International Journal of Advanced Engineering and Management Research
Vol. 3 Issue 1; 2018

http://ijaemr.com/ ISSN: 2456-3676

TESTING FOR THE DIFFERENCE IN SENSITIVITIES OF BINARY CLASSIFIERS USING SURVIVAL ANALYSIS

Kathare Alfred

Department of Mathematics and Computer Sciences

University of Eldoret

Otieno Argwings

Department of Mathematics and Computer Sciences

University of Eldoret

Kimeli Victor

Department of Mathematics and Computer Sciences

University of Eldoret

Abstract

Sensitivity is an important measure of the performance of a binary classifier in a disease control program among populations at risk. Where one has to choose between binary classifiers, on basis of their sensitivities, a test grounded on the theory is important for viability of the test results. The commonly used procedure namely the area under the receiver operating characteristic curve does not only lack strong theoretical basis but also trades the sensitivity of a classifier with its specificity. Also, the use of relative sensitivity is limited to comparison at single cut-off point of the classifiers. In this study was we provide a procedure for testing for the difference in sensitivities of two or more binary classifiers over a set of cut-offs without reference to their specificities. By observing the cumulative sensitivities over ordered cut-offs, we defined the survival function of the diseased individuals. Low sensitivity results to high survivability. To test for the difference in sensitivities we tested for the difference of the survival curve using the log-rank test. We subjected our approach to two pancreatic cancer classifiers and the test results show that the two were statistically difference. Further studies should focus on survival curves that close at some point.

Key Words: Binary classifier, sensitivity, survival function, risk of failure, “diseased” subject, “non-diseased” subject.

1.0 Introduction

Predictive assessments are common in many spheres of life today. Evaluation and comparison of the accuracy of available different binary classifiers is therefore an essential integral part in decision making. The conventional approach to evaluate the performance of such classifiers is the use of sensitivity and specificity as measures of accuracy of test in comparison to the gold standard. Sensitivity is a measure of how well a binary predictive classifier correctly identifies a positive case while specificity measures how well a predictive classifier correctly identifies a negative case. In a situation where the test results are recorded over some ordered cut-offs of binary predictive classifier, the overall performance of the classifier is evaluated using the receiver operating characteristic (ROC) curve. The ROC curve is a graph of sensitivity against 1-specificity. In this case, the area under the curve (AUC) is used as a summative measure of the performance of the classifier. The importance of sensitivity and specificity may be significantly different in some programs such that the interest is on one and not the other. Faced with two or more predictive classifiers, one is bound to know how different their sensitivities or specificities are. In this study we present application of survival analysis that allows for the test of difference between sensitivities of two classifiers.

2.0 Methodology

2.1 Study Design

A simple randomized screening test design was used for the study. In this design, the study individuals are first classified by gold standard as “diseased” or “non-diseased”. These individuals are then randomly assigned for screening to one of the two or more binary predictive classifiers over a set of ordered cut-offs.

Let C denote the outcome of a binary predictive classifier and G be the outcome of the gold standard such that

$$C = \begin{cases} 1\, if \;\; the\; test\; result\; is \;\; positive \\ \\ 0\, if \;\; the\; test\; results\; is\; negative \end{cases} \quad \text{and } G = \begin{cases} 1\, if \;\; a\; subject\; is\; diseased \\ \\ 0\, if \;\; a\; subject\; is\; non-diseased \end{cases}$$

Further let $i = 1, 2, ..., k$ denote a particular value of the random variable X representing the cut-off point and $q = 1, 2, ..., s$ be the particular predictive classifier.

The quantity $n_{cg}^{i\,q}$ represents results of the gold standard classification and screening by the predictive classifier where c and g represents the binary classifier and gold standard respectively. Then, the ordered tables of these values for a fixed q form cumulative partial tables for given reference cells. Thus, for true positive we have $n_{11}^{k\,q} > \mathrm{n}_{11}^{(k-1)\,q} > \ldots > \mathrm{n}_{11}^{2\,q} > n_{11}^{1\,q}$. Essentially, this means that if a diseased case was correctly identified at cut-off X_i it will also be correctly identified at cut-off X_{i+1}. Similarly for false negative we have $n_{01}^{1q} > \mathrm{n}_{01}^{2q} > \ldots > \mathrm{n}_{01}^{(k-1)q} > n_{01}^{kq}$.

For false positive we have $n_{10}^{kq} > \mathrm{n}_{10}^{(k-1)q} > \ldots > \mathrm{n}_{10}^{2q} > n_{10}^{1q}$ and for true negative we have $n_{00}^{1q} > \mathrm{n}_{00}^{2q} > \ldots > \mathrm{n}_{00}^{(k-1)q} > n_{00}^{kq}$.

2.2 Estimation of Curve Curves

Now suppose n out of N study cases were confirmed to have the disease. Let n_{iq} denote the number of subjects correctly identified by the predictive classifier q; $q = 1, 2, \ldots, s$ as having the disease and let $n_{iq}^{/}$ denote the number of subjects incorrectly identified by the same predictive classifier as not having the disease at cut-off X_i: $i = 1, 2, \ldots, k$ and that $n_{iq} + n_{iq}^{/} = n$. Basically n_{iq} are the true positives and $n_{iq}^{/}$ are the false negatives. Instead of observing the cumulative function of n_{iq}, we then observe n_{xq}, the number of diseased individuals who are correctly identified as having the disease within the interval X_{i+1} and X_i. For the purpose of modelling the survivorship curve we refer n_{xq} as failures and $n_{iq}^{/}$ as survivors. The proportion of individuals surviving at cut-off x is then given by $p_x = \dfrac{n_{i+1,q}^{/}}{n_{i,q}^{/}}$.

We estimate the survival curve using Kaplan-Meier method. This method does not require the assumption of some underlying probability distribution. In this case the cumulative proportion surviving up to $X = k$ having survived $X = k-1$ is given by $S(X) = \prod_{i=1}^{i=k} p_i$.

2.3 Testing for the Difference in Sensitivities of Binary Classifiers

In order to test for the difference between sensitivities of two or more predictive classifiers, we used the long rank test. We hypothesized that there was no difference between the two survival curves. In this case $H_0 : S_1(x) = S_2(x)$.

The test statistic is given by $\chi^2(\log rank) = \sum \dfrac{\left(O_q = E_q\right)^2}{E_q}$ by where O_q is the total number of observed failures and E_q is the total number of expected failure in classifier q.

The expected number of events at cut-off x is the product of the risk of failure at x and the number of survivors at x. For $q = 1, 2$ risk of failure is given by $r_i = \dfrac{n_{1x} + n_{2x}}{n'_{n1i} + n'_{2i}}$. The expected number of failures for classifiers at x is given by $e_{iq} = r_i n'_{iq}$.

3.0 Application to Real Data

We compared sensitivities of two classifiers measuring the carbohydrate antigen 19-9 (CA 19-9). Elevated levels of CA 19-9 (> 37 U/mL) has been found to be associated with gastrointestinal carcinomas particularly in pancreatic cancer. We thus, bench marked our cut-off point at 40 U/mL and weighted cut-off above 40 U/mL nearly twice to spread the possibility of cancer detection. Our cut-offs thus ranged between X = 110 and $x = 0$ at arbitrarily interval of 10 U/mL. Figure 1 shows the survival curves for the two classifiers.

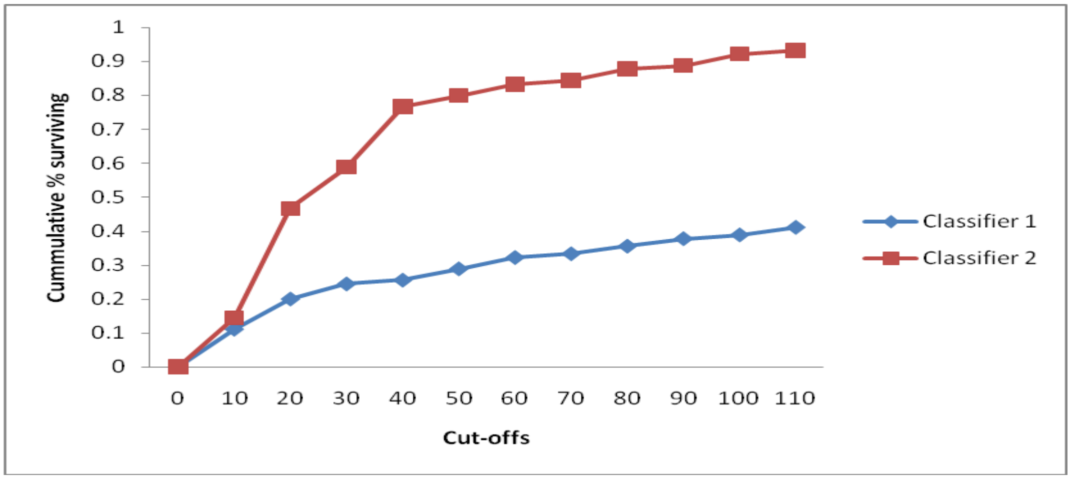

Figure 1: Survival curves for classifier 1 and 2

At cut-off X > 110, less than a half (41%) of the diseased were identified by classifier 1 while classifier 2 identified over 93% of them. By widening the possibility of detection to X > 0 both classifiers were able to detect all the diseased. The results show that classifier 2 has a higher survivorship rate compared to classifier 1. The greater the survivorship, the lower the ability of a classifier to correctly identify the "diseased" individuals as "diseased".

Further, we tested the null hypothesis that the two classifiers have the same sensitivity. In this case the two survival curves are the same: $H_O : S_1(x) = S_2(x)$. The $\chi^2 = 11.98\ (p < 0.05)$. In this case, we conclude that classifier 1 has higher sensitivity compared to classifier 2.

4.0 Discussions

In survival analysis, the study subjects with a defined state are followed over non-negative real time variable. In this case, the outcome variable is the time until the occurrence of an event of interest. The event of interest would be death, occurrence of disease, divorce, marriage etc.

By adopting a simple screening design, and observing diseased individuals over a set of ordered non-negative cut-offs until they are correctly classified, we mapped sensitivity into survival curve. The approach provides possibilities of testing for the difference in sensitivities of two or more binary classifies without trading with their specificities and vice versa. Our test results for the two pancreatic cancer classifiers showed that classifier 1 has a significantly higher sensitivity than classifier 2.

REFERENCE

Adams, N., M., & Hand, D., J., (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition, 32*, 1139–1147.

Akobeng, A., K., (2006). Understating diagnostic tests 1: Sensitivity, specificity and predictive values. *Act Pediatric*, 96, 338-341.doi:10.1111/j.1651-2227. 2006. 00180. x.

Alonzo, T., A., & Kittelson, J., M., (2006). A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible.

International Biometric Society, 62(2), 605–612. doi:0.1111/j.15410420.5.00445.x.

Bradley, P., A., (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *The Pennsylvania State University, 30*(7), 1145--1159. doi:

10.1016/ S0031-3203(96)00142-2. x.

Cheng, H., & Macaluso, M., (1996). Comparison of the accuracy of two tests with a

confirmatory procedure limited to positive results. *Epidemiology, 8*, 104–106.

Griner, P., Mayewski, R., J., Mushlin, A. I., & Greenlan, P.,(1981). Selection and interpretation of diagnostic tests and procedures: Principles and applications. *Annals of Internal Medicine*, *94*, 557–592.

Halligan, S., Douglas, G., A., & Mallet, S., (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A

discussion and proposal for an alternative approach. *European Radiology,* 25(4), 932–939. doi: 10.1007/s00330-014-3487-0. x.

Kumar, R., & Indrayan, A., (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics, 48*(4), 277–287.

Lobo, J., M., Jiménez-Valverde, A., & Real, R., (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. doi: 10.1111/j.1466-8238.2007.00358. x.

Pepe, M., S.,(2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction* (1st ed.). United Kingdom: Oxford University Press.

Pepe, M.,S., (2000) Receiver operating characteristic methodology. *Journal of the*

American Statistical Association, 95, 308–311.

Walter, S., D., (2005). The partial area under the summary ROC curve. *Statistical Medicine,* 24(13), 2025–2540. doi: 10.1002/sim.2103. x.

Zweig, M., H., & Campbell, G., (1993). Receiver-operating characteristic (ROC) plots: