Curve

The Kullback - Leibler Divergence (KLD) is an essential equation of information theory that quantifies the proximity of two probability density functions and also based on the likelihood theory (Cover and Thomas 1991). The KLD is usually based on two probability density functions (Kullback & Leibler 1951) and is defined as,

$$\text{KLD}[f||g]=E_{f}\left[\log\frac{f(x)}{g(x)}\right] = \int f(x) \log\frac{f(x)}{g(x)} \, dx$$  \hspace{1cm} (7)

$$\text{KLD}[g||f]=E_{g}\left[\log\frac{g(x)}{f(x)}\right] = \int g(x) \log\frac{g(x)}{f(x)} \, dx$$  \hspace{1cm} (8)
To study and interpret the STROC Curve, the information measure KLD has been used to identify the closeness between both distributions of normal and abnormal populations. In this paper, few results are proposed which focuses on the functional relationship between slope of STROC Curve and KLD and the asymmetric properties of STROC Curve.

3.1. Result: The expected value of log likelihood ratio of STROC Curve is equivalent to the divergence measure KLD.

\[
E_g \left[ \log \frac{dy}{dx} \right] = KLD[g||f] \tag{9}
\]

\[
E_f \left[ \log \left( \frac{dy}{dx} \right)^{-1} \right] = KLD[f||g] \tag{10}
\]

Proof: Let us consider the likelihood ratio of STROC Curve as,

\[
\frac{dy}{dx} = \frac{dx}{dt}
\]

where, \( \frac{dy}{dt} = \frac{d}{dt} \{ P(X>t|D) \} = -g(t) \)

and \( \frac{dx}{dt} = \frac{d}{dt} \{ P(X>t|H) \} = -f(t) \)

\[
\therefore \frac{dy}{dx} = \frac{-g(t)}{-f(t)} = \frac{g(t)}{f(t)}
\]

on substituting the above likelihood ratio of STROC Curve value in equation (8), one can get the divergence measure for STROC Curve as,

\[
E_g \left[ \log \frac{dy}{dx} \right] = KLD[g||f] \]

Similarly,

\[
E_f \left[ \log \left( \frac{dy}{dx} \right)^{-1} \right] = KLD[f||g]
\]

i.e. The proximity measure KLD is equivalent to the expected value of log likelihood ratio of STROC Curve.

Conventionally, the abnormal mean will always be greater than the normal mean. Hence, the measure KLD[g||f] is prominent to study ROC curve (Hughes and Bhattacharya 2013). The slope of STROC Curve is defined as,

\[
\frac{dy}{dx} = \frac{\sigma_H}{\sigma_D} e^{\frac{1}{2} \left[ \frac{(t-\mu_H)^2}{\sigma_H^2} - \frac{(t-\mu_D)^2}{\sigma_D^2} \right] + \frac{1}{\phi \left( \frac{b-\mu_H}{\sigma_H} \right) - \frac{1}{\phi \left( \frac{b-\mu_D}{\sigma_D} \right)}}} \tag{11}
\]

applying log and taking expectation on both sides for above equation (11), we get
Similarly, one can find the expression for KLD[f||g],

$$KLD[f||g] = \frac{1}{2} \left( \frac{\sigma_f^2}{\sigma_D^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_D^2} - \log \left( \frac{\sigma_f^2}{\sigma_D^2} \right) + \log \left[ \frac{1 - \Phi \left( \frac{a - \mu_H}{\sigma_H} \right)}{\Phi \left( \frac{b - \mu_D}{\sigma_D} \right)} \right]$$

The above equation is known as the KLD expression for STROC Curve, which is used to measure the distance between normal and abnormal populations of STROC curve.

In particular, it can be shown that when $b \to \infty$, then $\Phi \left( \frac{b - \mu_D}{\sigma_D} \right) = 1$ and $a \to -\infty$, then $\Phi \left( \frac{a + \mu_H}{\sigma_H} \right) = 0$, then the expression in (12) reduces to,

$$KLD[g||f] = \frac{1}{2} \left( \frac{\sigma_f^2}{\sigma_H^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_H^2} - \log \left( \frac{\sigma_f^2}{\sigma_H^2} \right) + \log \left[ \frac{1 - \Phi \left( \frac{a - \mu_H}{\sigma_H} \right)}{\Phi \left( \frac{b - \mu_D}{\sigma_D} \right)} \right]$$

which is the KLD expression for Binormal ROC curve (Hughes and Bhattacharya 2013). In case of Binormal ROC Curve if $\sigma_H = \sigma_D$, then their KLD's are found to be equal, and it is known as symmetric. It is well known that KLD[f||g] ≠ KLD[g||f] and KLD[f||g] ≥ 0 and equality holds if and only if f = g (Kenneth et al. 2002).

The ROC Curves may be symmetric or asymmetric in nature. The measure KLD is shown to be useful for detecting asymmetry in distributions of categorical or continuous random variables. Whenever the sum of $1-S_p$ and $S_n$ equals to one then the corresponding ROC Curve is symmetric about the negative diagonal (Peter and Thomas 2004). Bi-normal ROC Curves are symmetric when standard deviations of both populations equal and imply that their KLD's are also equal (Johnson and Dudgeon 1993). Further, the KLD measure is used as a discriminate measure to classify the two different distributions such as Weibull and Log-normal (Ali Akbar, 2012).

Here, using a result it is shown that the STROC Curve is asymmetric and a particular case is also presented to show when the STROC Curve will be symmetric in nature. The asymmetric ROC Curves are of TPP asymmetric and TNP asymmetric about the negative diagonal. The TPP asymmetric ROC Curve leaned to the top left edge of ROC space longer than it does to top whereas TNP asymmetric ROC Curve is leaned to the top edge of ROC space longer than it does to the left (Hughes and Bhattacharya 2013).

3.2. **Result:** If KLD[g||f]=KLD[f||g], then $\Phi \left( \frac{a - \mu_H}{\sigma} \right) = \Phi \left( \frac{b - \mu_D}{\sigma} \right)$, which is a particular case to prove that the STROC Curve is symmetric, here $\sigma_H = \sigma_D = \sigma$.

**Proof:** Kenneth et al. (2002) proved that KLD[f||g]≠KLD[g||f] and equality holds if f=g.

Let us assume that the pdf’s of two populations are equal, i.e. f=g then the following equality holds.

$$KLD[g||f]=KLD[f||g]$$

Now, using equation (14) we prove that STROC is symmetric.

Recall the expressions (12) & (13), we have
when \( f=g \) condition holds, which means that \( \sigma_H = \sigma_D = \sigma \). Therefore, on further simplification, we can write the above equality as

\[
\log \left[ \frac{1 - \Phi \left( \frac{a - \mu_H}{\sigma} \right)}{\Phi \left( \frac{b - \mu_D}{\sigma} \right)} \right] = \log \left[ \frac{\Phi \left( \frac{b - \mu_D}{\sigma} \right)}{1 - \Phi \left( \frac{a - \mu_H}{\sigma} \right)} \right]
\]

This clearly states that STROC Curve is symmetric in nature. If the condition \( f=g \) fails, then STROC Curve possesses asymmetric nature.

Therefore, the STROC Curve is asymmetric towards the anti diagonal of the STROC plot. The asymmetric properties of STROC Curve are explained in the following result.

3.3. \textbf{Result:} The STROC Curve is TPP asymmetric if \( \beta < 1 \) & \( k < 1 \) and is TNP asymmetric if \( \beta > 1 \) & \( k > 1 \), where \( \beta = \frac{\sigma_H}{\sigma_D} \) and \( k = \frac{1 - \phi \left( \frac{a - \mu_H}{\sigma_H} \right)}{\phi \left( \frac{b - \mu_D}{\sigma_D} \right)} \).

\textbf{Proof:} Again reconsider the expressions (12) & (13),

\[
KLD[g||f] = \frac{1}{2} \left[ \left( \frac{1}{\beta^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_H^2} - \log \left( \frac{1}{\beta^2} \right) \right] + \log[k]
\]

similarly,

\[
KLD[f||g] = \frac{1}{2} \left[ (\beta^2 - 1) + \frac{(\mu_D - \mu_H)^2}{\sigma_D^2} - \log(\beta^2) \right] + \log \left[ \frac{1}{k} \right]
\]

Under the Binormal case, if \( \beta \gg 1 \), then \( KLD[g||f] \gg KLD[f||g] \).

\[
i. e. \frac{1}{2} \left[ \left( \frac{1}{\beta^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_H^2} - \log \left( \frac{1}{\beta^2} \right) \right] \gg \frac{1}{2} \left[ \left( \beta^2 - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_D^2} - \log(\beta^2) \right]
\]

Let us consider the conditions \( a < \mu_H, b > \mu_D, a < b \) and \( \mu_D > \mu_H \).

Initially, using the above conditions we prove that STROC Curve is TPP asymmetric. Let us assume that \( \beta < 1 \) which implies \( \sigma_H < \sigma_D \).

Consider the condition \( a < b \), which can also be expressed as,

\[
i. e. \mu_H - a < b - \mu_D
\]

on further simplification using the condition \( \sigma_H < \sigma_D \), one can get
when \( k < 1 \), then \( \log(k) < \log(1/k) \) and thereby using expression (17), we can see that

\[
\frac{1}{2} \left( \frac{1}{\beta^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_H^2} \log \left( \frac{1}{\beta^2} \right) + \log[k] > \frac{1}{2} \left( (\beta^2 - 1) \ast \left( \frac{\mu_D - \mu_H}{\sigma_D^2} \right)^2 \log(\beta^2) \right) + \log \left( \frac{1}{k} \right)
\]

i.e. \( \text{KLD}[g||f] > \text{KLD}[f||g] \) for \( \beta < 1 \) and \( k < 1 \), therefore, STROC Curve is TPP asymmetric.

Similarly, when \( \beta > 1 \), \( k > 1 \), then \( \log(k) > \log(1/k) \) and thereby using expression (17), we can see that

\[
\frac{1}{2} \left( \frac{1}{\beta^2} - 1 \right) + \frac{(\mu_D - \mu_H)^2}{\sigma_H^2} \log \left( \frac{1}{\beta^2} \right) + \log[k] < \frac{1}{2} \left( (\beta^2 - 1) \ast \left( \frac{\mu_D - \mu_H}{\sigma_D^2} \right)^2 \log(\beta^2) \right) + \log \left( \frac{1}{k} \right)
\]

i.e. \( \text{KLD}[g||f] < \text{KLD}[f||g] \) for \( \beta > 1 \) and \( k > 1 \), therefore, STROC Curve is TNP asymmetric. From these results, it is clear that the STROC Curve is TPP asymmetric if \( \beta < 1 \) \& \( k < 1 \) and is TNP asymmetric if \( \beta > 1 \) \& \( k > 1 \).

4. Results and Discussion

In this section, the proposed methodology is demonstrated using a Pancreatic Cancer data set conducted at Mayo Clinic with the relative accuracy of a biomarker namely CA19-9 with cutoff >37 units/mL (Del Villano et al. 1983 & Zhou et al. 2002). The serum concentrations were collected for the biomarker from 51 "control" patients with pancreatitis and 90 "case" patients measured on a continuous scale.

### Table 1 Results of KLD when truncation points are varied on Pancreatic Cancer Dataset

| a   | b   | \( \mu_H \) | \( \mu_D \) | \( \sigma_H \) | \( \sigma_D \) | Slope (\( \beta \)) | AUC  | KLD[g||f] | KLD[f||g] |
|-----|-----|-------------|-------------|----------------|----------------|---------------------|------|-----------|-----------|
| 6.5 | 1079| 22.1820     | 218.1628    | 22.2476        | 271.3892       | 0.0819              | 0.7379| 109.9269  | 2.5396    |
| 6.5 | 592 | 22.1820     | 152.7047    | 22.2476        | 178.1096       | 0.1249              | 0.7276| 46.4077  | 2.1246    |
| 6.5 | 369 | 22.1820     | 95.9623     | 22.2476        | 110.7215       | 0.2009              | 0.7047| 15.5102  | 1.6151    |
| 6.5 | 251 | 22.1820     | 69.8278     | 22.2476        | 79.3855        | 0.2802              | 0.6715| 6.6237   | 1.2551    |
| 11.5| 1079| 33.0090     | 218.1628    | 24.7418        | 271.3892       | 0.0911              | 0.7194| 85.0506  | 2.3447    |
| 11.5| 592 | 33.0090     | 152.7047    | 24.7418        | 178.1096       | 0.1389              | 0.7102| 34.9322  | 1.9161    |
| 11.5| 369 | 33.0090     | 95.9623     | 24.7418        | 110.7215       | 0.2234              | 0.6677| 11.0448  | 1.3918    |
| 11.5| 251 | 33.0090     | 69.8278     | 24.7418        | 79.3855        | 0.3116              | 0.6202| 4.3865  | 1.0242    |
In order to define the left and right truncation points, scatter plots of the data are plotted. From that, gaps at the extreme left and right points are identified. Basing on the gaps in the data, the truncation points are defined. Another point to note is that the extreme values have influence in suppressing the actual accuracy of the biomarker and that has been improved by defining truncation points. In this work, the proposal is to execute the importance of truncation where it will help to increase the accuracy of a biomarker. This dataset’s accuracy is observed to be 0.6903 (69.03%). On incorporating the concept of truncation, the true performance of CA19-9 is obtained. In Table 1, the means and standard deviations of both populations are reported along with the AUC at various truncation points. At \(a = 6.5\) and \(b = 1079\), the AUC is observed to be 0.7379 (73.79%) which is comparatively better percentage of accuracy. This is due to restricting the influence of extreme values on the left and right of populations respectively. This exhibits the point that the AUC obtained from the full data may get influenced by the extreme values. In other words, when there is a stretch in the tails of the population density, then we can incorporate the concept of truncation without loss of information; there by the true accuracy of the test procedure can be explained. Figure 3 visualizes the shapes of the two populations as we change the left and right truncation points. The point to highlight only the truncation point \(a = 6.5, 11.5\) and \(b = 1079\) is that a less percentage of overlapping area is observed and this is also an important objective in classification problems. If we look into the density curves of the remaining truncation points, the overlapping area is gradually increasing and this will increase the percentage of misclassification rate too. So, the truncation points are to be defined in such a way that it should explain the true accuracy of the test as well should minimize the overlapping area.

The ROC curves are plotted at each truncation point and are graphed in Fig. 2(a), 2(b). The ROC curves at \(a = \{6.5, 11.5\}\) and \(b = 1079\) tend to move towards top left corner and rest of the points are under these two curves. This is another way of explaining the importance truncation rather than handling the data with conventional Binormal ROC curve. In the Fig. 2(c), the Binormal ROC curve is closer to the chance line where as the STROC curve at truncation points provide better visualization and interpretation about the biomarker.
(c). Plot of Binormal and STROC curves

Fig. 2. STROC Curves for Pancreatic Cancer Dataset at various truncation points

Fig. 3. Plots of Density Curves for Pancreatic Cancer Data at various truncation points
Since the data follows heavy tailed pattern, therefore, we have defined the truncation points and then applied the concept of Kullback Leibler Divergence to explain the phenomenon of STROC Curve and its symmetric, asymmetric properties. Table 1, reports the basic characteristics such as means and standard deviations at each truncation point along with slope ($\beta$) & KLD values. In the context of STROC Curve, k is less than one ($k<1$), since abnormal test scores are always higher than the normal test scores. It is observed that as the slope of STROC Curve is less than one ($\beta<1$) and $k<1$ then $\text{KLD}[g||f] > \text{KLD}[f||g]$. From the result (3.3) of asymmetry, when $\beta<1$ & $k<1$, then the STROC Curve is TPP asymmetric about the negative diagonal. It is also observed that the slope of STROC is high at extreme truncation points compared to the other truncation points. Thereby with the influence of slope and truncation points (see equations 9 & 10), KLD will attain maximum value at extreme truncation points. The $\text{KLD}[g||f]$ value for the case of Binormal ROC Curve is found to be very high (16782.22 bits) because of the influence of extreme points of CA19-9 biomarker.

From Table 1, the KLD measure attains the maximum value 109.9269 at extreme truncation points $a=6.5$ and $b=1079$ and minimum value 4.3865 at closer truncation points $a=11.5$ and $b=251$. Which means as the truncation points moves closer to each other, both KLD and AUC values diminishes. Here, whatever the phenomenon explained with the help of KLD of STROC Curve, the similar kind of information is observed with AUC. Therefore, the AUC and KLD are directly proportional to each other. i.e. as the AUC increases, the KLD also increases at the defined truncation points (Table 1). It is well known that whenever, the KLD has higher value then the proximity distance between both populations will be more. In the concept of ROC Curve analysis, whenever the distance between both populations is more then there is less percentage of misclassification rate thereby the true accuracy measure will be more.

![KLD's for STROC Curve at a=6.5](image1)

![KLD's for STROC Curve at a=11.5](image2)

**Fig. 4.** KLD Plots of Pancreatic Cancer Dataset at various truncation points

To observe the effect of truncation points, the KLD’s of STROC curve are also drawn with varying slope (Table 1) of STROC Curve at different combinations of truncation points (Fig. 4). From Figure 4, it can be visualized that as the slope of STROC increases, then the KLD measure attains the decrement pattern. Clearly, it can be seen that when $\sigma_R < \sigma_N$ (i.e. $\beta<1$) then $\text{KLD}[g||f] > \text{KLD}[f||g]$. 

78
Therefore, the STROC Curve is TPP asymmetric about the negative diagonal and the same is depicted in Fig 5. In some cases, if slope of STROC Curve is greater than one ($\beta > 1$) and $k > 1$, then the STROC Curve is TNP asymmetric about negative diagonal.

![The TPP Asymmetric property of STROC Curve](image)

**Fig. 5.** The TPP Asymmetric property of STROC Curve

### 5. Conclusion

In classification problems, the importance and need of truncation is taken into account and it is demonstrated using a real data set. The overall accuracy of full data (untruncated) is obtained as 0.6903 where as the AUC values obtained after defining various truncation points are reported in table 1. In that, at point $a=6.5$ and $b=1079$, the AUC is obtained as 0.7379, which seems to be better than the AUC of conventional Binormal ROC Curve. To support the above information, the density curves are plotted at various truncation points, of these minimum overlapping area is obtained at $a=6.5$ and $b=1079$ and rest of truncation points have a larger overlapping area. Similar kind of phenomenon, is observed at $a=11.5$ and $b=1079$, where the AUC is 0.7194 which is also better than the AUC of Binormal ROC Curve. The main focus of the paper is to show that whenever data has stretched tails then that will not provide true information about the percentage of correct classification. In such situations, the concept of Truncated ROC Curve can be used to extract the true accuracy of the test rather than the conventional Binormal ROC curve.

It is also observed that whenever the slope of STROC Curve is higher, the corresponding KLD is found to be high and vice versa. It is also observed that when $\beta < 1$ and $k < 1$ then $\text{KLD}(g||f) > \text{KLD}(f||g)$ and when $\beta > 1$ and $k > 1$ then $\text{KLD}(g||f) < \text{KLD}(f||g)$. In the context of ROC, slope is always less than or equal to one, thereby the STROC Curve is TPP asymmetric about the negative diagonal. Therefore, the proposed methodology will be very useful to study the asymmetry properties of STROC Curve and to explain the overlapping area of two populations in classification.
Acknowledgement

The authors would like to thank the reviewers for their valuable comments and suggestions for better improvisation of the paper.

References


http://dx.doi.org/10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z


http://dx.doi.org/10.1007/BF02289677


http://dx.doi.org/10.1016/0022-2496(69)90019-4


http://dx.doi.org/10.1016/S1076-6332(97)80013-X


http://dx.doi.org/10.3390/e15041342


http://dx.doi.org/10.1177/0272989X8800800308


http://dx.doi.org/10.1016/j.jbi.2005.02.004


http://dx.doi.org/10.1126/science.171.3977.1217

http://dx.doi.org/10.1097/00004424-198903000-00012

http://dx.doi.org/10.1093/biomet/84.3.595

http://dx.doi.org/10.2307/2534001

http://dx.doi.org/10.1016/j.jmp.2004.08.005

http://dx.doi.org/10.1109/TIT.1954.1057460

http://dx.doi.org/10.1214/aoms/1177729694


http://dx.doi.org/10.1002/0471200611

http://dx.doi.org/10.1177/0272989X8800800208

http://dx.doi.org/10.1002/9780470317082