Published in final edited form as:

Stat Biosci. 2017 June ; 9(1): 91–104. doi:10.1007/s12561-016-9159-7.

ROC Curve Analysis in the Presence of Imperfect Reference Standards

Peizhou Liao,

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, Tel.: +1(404)747-8400, pliao3@emory.edu

Hao Wu, and

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, Tel.: +1(404)727-8633, Fax: +1(404)727-1370, hao.wu@emory.edu

Tianwei Yu

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, Tel.: +1(404)727-7671, Fax: +1(404)727-1370, tyu8@emory.edu

Abstract

The receiver operating characteristic (ROC) curve is an important tool for the evaluation and comparison of predictive models when the outcome is binary. If the class membership of the outcomes are known, ROC can be constructed for a model, and the ROC with greater area under the curve (AUC) indicates better performance. However in practice, imperfect reference standards often exist, in which class membership of every data point are not fully determined. This situation is especially prevalent in high-throughput biomedical data because obtaining perfect reference standards for all data points is either too costly or technically impractical. To construct ROC curves for these data, the common practice is to either ignore the uncertainties in references, or remove data points with high uncertainties. Such approaches may cause bias to the ROC curves and generate misleading results in method evaluation.

Here we present a framework to incorporate membership uncertainties into the construction of ROC curve, termed the expected ROC or “eROC” curve. We develop an efficient procedure for the estimation of eROC curve. The advantages of using eROC are demonstrated using simulated and real data.

Keywords

ROC curve; High-throughput Data; Imperfect Reference Standards

1 INTRODUCTION

An active field in biomedical and statistical research is to construct binary classifier (predictive model for binary outcomes), for example, to look for biomarkers for predicting

the presence/absence of certain diseases. To evaluate the predictive performance of different types of biomarkers or predictive models, a common practice is to compute some measurements to represent the classification accuracies, including true positive rate (TPR) and false positive rate (FPR). Receiver Operating Characteristic (ROC) curve [9] is a popular tool to comprehensively evaluate a classifier, which plots the TPR versus FPR under various thresholds for predictive results. Classifier with ROC having higher area under the curve (AUC) is usually more desirable.

In order to construct an ROC curve, a necessary prerequisite is to fully determine the class membership for each instance in the data. The class membership are obtained from various mechanisms depending on the application. In some situations, the true status of each instance is known. For example, the disease status for a group of people are determined by established procedure with high accuracy. Another example in high-throughput biology is the spike-in experiment, where each instance being manipulated is known to the experimenter [5]. In many other situations, the reference standard fails to supply 100% definitive membership values for every individual instance. In literatures on diagnostic test, the best performing test or expert's opinion in establishing a patient's true disease status is referred to as the "gold standard" [12, 26], even though non-trivial uncertainties still exist within these results. In some cases, the reference standard would provide scores or class membership probabilities that reflect uncertainties. For example, some reference standard based on experts' experience or quality scores generated by machines actually indicate the probabilistic confidence for true class membership. Here we refer to the reference standard with certain degree of uncertainty by "*imperfect reference standard*". For all imperfect reference standards, the degree of uncertainty can be quantified by two factors: the proportion of instances having indefinite class membership, and confidence level of assigning each instance to its class. Basically, the more instances having indefinite membership and the lower the average confidence in class membership, the greater uncertainty exists for the imperfect reference standard. Note that imperfect reference standard is different from "verification biased" sample. For the imperfect reference standard, the class membership of all instances have been established but a fraction of them lack definitive class labels. In contrast, not all the instances in the verification biased sample have undergone an assessment of class membership, thus classification information is missing for those unverified instances [25].

Situations of imperfect reference standard are common in high-throughput data analyses. For example, to evaluate transcription factor (TF) binding site prediction algorithms, ChIP-seq (chromatin immunoprecipitation followed by deep sequencing) results are often used as the gold standard [15,20]. However, the ChIP-seq results depend on peak calling algorithms and the thresholds for defining peaks, while the confidence levels of the peaks are ignored during evaluation. Another example is the evaluation of peak detection algorithms in the processing of liquid chromatography-mass spectrometry (LC/MS) metabolomics data [10]. Because each peak detected is not tagged with a chemical identity, it is a common practice to match the mass-to-charge ratio (m/z) of peaks to known metabolome databases, in order to gain some knowledge whether detected peak is real signal or noise [23,24]. However, databases are incomplete, thus unmatched peaks could still be a true signal. On the other hand, peaks

matched to the database still have a small chance to be false due to high levels of experimental noise.

In the presence of imperfect reference standard, it is difficult to obtain unbiased estimates of ROC curve to evaluate the classifier accuracy. When the degree of uncertainty is low, people directly use the class membership in the imperfect reference standard as the truth to assess the classifiers. In this process, the instances with indefinite class membership would either be discarded, or their uncertainties ignored. However, dropping instances or ignoring uncertainties could lead to a loss of information about classification performance. When the degree of uncertainty is high, one common practice to deal with uncertainty is to dichotomize uncertainty measures at some cutoff point (e.g., median), so that imperfect reference standard would be converted into binary “gold standard” [1]. However, the estimation of classifier accuracy could be severely biased by dichotomizing uncertainty measures. Another way to take uncertainty into account is to use fuzzy theory to compute classifier accuracy [6], which is not statistically rigorous as it simply creates continuous scale class membership. Given the class label uncertainties can be different for different samples, a more statistical approach incorporating class probabilities can better reflect the performance of classifiers.

In this work, we aim to develop a rigorous statistical method to incorporate the uncertainties in reference standard in estimating the ROC curve. To include the full information on the indefinite instances in the evaluation of classifiers, we propose to compute the expected value of TPR and FPR, i.e., eTPR and eFPR. By integrating out the uncertainty carried by the indefinite instances, we can use eTPR and eFPR to assess the average performance of new classifiers. Furthermore, by varying the values of threshold, we are able to construct the eROC curve, which is a natural generalization of ROC curve given the class membership uncertainties. By incorporating the uncertainties of reference standard into the construction of ROC curves, the eROC approach we develop here will produce more unbiased and realistic estimates of the predictors' performance.

2 METHODS

2.1 Definition of the eROC Curve

Suppose the dataset consists of n independent instances (x_i, μ_i) for $i = 1, \dots, n$, where x_i is the score yielded by the classifier, and μ_i is the probabilistic confidence assigned by the imperfect reference standard,

$$\mu_i = \Pr(\text{instance } i \text{ belongs to the positive class}). \quad (1)$$

For instance, (1.23, 0.95) corresponds to an observation with assigned score 1.23 and the probability of this observation coming from the positive class is 0.95 according to the imperfect reference standard. In biomedical research, extreme scores usually provide sufficient evidence to identify the class membership of an instance; however, values between extremum, especially those which are either close to the threshold or near the center of the sample, can cause much uncertainty in establishing their class membership.

Denote S_i as the indicator variable of i -th instance belonging to the true positive population, then S_i follows a Bernoulli distribution, $S_i \sim \text{Bernoulli}(\mu_i)$. If no uncertainty exists in the class membership (i.e., $\mu_i = 0$ or 1), S_i has a degenerate distribution. Assuming that S_i 's are mutually independent and that S_i 's and x_i 's are independent conditional on μ_i for $i = 1, \dots, n$, at a cut-off point α , we define the expected true positive rate of the classifier as

$$\text{eTPR}_\alpha = E \left\{ \frac{\sum_{i=1}^n S_i I_{\{x_i \geq \alpha\}}}{\sum_{i=1}^n S_i} \middle| \sum_{i=1}^n S_i > 0 \right\} \quad (2)$$

It can be easily demonstrated that eTPR_α is monotonically decreasing with respect to α . Also, $\text{eTPR}_\alpha = 0$ when $\alpha > \max_{i=1, \dots, n} x_i$ and $\text{eTPR}_\alpha = 1$ when $\alpha \leq \min_{i=1, \dots, n} x_i$. In equation (2), the conditional expectation rules out the undefined case $0/0$, which can be shown to have little impact on computation under large sample size. The proposed eTPR_α is a natural estimate for TPR as the indefinite class membership is well modeled through indicator variables S_i 's. Additionally, eTPR_α provides an accurate estimate of the true TPR_α as long as $\mu_i \rightarrow 1(0)$ for true positive (negative) instances and the sample size is relatively large.

Let R_i 's be indicator variables of i -th instance belonging to the true negative population, similarly, we define the expected false positive rate and true discovery rate of the classifier as

$$\text{eFPR}_\alpha = E \left\{ \frac{\sum_{i=1}^n R_i I_{\{x_i \geq \alpha\}}}{\sum_{i=1}^n R_i} \middle| \sum_{i=1}^n R_i > 0 \right\} \quad (3)$$

$$\text{eTDR}_\alpha = E \left\{ \frac{\sum_{i=1}^n S_i I_{\{x_i \geq \alpha\}}}{\sum_{i=1}^n I_{\{x_i \geq \alpha\}}} \right\}$$

Notice that eTDR_α is equivalent to $\sum_{i=1}^n \mu_i I_{\{x_i \geq \alpha\}} / \sum_{i=1}^n I_{\{x_i \geq \alpha\}}$, which has the same form as TDR based on fuzzy theory [13].

The proposed eTPR_α and eFPR_α can be tailored to usual ROC-based analysis. Following the standard procedure to construct an ROC curve, we vary the value of threshold α from $-\infty$ to $+\infty$. The various α_i 's generate a sequence of $\{\text{eTPR}_{\alpha_i}\}$ and $\{\text{eFPR}_{\alpha_i}\}$, each pair of which represents a different discrete classifier in ROC space. By connecting the points, we obtain a step function that resembles ROC curve and will be referred to as an eROC curve. The eROC curve shows the averaged ability of the classifier to rank the positive instances relative to the negative instances given the uncertainties in the reference standard.

Furthermore, to compare different classifiers we can reduce the eROC performance to a single scalar value by calculating the area under the eROC curve (eAUC). In the presence of uncertainty, the eAUC no longer equals the probability of ranking a randomly chosen positive instance higher than a randomly chosen negative instance by the classifier. However,

eAUC measures how large, on average, such a probability of correctly ranking random instances from different classes would be.

2.2 Calculation of eTPR_α

The n Bernoulli variables S_j are independent but not necessarily identical, therefore,

$\sum_{i=1}^n S_i$ follows the Poisson binomial distribution. The calculation of eTPR_α is facilitated by the following lemma.

Lemma 1—Given that the sample size n is relatively large and $\tilde{n}/n \rightarrow c > 0$ ($\tilde{n}$ is the number of instances with nonzero μ), then

$$E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \mid \sum_{i=1}^n S_i > 0 \right\} \approx \mu_k E \left\{ \frac{1}{1 + \sum_{i \neq k}^n S_i} \right\} \quad (4)$$

Proof of the lemma:

$$\begin{aligned} E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \right\} &= E \left\{ E \left[\frac{S_k}{\sum_{i=1}^n S_i} \mid S_k \right] \right\} \\ &= E \left[\frac{S_k}{\sum_{i=1}^n S_i} \mid S_k = 1 \right] \Pr(S_k = 1) + \\ &\quad E \left[\frac{S_k}{\sum_{i=1}^n S_i} \mid S_k = 0 \right] \Pr(S_k = 0) \\ &= \mu_k E \left\{ \frac{1}{1 + \sum_{i \neq k}^n S_i} \right\} \end{aligned}$$

Note that

$$\begin{aligned} E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \right\} &= E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \mid \sum_{i=1}^n S_i > 0 \right\} \Pr \left(\sum_{i=1}^n S_i > 0 \right) + \\ &\quad E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \mid \sum_{i=1}^n S_i = 0 \right\} \Pr \left(\sum_{i=1}^n S_i = 0 \right) \\ &= E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \mid \sum_{i=1}^n S_i > 0 \right\} \Pr \left(\sum_{i=1}^n S_i > 0 \right) \\ &\approx E \left\{ \frac{S_k}{\sum_{i=1}^n S_i} \mid \sum_{i=1}^n S_i > 0 \right\} \end{aligned}$$

The approximation comes from the fact that

$$\Pr \left(\sum_{i=1}^n S_i > 0 \right) = 1 - \Pr \left\{ \sum_{i=1}^n S_i = 0 \right\} = 1 - \prod_{i=1}^n (1 - \mu_i) \geq 1 - (1 - \min_{\mu_i \neq 0} \mu_i)^{\tilde{n}} \approx 1 \text{ as } n \rightarrow \infty$$

We further define the Poisson binomial variable $T_k = \sum_{i \neq k}^n S_i$. Let $F_{T_k}(t)$ be the cumulative distribution function (CDF) of T_k , then

$$E \left\{ \frac{1}{1+T_k} \right\} = \sum_{t=1}^{n-1} \frac{1}{1+t} [F_{T_k}(t) - F_{T_k}(t-1)] + F_{T_k}(0) = \sum_{t=0}^{n-2} \frac{F_{T_k}(t)}{(t+1)(t+2)} + \frac{1}{n}$$

Define $E\{1/(1+T_k)\} \equiv v_k$, the calculation of $eTPR_\alpha$ proceeds as

$$eTPR_\alpha = E \left\{ \frac{\sum_{i=1}^n S_i I_{\{x_i \geq \alpha\}}}{\sum_{i=1}^n S_i} \middle| \sum_{i=1}^n S_i > 0 \right\} = \sum_{i=1}^n \left[E \left\{ \frac{S_i}{\sum_{i=1}^n S_i} \middle| \sum_{i=1}^n S_i > 0 \right\} I_{\{x_i \geq \alpha\}} \right] = \sum_{i=1}^n \mu_i v_i I_{\{x_i \geq \alpha\}}$$

(5)

Regarding the computing of the CDF of Poisson binomial distribution, many methods have been proposed in the literature including Le Cam's in-equality, normal approximation and recursive formula [3, 14, 18]. Hong (2013) utilizes the discrete Fourier transformation of the characteristic function to derive an exact formula with a closed-form expression for the CDF of Poisson binomial distribution [11]. The proposed algorithm in Hong (2013) has been implemented in the R package "poibin", which would be used in the calculation of $eTPR_\alpha$ when $n \leq 1000$. When $n > 1000$, we use the refined normal approximation method [18], which not only approximates the CDF well but also reduces much computational time.

We develop an R function `eROC`, which is freely available as part of the `ROCS` package available at CRAN [22].

3 RESULTS

3.1 Simulation Studies

We conduct extensive simulation studies to assess the performance of $eROC$ curves, and compare it with ROC curves using binary reference standards created by dichotomizing the probabilistic confidence of the imperfect reference standard (referred to as "dROC"). We consider a number of simulation scenarios under different sample sizes (n is set to be 500, 1,000, 5,000, and 10,000), and different percentages of positives (ratios of positive to negative classes are set to be 10:1, 1:1, 1:10, 1:50, and 1:100, as in genomic studies, usually true positives account for a very small percentage of the data).

Under each simulation scenario, we first generate a set of "true" reference scores for all instances, denoted by z_p from normal distributions with different parameters for positive and negative classes: $N(1, 2)$ for scores from positive class, and $N(-1, 2)$ for scores from negative class. The reference scores are the underlying, unobserved scores for each instance. It mimics the process from which an imperfect reference standard is generated through an experiment. Given z_p , the measurements of uncertainties (μ_j) are generated from a logistic regression model: $\text{logit}(\mu_j) = b_0 + b_1 * z_p$. We fix the logistic regression coefficients b_0 and

b_1 so that we have $\mu_i = 0.85$ when $z_i = 1$, and $\mu_i = 0.2$ when $z_i = -1$. Furthermore, we simulate the observed scores (x_i) from a normal distribution conditional on the reference scores, i.e., $x_i|z_i \sim N(1 + r(z_i - 1), 2(1 - r^2))$, where r is the correlation coefficient between reference and observed scores. We set $r = 0.8$ in all simulations, which reflects the positive correlation between the observed scores and the underlying reference scores.

Different ROC curves are constructed based on these simulated data. To be specific, eROC is computed based on x_i and μ_i using the proposed procedure. dROC is constructed by first dichotomizing μ_i 's using the cutoff 0.5 to obtain binary outcome, and then compute ROC accordingly. In addition, we also construct true ROC curve using x_i and true class labels, and use this as the benchmark to compare different ROC curves. ROC curve closer to the truth is deemed better. Figure 1 displays three ROC curves from one set of simulations with a sample size of 1,000 and under different percentages of positives. They show that eROC is much closer to true ROC compared with dROC, regardless of the percentage of positives. It also shows that the dROC consistently over-estimates the truth, which provides an unrealistically optimistic result.

To further systematically compare the ROC estimates, we repeat the data simulation and ROC construction steps for 100 times under each sample size/class ratio setting. For each simulation, we restrict the true AUC from using the reference scores to be greater than 0.8 so that the reference scores has good accuracy in prediction. Simulated data with AUC less than 0.8 from using reference scores are discarded and regenerated. We first evaluate the performance of dROC and eROC based on the biases of AUC from the estimated ROC curves. The box plots of the bias of estimated AUCs from 100 simulations are displayed in Figure 2, where the upper panel shows plots for results from smaller sample size (1, 000), and the lower panel are for larger sample size (10, 000). Columns represent different class ratios (percentages of positives among all instances). The same results for other sample sizes ($n = 500$ and $5, 000$) are provided in Supplemental Figure S1 and S2. These results show that dROC consistently overestimates AUC and the overestimation bias can be as large as 0.09 on average. This significant overestimation is the result of the strong correlation between observed scores and reference scores as well as the fact that uncertainties are generated from a monotone transformation of reference scores. Also, the underlying reference scores are simulated from two partially overlapping densities. When a sample that belongs to the positive class has a very small reference score, the dROC approach will falsely consider the sample as belonging to the negative class given the observed score is strongly correlated with the reference score. Hence, the AUC value of the dROC approach becomes overly optimal. In contrast, for $n = 10, 000$, biases from eROC are always smaller than that from dROC across all scenarios. For $n = 1, 000$, eROC has significantly smaller biases than that from dROC, except for ratios 1:50 and 1:100 at which the two methods have comparable biases. In some cases, the average biases in AUC from eROC are very close to 0. These results indicate favorable performances of eROC compared with dROC. It is also shown that when the class membership are very unbalanced, the biases from eROC seems to be slightly greater than those from cases with less imbalanced class membership.

We further compare the two methods in terms of FDR-Controlled Area Under the Curve (FCAUC), which is defined as the area under the ROC curve corresponding to a

predetermined FDR threshold [21]. In many applications of class-skewed data, people often focus on the region of the ROC curve with low FDR. For example, in differential expression (DE) analysis, the proportion of true DE genes is usually low ($\sim 5\%$). In evaluating methods for DE detection, the starting portion of the ROC curve (corresponding to the top ranked genes) is more important than the whole curve. Figure 3 shows the box plots of biases in estimating FCAUC at the FDR level of 0.2. As shown in the figure, dROC severely overestimates the FCAUC because dichotomization at 0.5 mistakenly increases the number of true positives, leading to more FDR values falling below 0.2. On the other hand, eROC provides results with much lower biases, compared with dROC. It also shows the pattern that higher percentage of negatives leads to more severe overestimation of FCAUC. This is explained by the fact that more variability exists in the estimation of FDR values when there is only a small proportion of positives.

The eROC approach assumes that the probabilistic confidence μ_j 's are free of measurement errors. However, in practice, it is not uncommon that some uncertainty could exist in μ_j 's. To assess the impact of the uncertainty in μ_j 's, we performed the following sensitivity analysis:

1. Apply the logit transformation on μ_j i.e., $m_j = \log(\mu_j/(1 - \mu_j))$
2. Add measurement error to m_j by $m_i^* = m_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$
3. Obtain perturbed probabilistic confidence $\mu_i^* = e^{m_i^*} / (1 + e^{m_i^*})$
4. Compare eROC curves based on (x_j, μ_j) with that based on (x_j, μ_i^*)

We adjust the parameter σ to examine whether the calculation of eROC would be severely affected by a certain amount of measurement errors.

To carry out sensitivity analysis, we consider the scenarios where the sample size is 1,000 and the class ratios are 1:10 and 1:50. We vary the amount of measurement error by using different σ . In the presence of mismeasured probabilistic confidence μ_j 's, we examine the difference in AUCs (i.e., AUCs based on (x_j, μ_j) and (x_j, μ_i^*)) from the estimated ROC curves. Figure 4 shows the box plots of difference in AUCs. For all noise level considered, the median difference in AUC is almost zero from both methods, but larger measurement errors lead to ROC curves based on μ_i^* 's significantly deviating from that based on μ_j 's, as expected. However, eROC has much smaller variance in such differences, indicating that eROC is less sensitive to measurement errors in μ_j 's compared to dROC.

Overall, the simulation results demonstrate that in the presence of imperfect reference standards, the proposed eROC method provides more realistic results than dROC. The improvement is very consistent and robust under various sample sizes and class ratio settings. The procedure for calculating eROC is computationally efficient. For example, for 1,000 observations, it takes 2 seconds and 128 MB memory on a single thread of an Intel Xeon X5650 machine with 2.67GHz to construct the eROC. The computation increases linearly with the number of observations. For larger studies such as gene expression analysis, the construction of eROC can be finished in less than 1 minute.

3.2 Real Data Analysis

We demonstrate the performance of eROC in a set of real data to compare methods for detecting differentially expressed genes (DEG) from microarray and RNA-seq experiments. The data are generated by the MicroArray Quality Control (MAQC) project, which is initiated by FDA to evaluate different technologies for measuring gene expressions [17]. The third phase of the project is also known as the Sequencing Quality Control (SEQC) project, which adds RNA-seq technology to the comparison. In all the experiments, the gene expressions of two biological samples: human brain and human universal reference RNA were profiled. Different technologies were applied, including two types of gene expression microarrays (Affymetrix Human Genome U133A and U133Plus2), and RNA-seq from Illumina's Genome Analyzer II high-throughput sequencing system. The data were obtained from Gene Expression Omnibus (GEO) database, under accession number GSE23906 for microarray and GSE30250 for RNA-seq.

In addition, as part of the MAQC Project, low-throughput measurements of gene expressions based on quantitative real-time polymerase chain reaction (qRT-PCR) data are available for 992 genes from the same samples using Taqman technology. These data are often used to construct reference standard for comparing the DE detection methods [4, 19]. In these comparisons, the fold changes of the Taqman expression measurements from two groups are first computed, and then the DE status for these genes are determined by using arbitrary thresholds, for example, absolute fold changes greater than 2 is deemed DE, less than 0.2 is deemed non-DE, and others are no-call [4, 19].

We use this data set to demonstrate the performance of eROC procedure, and compare that with dROC which uses reference standard from arbitrary thresholding. To construct the eROC curve, we first derived the uncertainty in class labels using the local false discovery rate (FDR) approach [7]. The log fold changes of all genes measured by qRT-PCR are assumed to follow a two-component mixture model. The local FDR estimates the posterior probability that a gene belongs to the positive group given the observed statistic of all genes. We first computed the log fold changes of all genes, and then fitted a two-component mixture model and obtained the posterior probabilities of being DE for each gene. The posterior probabilities were used as uncertainties for the reference standards (μ_j) to construct the eROC curves. We applied `limma` package to call DE from the data from two microarray platforms, and applied three different RNA-seq DE detection methods: edgeR [16], DESeq [2] and DSS [19]. The genes were ranked based on the reported p-values, and the eROC and dROC curves were constructed according to the rank.

Figure 5(A) shows the eROC curves from five different lists of DE calling results. Overall, the eROC curves from RNA-seq data are higher than those from the microarray data when the eFPR is low, demonstrating the RNA-seq provides better DE results for the top ranked genes. Comparing the eROC curves from two microarray platforms, the U133 plus 2 is slightly better than U133A, which is as expected since U133 plus 2 is a newer technology. Among the three RNA-seq DE detection method, it seems DSS and edgeR are comparable, and both are better than DESeq. In contrast, Figure 5(B) shows the dROC curves from reference standard based on thresholding the fold changes. The no-call genes (absolute log

fold change between 0.2 and 2) are discarded. From this set of curves, the RNA-seq data don't show improvement from microarray technologies. In fact, the two dROC curves from microarray data are slightly better for the top ranked genes. This is contradictory to the belief that RNA-seq provides data with better quality than microarray. It also shows that the AUCs from the dROC curves are considerably higher than the eAUCs from eROC curves, which is overly optimistic. Overall, comparing with dROC, the eROC provides more objective evaluation of the data and methods for DE detection.

4 DISCUSSION

In this work, we develop a novel statistical method to evaluate the classifier accuracy in the presence of imperfect reference standard. With the wide application of high-throughput technology in genome-wide studies, such data have become increasingly prevalent. This is because the true gold standards are difficult to obtain due to technological limitation or economical consideration, and many reference standards are obtained by experiments, rather than from biological truth. As we demonstrated in the simulation study, simply dichotomizing the measurement from experimental results to generate class labels (dROC approach) can cause substantial bias in the estimation of classifier performance, often to the positive side when the new method being assessed yields data that positively correlate with the reference standard data.

To overcome such problem, we propose a new method termed "eROC", which extends the traditional ROC method with the consideration of uncertainties of the reference standard. We use binary variables to incorporate uncertainties, and calculate the expected classification accuracy. Our proposed procedure to construct eROC is computationally efficient by taking advantage of the existing package that computes the CDF of poisson binomial distribution. In the presence of mismeasured probabilistic confidence, we suggest to perform sensitivity analysis outlined in the Results section. We show by simulation and real data analysis that eROC provides more accurate and objective measurement for the performance of a classifier. For data with imperfect reference standard, we envision that the eROC method we present in this work will potentially be very helpful.

In order to use eROC, an objective measure of the class label uncertainty is necessary. In some cases, the uncertainties are provided by either the experimental instruments or expert's opinions. When the uncertainties are unavailable, they sometimes can be estimated through a mixture model as long as there are some type of scores available. For example, in the gene expression analysis we presented in the real data analysis, mixture model is applied on the qRT-PCR measurements which gives more uncertainty to genes with scores near the class boarder. However, the mixture modeling procedure requires reasonably large number of samples to provide stable results. For data with limited sample size, some heuristic estimates of the reference standard uncertainty may be necessary in order to use eROC.

The R code for computing eROC is a component of the ROCS package, which is freely available from The Comprehensive R Archive Network (CRAN). The R function includes an option to construct the bootstrap percentile confidence interval for the AUC of eROC [8]. The bootstrap estimate of the standard error can be obtained accordingly. For future work,

we plan to incorporate the eROC approach into the ROC surface (ROCS) and the calculation of the volume under the surface (VUS) for class-skewed high-throughput data [21]. Also, we plan to develop normal-approximation based confidence intervals for the AUC of eROC as our numerical studies show that the bootstrap estimates appear to be asymptotically normally distributed. In addition, the indicator for being positive S_j 's are assumed to be mutually independent. In many cases such as in genome-wide association studies (GWAS), S_j 's could be correlated. To account for such correlation and estimate eROC is our future research plan.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

TY is partially supported by NIH grants P20HL113451 and U19AI090023. HW is partially supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454.

References

1. Albert, Paul S., Liu, Aiyi, Nansel, Tonja. Efficient logistic regression designs under an imperfect population identifier. *Biometrics*. 2014 Mar; 70(1):175–184. [PubMed: 24261471]
2. Anders, Simon, Huber, Wolfgang. Differential expression analysis for sequence count data. *Genome Biology*. 2010; 11(10):R106. [PubMed: 20979621]
3. Barlow, Richard E., Heidtmann, Klaus D. Computing k-out-of-n system reliability. *IEEE Transactions on Reliability*. 1984; 33:322–323.
4. Bullard, James H., Purdom, Elizabeth, Hansen, Kasper D., Dudoit, Sandrine. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*. 2010; 11(1):94. [PubMed: 20167110]
5. Choe, Sung E., Boutros, Michael, Michelson, Alan M., Church, George M., Halfon, Marc S. Preferred analysis methods for affymetrix genechips revealed by a wholly defined preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*. 2005; 6(2):R16. [PubMed: 15693945]
6. DeLeo, James M., Campbell, Gregory. The fuzzy receiver operating characteristic function and medical decisions with uncertainty; Proceedings of the First International Symposium on Uncertainty Modeling and Analysis; 1990. p. 694-699.
7. Efron, Bradley. Local false discovery rates. Division of Biostatistics, Stanford University; 2005.
8. Efron, Bradley, Tibshirani, Robert J. An Introduction to the Bootstrap. Chapman and Hall; 1993.
9. Fawcett, Tom. An introduction to roc analysis. *Pattern Recognition Letters*. 2006; 27:861–874.
10. Fiehn, Oliver. Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*. 2002; 48(1–2):155–171. [PubMed: 11860207]
11. Hong, Yili. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*. 2013; 59:41–51.
12. Hui, Siu L., Zhou, Xiaohua. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*. 1998; 7:354–370. [PubMed: 9871952]
13. Kandel, A. Fuzzy mathematical techniques with applications. Menlo Park, California: Addison-Wesley; 1986.
14. Le Cam, Lucien. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*. 1960; 10(4):1181–1197.

15. Pique-Regi, Roger, Degner, Jacob F., Pai, Athma A., Gaffney, Daniel J., Gilad, Yoav, Pritchard, Jonathan K. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Research*. 2011; 21(3):447–455. [PubMed: 21106904]
16. Robinson, Mark D., McCarthy, Davis J., Smyth, Gordon K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26(1):139–140. [PubMed: 19910308]
17. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*. 2006; 24(9):1151–1161.
18. Volkova, A Yu. A refinement of the central limit theorem for sums of independent random indicators. *Theory of Probability & Its Applications*. 1996; 40:791–794.
19. Wu, Hao, Wang, Chi, Wu, Zhijin. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*. 2013; 14(2):232–243. [PubMed: 23001152]
20. Xu, Tianlei, Li, Ben, Zhao, Meng, Szulwach, Keith E., Craig Street, R., Lin, Li, Yao, Bing, Zhang, Feiran, Jin, Peng, Wu, Hao, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Research*. 2015:gkv151.
21. Yu, Tianwei. ROCS: receiver operating characteristic surface for class-skewed high-throughput data. *PLoS ONE*. 2012; 7(7):e40598. [PubMed: 22792381]
22. Yu, Tianwei. ROCS: receiver operating characteristic surface. 2015 <http://CRAN.R-project.org/packagename=ROCS>.
23. Yu, Tianwei, Park, Youngja, Johnson, Jennifer M., Jones, Dean P. apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics*. 2009 May; 25(15):1930–1936. [PubMed: 19414529]
24. Yu, Tianwei, Park, Youngja, Li, Shuzhao, Jones, Dean P. Hybrid feature detection and information accumulation using high-resolution lc-ms metabolomics data. *Journal of Proteome Research*. 2013; 12(3):1419–1427. [PubMed: 23362826]
25. Zhou, Xiaohua. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research*. 1998; 7:337–353. [PubMed: 9871951]
26. Zhou, Xiaohua, Castelluccio, Pete, Zhou, Chuan. Nonparametric estimation of roc curves in the absence of a gold standard. *Biometrics*. 2005; 61(2):600–609. [PubMed: 16011710]

Fig. 1.
Comparison of ROC curve estimates from simulation with 1,000 instances.

Fig. 2.

Comparison of the biases of AUC from different methods for ROC curve estimation. SS is the sample size. Ratio represents the class skewness, i.e., the number of positive instances over the number of negative instances. The horizontal dashed line is at 0.

Fig. 3. Comparison of the biases of FCAUC. Figure setting is the same as in Figure 2.

Fig. 4.

Comparison of the difference in AUC based on μ_i 's and μ_i^* 's from sensitivity analysis.

Figure setting is the same as in Figure 2 with 1,000 instances. The horizontal dashed line at 0 corresponds to $\sigma = 0$.

Fig. 5.
eROC and dROC curves from a set of gene expression data.