Kernel Smoothing For ROC Curve And Estimation For Thyroid Stimulating Hormone

Tazhibi Mehdi, Bashardoost N & Ahmadi M*

Department of Biostatistics and Epidemiology, School of Health, Isfahan University of Medical Science.

**For reprint and all correspondence: Tazhibi Mehdi, Biostatistics and Epidemiology Department, School of Health, Isfahan University of Medical Science.*
Email: tazhibi@hlth.mui.ac.ir

ABSTRACT

Receiver Operating Characteristic (ROC) Curves are frequently used in biomedical informatics research to evaluate classification and prediction models to support decision, diagnosis, and prognosis. ROC analysis investigates the accuracy of models and has ability to separate positive from negative cases. It is especially useful in evaluating predictive models and compare to other tests which produce output values in a continuous range. Empirical ROC curve is jagged but a true ROC curve is smooth. For this purpose kernel smoothing were used. The Area Under ROC Curve (AUC) frequently is used as a measure of the effectiveness of diagnostic markers. In this study we compare estimation of this area based on normal assumptions and kernel smoothing. This study used measurements of TSH from patients and non-diseased people of congenital hypothyroidism screening in Isfahan province. Using the method, TSH ROC curves from Isfahani's infants were fitted. For evaluating of accuracy of this test, AUC and its standard error calculated. Also effectiveness of the kernel methods in comparison to other methods showed.

Keyword: Kernel Smoothing-ROC Curve-Thyroid Stimulating Hormone

INTRODUCTION

A receiver operating characteristics (ROC) curves is a technique for visualizing, organizing and selecting classifiers based on their performance. ROC curves have long been used in signal detection theory to depict the trade-off between hit rates and false alarm rates of classifiers (Egan, 1975; Swets et al., 2000).

ROC Curves are frequently used in biomedical informatics research to evaluate classification and prediction models to support decision, diagnosis, and prognosis.

ROC analysis investigates the accuracy of models ability to separate positive from negative cases. It is especially useful in evaluating predictive models or other tests that produce output values over a continuous range. A basic classification tool in medicine is the binary test, which yields two discrete results to infer as unknown. The accuracy of these tests is commonly assessed using measures of sensitivity and specificity.

The medical decision making community has an extensive literature on the use of ROC curves for diagnostic testing (Zou, 2002) Swets et al. (2000) brought ROC curves to the attention of the wider public with their Scientific American article.

For continuous and ordinal tests, there is no particular value of sensitivity or specificity that characterizes the overall accuracy of the test, but rather an entire range of values that very depending on what we use as the threshold for discrete the test results. Analysis investigates the accuracy of models ability to separate positive from negative cases. It is especially useful in evaluating predictive models or other tests that produce output values over a continuous range. A basic classification tool in medicine is the binary test, which yields two discrete results to infer as unknown. The accuracy of these tests is commonly assessed using measures of sensitivity and specificity. Therefore for continuous and ordinal tests, there is no particular value of sensitivity or specificity that characterizes the overall accuracy of the test, but rather an entire range of values that very depending on what we use as the threshold for discrete test results. The ROC captures in a single graph is the trade-off between a test sensitivity and specificity over this entire range. Area under ROC curve is the most commonly used index. Empirical ROC curve is jagged but a true ROC curve is smooth. Kernel smoothing was used for this target.

Object

Empirical ROC curve is jagged but a true ROC curve is smooth. Using Kernel smoothing method to smooth the jagged ROC curve.

Definition

ROC analysis investigates and employs the relationship between sensitivity and specificity of a binary classifier.

Sensitivity or true positive rate measures the proportion of positives correctly classified or

$\overline{\quad=\qquad\qquad\qquad}$ and specificity or true negative rate measures the proportion of negatives correctly classified or $\overline{\quad=\qquad\qquad\qquad}$.

Conventionally, Sensitivity is plotted against the false positive rate, which is one minus specificity.

If a classifier outputs a score proportional to its belief that an instance belongs to the positive class, decreasing the decision threshold – above which an instance is deemed to belong to the positive class – will increase both true and false positive rates. Varying the decision threshold from its maximal to its minimal value results in a piecewise linear curve from (0; 0) to (1; 1), such that each segment has a non-negative slope .This ROC curve is the main tool used in ROC analysis.

It can be used to address a range of problems, including: (1) determining a decision threshold that minimizes error rate or misclassification cost under given class and cost distributions; (2) identifying regions where one classifier out performs another; (3) identifying regions where a classifier performs worse than chance; and (4) obtaining calibrated estimates of the class posterior.

METHODS

ROC Analysis:

ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Swets, 1988). The medical decision making community has an extensive literature on the use of ROC curves for diagnostic testing (Zou, 2002).

ROC methodology is appropriate in situations where there are 2 possible "truth states" (i.e., diseased/normal, event/non-event, or some other binary outcome), "truth" is known for each case, and "truth" is determined independently of the diagnostic tests / predictor variables / etc. under study.

Rating data vs Continuous data

The term "rating data" is used to describe data based on an ordinal scale. For example, it is common in radiology studies to use a 5-point scale such as 1=disease definitely absent, 2=disease probably absent, 3=disease possibly present, 4=disease probably present, 5=disease definitely present. "Continuous data" refers to either truly continuous measurements or "percent confidence" scores (0-100).

Interpreting the Area Under the ROC Curve (AUC)

The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. It can take values from 0.0 to 1.0. The AUC can be interpreted as the probability that a randomly selected diseased case (or "event") will be regarded with greater suspicion (in terms of its rating or continuous measurement) than a randomly selected non-diseased case (or "non-event"). So, for example, in a study involving rating data, an AUC of 0.84 implies that there is an 84% likelihood that a randomly selected diseased case will receive a more-suspicious (higher) rating than a randomly selected non-diseased case. Note that an AUC of 0.50 means that the diagnostic accuracy in question is equivalent to that which would be obtained by flipping a coin (i.e., random chance). It is possible but not common to run into AUCs less than 0.50. It is often informative to report a 95% confidence interval for a single AUC in order to determine whether the lower endpoint is > 0.50 (i.e., whether the diagnostic accuracy in question is, with some certainty, any better than random chance).

Designing an ROC study: Which scale to use?

While ordinal (1-5) rating scales are probably the most widely used in radiology studies, there are advantages to using "percent confidence" (0-100) scales. (Of course, if you are dealing with a continuous measurement, you don't have to worry about which scale to use.) For continuous data, nonparametric methods are quite reasonable. With rating data, parametric methods are recommended, as nonparametric methods will be biased (i.e., tend to underestimate the true AUC). The standard error of the estimated area under the ROC curve is smaller using a continuous scale.

Parametric vs Nonparametric methodology

"Parametric" methodology refers to inference (MLEs) based on the bivariate normal distribution (i.e., this estimate assumes one normal distribution for cases with the disease and one normal distribution for cases without, or that the data has been monotonically transformed to normal). When this assumption is true, the MLE is unbiased.

"Nonparametric" refers to inference based on the trapezoidal rule (which is equal to the Wilcoxon estimate of the area under the ROC curve, which in turn is equal to the "c"-statistic in SAS PROC LOGISTIC output). Nonparametric estimates of the area under the ROC curve (AUC) tend to underestimate the "smooth curve" area (i.e., parametric estimates), but this bias is negligible for continuous data.

The Area Under an ROC Curve

By definition from (http://gim.unmc.edu/dxtests/roc3.htm) we have the most important statistic associated with ROC curves is the Area Under ROC Curve or AUC. Since the curve is located in the unit square, we have 0 _ AUC _ 1.

AUC = 1 is achieved if the classifier scores every positive higher than every negative; AUC = 0 is achieved if every negative is scored higher than every positive. AUC =1=2 is obtained in a range of different scenarios, including: (i) the classifier assigns the same score to all test examples, whether positive or negative, and thus the ROC curve is the ascending diagonal; (ii) the per-class score distributions are similar, which results in an ROC curve close (but not identical) to the ascending diagonal; and (iii) the classifier gives half of a particular class the highest scores, and the other half the lowest scores. Notice that, although a classifier with AUC close to 1/2 is often said to perform randomly, there is nothing random in the third classifier: rather, its excellent performance on some of the examples is counterbalanced by its very poor performance on some others.

The graph at right shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of .5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

Univariate kernel density estimator:
Given a random sample X1; : : : ;Xn with a continuous, univariate
density f. The kernel density estimator is

$$(\ \ ,h) = \frac{1}{h} \qquad (\frac{-}{h})$$

with kernel K and bandwidth h. Under mild conditions (*h* must decrease with increasing *n*) the kernel estimate converges in probability to the true density.

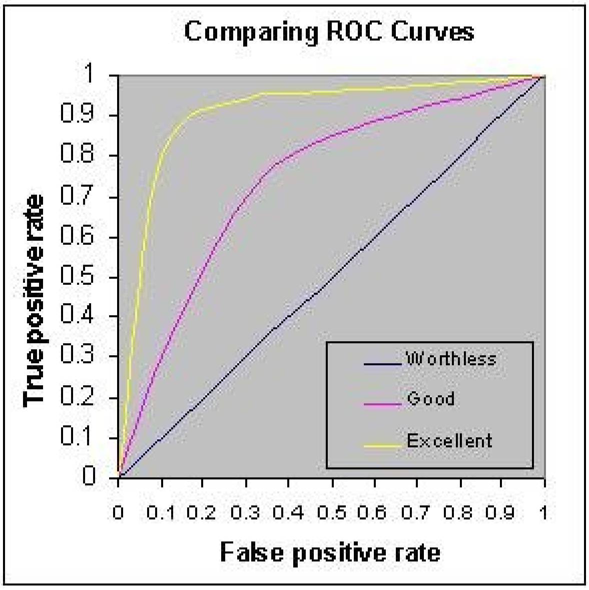
Let we have a sample of *N* which *m* of them are diseased and *N-m* are non diseased. Also let $(\ \ = 1, \ldots ,\ \ )$ for diseased people and $(\ \ = 1, \ldots ,\ \ )$ for non diseased people. Let F(0) and G(0) are cumulative functions of X and Y respectively and *p* is the false positive rate then we have

$$(\ ) = 1 - \quad (\quad (1 - \quad)) \ .$$

Now kernel estimator of *F(x)* and *G(y)* with kernel function *K(t)* are found. Since the ROC curve is gagged by use of Faraggi et al (2002) and kernel estimation this jagged curve become smooth.

RESULTS

Using this method, ROC curves of TSH from Isfahan's Infant were fitted. AUC and standard error are calculated. For evaluating of accuracy of this test by the Kernel methods AUC and SE were 0.843 and 0.02 and by the empirical methods AUC and SE were 0.847 and 0.017 respectively .Optimum cut off point was equal to 7.7 with 76 percent sensitivity and 81% specificity.



DISCUSSION

Faraggi and Reiser performed Monte Carlo simulation in large variety of different distribution for AUC, compared in terms of bias and root mean square error. They found that transform of variable to normality usually are preferred except for bimodal cases where Kernel methods can be effective and empirical methods as a robust method for continues data when diseased and healthy population sizes are at least 20.

REFERENCES

1. Egan, J.P., 1975. Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York.
2. Faraggi D, Reiser, B., 2002 Estimation of the area under the ROC curve, Statistics in medicine 21, 3093-3106 Flach, Peter ROC Analysis.
3. University of Bristol An entry to appear in the forthcoming Encyclopedia of Machine Learning (Springer) http://gim.unmc.edu/dxtests/roc3.htm.

4. Swets, J., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

5. Zou, K.H., 2002. Receiver operating characteristic (ROC) literature research. On-line bibliography available from: http://splweb.bwh. harvard.edu:8000/pages/ppl/zou/roc.html.