Communications for Statistical Applications and Methods
2018, Vol. 25, No. 3, 307–319

https://doi.org/10.29220/CSAM.2018.25.3.307
Print ISSN 2287-7843 / Online ISSN 2383-4757

Applying a modified AUC to gene ranking

Wenbao Yu[a], Yuan-Chin Ivan Chang[b], Eunsik Park[1,a]

[a]Department of Statistics, Chonnam National University, Korea;
[b]Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

High-throughput technologies enable the simultaneous evaluation of thousands of genes that could discriminate different subclasses of complex diseases. Ranking genes according to differential expression is an important screening step for follow-up analysis. Many statistical measures have been proposed for this purpose. A good ranked list should provide a stable rank (at least for top-ranked gene), and the top ranked genes should have a high power in differentiating different disease status. However, there is a lack of emphasis in the literature on ranking genes based on these two criteria simultaneously. To achieve the above two criteria simultaneously, we proposed to apply a previously reported metric, the modified area under the receiver operating characteristic cure, to gene ranking. The proposed ranking method is found to be promising in leading to a stable ranking list and good prediction performances of top ranked genes. The findings are illustrated through studies on both synthesized data and real microarray gene expression data. The proposed method is recommended for ranking genes or other biomarkers for high-dimensional omics studies.

Keywords: gene ranking, Modified AUC ROC curve

1. Introduction

High-throughput studies simultaneously provide a measurement of thousands of genes; however, many of them are not important for a specific study. Hence, how to select informative ones or to provide a rank of them for further studies is an important problem in many biomedical studies. Along with detecting differentially expressed genes (DEG), a major interest in many genetic studies (Benjamini and Hochberg, 1995; Storey, 2003), ranking genes based on some relevant metric was also helpful to form relevant and integrated gene candidates for further analysis (Noma and Matsui, 2013). Many ranking methods have been proposed in the literature, these methods range from simple methods such as fold change, classical t-statistics, ad hoc modification of ordinary t-statistics (Tusher *et al.*, 2001) and Efron's 90% rule (Efron *et al.*, 2001) to complex hierarchical Bayes model-based approaches (Newton *et al.*, 2004), such as the moderated t-statistics (Smyth, 2004; Noma *et al.*, 2010; Noma and Matsui, 2013). There are some good provided found in Boulesteix and Slawski (2009) and Jeffery *et al.* (2006).

To evaluate a ranked gene list, the prediction performance of the top genes is a commonly used criterion (Furlanello *et al.*, 2003; Jeffery *et al.*, 2006). The selection probability, which quantifies the stability of the rank of a gene in the list gained more notice Pepe *et al.* (2003), Boulesteix and Slawski (2009). However, to our knowledge, there is inadequate emphasis on using these two criteria simultaneously to evaluate a ranking method.

[1] Corresponding author: Department of Statistics, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju, 61186, Korea. E-mail: espark02@chonnam.ac.kr

Published 31 May 2018 / journal homepage: http://csam.or.kr

The receiver operating characteristic (ROC) curve directly evaluates the differential ability of a gene as well as provides complete information about the relation between specificity and sensitivity that enable its utility in gene ranking Pepe *et al.* (2003). Two popular summary indexes of the ROC, the area under the ROC curve (AUC) and the partial area under the ROC curve (pAUC), have been used exclusively for such a purpose. Joober *et al.* (1999) and De Alava *et al.* (2000) used AUC to compare the diagnostic performance of different genes in clinical practice. Pepe *et al.* (2003) used pAUC given a prescribed specificity range for ranking genes. Both AUC and pAUC are rank-based statistics that do not incorporate differential magnitude of gene expressions and may lack the power to select some important genes. The modified AUC (mAUC), proposed in Yu *et al.* (2014), was shown to take advantage of both AUC and pAUC; therefore, we propose to use mAUC for gene ranking. We have demonstrated its properties using the two criteria mentioned above.

In Section 2, we reviewed the definition of AUC, pAUC, and mAUC, and proposed to rank genes by mAUC. We also gave the definitions of the two ranking evaluation measures. We then demonstrated our method using simulation study in Section 3 and two real microarray examples in Section 4. The Hierarchical Bayes method was popularly used in gene expression analysis; therefore, in our numerical studies, we also compare our method to a frequently used Hierarchical Bayes statistic – the moderated *t*-statistics (modT) (Smyth *et al.*, 2004). A discussion and conclusion are presented in Section 5.

2. Methods

2.1. mAUC for gene ranking

2.1.1. Definition of AUC, pAUC, and mAUC: a review

Here, we review the three ROC curve based summary indexes: AUC, pAUC, and mAUC, and propose applying mAUC to gene ranking. Let n and m be sample sizes of diseased and non-diseased subjects, respectively. Suppose there are p genes under study. Let $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{nk})^T$ and $\boldsymbol{X}_k = (X_{1k}, \ldots, X_{mk})^T$ denote the random scores of gene g_k for the diseased and non-diseased subjects, respectively for $k = 1, \ldots, p$. Let c be the fixed threshold. Then sensitivity and specificity of gene g_k is defined as $\mathrm{se}_k(c) = \Pr(\boldsymbol{Y}_k > c)$ and $1 - \mathrm{sp}_k(c) = \Pr(\boldsymbol{X}_k > c)$, respectively. The corresponding ROC curve is a plot of $\{(1 - \mathrm{sp}_k(c), \mathrm{se}_k(c)) : -\infty < c < \infty\}$, and the AUC of g_k is $\mathrm{AUC}_k = \int_0^1 \mathrm{se}_k(\mathrm{sp}_k{}^{-1}(1-t))dt$, which is equal to $P(\boldsymbol{Y}_k > \boldsymbol{X}_k)$ (Bamber, 1975). It follows that a pAUC, with a pre-fixed $t \in (0, 1)$ is defined as the integration of the ROC curve over a specific range $\mathrm{pAUC}_k(t_0) = \int_0^{t_0} \mathrm{se}_k(\mathrm{sp}_k^{-1}(1-t))dt$, $0 \le t_0 \le 1$. This implies that the perfect $\mathrm{pAUC}_k(t_0)$ is t_0. In practice, we are interested in the pAUC with high specificity; that is, t_0 is usually small.

The mAUC for gene g_k proposed in Yu *et al.* (2014) is defined as:

$$\mathrm{mAUC}_k = \Pr(\boldsymbol{Y}_k - \boldsymbol{X}_k > \delta) + (1 - \lambda)\Pr(0 < \boldsymbol{Y}_k - \boldsymbol{X}_k \le \delta)) \tag{2.1}$$

with some prefixed $0 \le \lambda \le 1$ and $\delta \ge 0$. It can be rewritten as a weighted average of two AUCs:

$$\mathrm{mAUC}_k = (1 - \lambda)\Pr(\boldsymbol{Y}_k > \boldsymbol{X}_k) + \lambda \Pr(\boldsymbol{Y}_k > \boldsymbol{X}_k + \delta), \tag{2.2}$$

where the first part is the original AUC of gene g_k, and the second part requires a larger difference between the diseased and non-diseased subjects. That is, for genes with the same AUCs, the larger the mAUC, the farther the separation between the two groups. From Equation (2.2), the estimate of mAUC can be obtained using the two conventional estimates of AUCs.

Note that, even though the definition of mAUC comes from a ROC curve, as in Equation (2.1), it is not necessarily based on the ROC curve. It can be treated as the weighted probabilities of two events, $\{\boldsymbol{Y}_k - \boldsymbol{X}_k > \delta\}$ and $\{0 < \boldsymbol{Y}_k - \boldsymbol{X}_k \leq \delta\}$; as λ increases, the importance of the event $\{\boldsymbol{Y}_k - \boldsymbol{X}_k > \delta\}$ increases; therefore, the larger difference between the two groups is emphasized more.

Define an empirical estimate of AUC as:

$$\widehat{\mathrm{AUC}}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\psi(Y_{ik}, X_{jk})}{nm},$$

where

$$\psi(y, x) = \begin{cases} 1, & y > x, \\ 0.5, & y = x, \\ 0, & y < x. \end{cases}$$

Then an empirical estimate of mAUC is

$$\widehat{\mathrm{mAUC}}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(1-\lambda)\psi(Y_{ik}, X_{jk}) + \lambda\psi(Y_{ik}, X_{jk} + \delta)}{nm}. \tag{2.3}$$

Based on Yu *et al.* (2014, 2015), a heuristic choice of δ is $z_{1-\alpha/2}\sigma_k$, where z_α is the α quantile of a standard normal distribution, and σ_k is the standard deviation of $\boldsymbol{X}_k$. From Equation (2.2), λ determines how much weight is put on the original AUC. In this work, we fix δ and vary λ for comparison. More details on mAUC can be found in Yu *et al.* (2014, 2015).

2.1.2. Parameter selection for mAUC in gene ranking

Generally, σ_k can be estimated from the sample variance. However, to obtain an accurate estimation of variance for each gene, a large number of samples is required, which is difficult for many studies. For example, it's hard to collect many patient samples for some complex diseases such as cancer. Therefore, we propose to set $\delta = \max(\sigma_0, \hat{\sigma})$, where $\hat{\sigma}$ is the pooled sample standard deviation from all genes, that is,

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{p}\sum_{k=1}^{m+n}\left(Z_{kj} - \bar{Z}_j\right)^2}{(n+m-1)p},$$

where $Z_{kj} = X_{kj}$ for $1 \leq k \leq m$, $Z_{kj} = Y_{kj}$ for $k \geq (m+1)$, and $\bar{Z}_j = \sum_{k=1}^{m+n} Z_{kj}/(n+m)$. The idea of borrowing information from other genes is commonly used in estimating variance of genes such as Cui and Churchill (2003), Cui *et al.* (2005), Smyth (2004). Also, the term σ_0 is pre-specified as the standard deviation of the gene with the largest AUC value. This will prevent δ from being too small in case the pooled variance is too small when a large portion of genes are non-differentially expressed in practice.

2.2. Evaluation metrics

In this section, we introduced two commonly used metrics that evaluate a ranked gene list in terms of its stability and prediction power, respectively.

2.2.1. Selection probability

A stable ranked gene list is usually preferred in clinical and genomic applications. However, the ranks of genes may vary when some genes change in their observations. To measure the ranking stability, Pepe *et al.* (2003) utilized a selection probability that quantifies the degree of confidence in choosing the i^{th} gene among the top K:

$$P_g(K) = \Pr\{\text{the gene } g \text{ is ranked in the top } K\} = \Pr\{\text{Rank}(g) \leq K\}.$$

Pepe *et al.* (2003) also proposed an estimate of $P_g(k)$ using the bootstrap method. Some other types of alternative measures to evaluate the stability of the ranked gene list are summarized in Boulesteix and Slawski (2009).

We propose using the following averaged selection probability (ASP) of the top K genes as a metric for the stability of a ranking method:

$$\text{ASP}(K) = \frac{\sum_{i=1}^{K} P_{g_i}(K)}{K}. \tag{2.4}$$

Pepe *et al.* (2003) proposed selection probability for a single gene, while ASP was the averaged selection probability proposed for the top K genes. ASP is a new metric; however, it is conceptually not because Pepe *et al.* (2003) also mentioned that the select probability for all top genes are informative for the stability of the ranked list. ASP is conceptually natural and simple; suppose the data was slightly perturbed. The selection probability for a single gene in the top list quantifies how often the gene is still in the top list after perturbation; while ASP quantifies the average frequency of the top ranked genes staying in the top after perturbation.

2.2.2. Relative classifier information metric

Another important characteristic of a gene list is its prediction accuracy as a classifier. The classifier could be constructed using some well-known supervised learning algorithms. Many studies present the success of a classifier by the accuracy of predicting responses of the test set, they may not work well for the imbalanced responses, where the ratio of sizes of two responses is far away from 1. Thus, we evaluate the prediction efficiency by the relative classifier information metric (RCI) defined as in Sindhwani *et al.* (2001).

For a given classifier's performance on a test set, the RCI measure is defined below:
Let q_{ij} be the number of times that an input class (I or true class) for a subject with actual label C_i is predicted as C_j. For a two-class problem, there are only two labels, C_1 and C_2; i.e., $i = 1, 2$, and $j = 1, 2$. The probability that the input class I has a true label C_i is given by

$$P(I \in C_i) = \frac{\sum_j q_{ij}}{\sum_{ij} q_{ij}}.$$

The Shannon's entropy of the data set before classification can be used to measure the uncertainty associated with a test set before a classification model has been applied and is calculated as

$$H_d(I) = \sum_i -P(I \in C_i) \log P(I \in C_i).$$

The probability that the output class (O or predicted class) for a subject is predicted to belong to class C_j is

$$P(O \in C_j) = \frac{\sum_i q_{ij}}{\sum_{ij} q_{ij}}.$$

The probability that a sample labeled as C_j by the classifier belongs to C_i is

$$P(I \in C_i | O \in C_j) = p_{ij} = \frac{q_{ij}}{\sum_i q_{ij}}.$$

Therefore, the uncertainty for a sample after classification is performed is

$$H_{o_j}(I|O \in C_j) = \sum_i -p_{ij} \log p_{ij},$$

and the overall uncertainty after classification is

$$H_o(I|O) = \sum_j P(O \in C_j) H_{o_j}(I|O \in C_j).$$

The reduction in uncertainty due to the classifier is used as the RCI score:

$$\text{RCI score} = H_d - H_o.$$

The RCI metric is an entropy-based measure that corrects for differences in prior probability due to unequal class sizes. By taking into account this prior probability, a better measure of classification is obtained (Sindhwani *et al.*, 2001). A higher RCI score indicates that greater reduction of the uncertainty for the test set is achieved after implementing the classifier. A detailed definition and more discussion can be found in Sindhwani *et al.* (2001) and Jeffery *et al.* (2006).

In this paper, we have built classifiers based on ranked gene lists through four supervised classification methods: support vector machines with linear kernel (SVMl) and radial kernel (SVMr), respectively, naive Bayes (NB) classification, and K-nearest neighbors (KNN). We used R package *class* for the implementation of KNN and package *e1071* for the others.

2.3. A highly related practical question: how many top genes to use?

In practice, choosing the number of genes to select is a key step after gene ranking. A possible way is to optimize ASP (optASP) across all possible Ks. We choose top d genes by

$$d = \arg \max_{K \in \{1,\ldots,p\}} \text{ASP}(K). \tag{2.5}$$

To get a sparse model, we propose using the following *optASP criterion*

$$d = \arg \min_{K \in \{1,\ldots,p\}} \{\text{ASP}(K) \geq \text{MASP} - \text{sASP}\}, \tag{2.6}$$

where MASP and sASP stand for the maximum and standard deviation of ASPs, respectively.

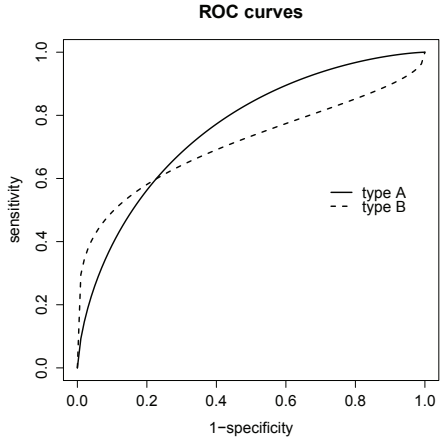


Figure 1: *The ROC curves for genes of types A and B in simulation studies. ROC = receiver operating characteristic.*

Table 1: Selection probabilities in simulation studies

Scenario	Method	$m = n = 10$		$m = n = 20$	
		$P_A(20)$	$P_B(20)$	$P_A(20)$	$P_B(20)$
$\rho = 0$	auc	0.568 (0.136)	0.332 (0.143)	0.896 (0.092)	0.634 (0.135)
	pauc	0.275 (0.129)	0.651 (0.145)	0.493 (0.144)	0.954 (0.064)
	modt	0.515 (0.134)	0.627 (0.146)	0.847 (0.131)	0.862 (0.129)
	mauc0.25	0.621 (0.134)	0.553 (0.151)	0.890 (0.137)	0.804 (0.147)
	mauc0.5	0.586 (0.138)	0.662 (0.147)	0.841 (0.183)	0.865 (0.187)
	mauc0.75	0.535 (0.142)	0.737 (0.142)	0.804 (0.188)	0.906 (0.194)
	mauc1.0	0.479 (0.152)	0.786 (0.134)	0.769 (0.163)	0.944 (0.159)
$\rho = 0.6$	auc	0.573 (0.306)	0.306 (0.275)	0.876 (0.214)	0.607 (0.279)
	pauc	0.272 (0.231)	0.640 (0.259)	0.494 (0.246)	0.959 (0.099)
	modt	0.512 (0.302)	0.609 (0.308)	0.731 (0.332)	0.754 (0.335)
	mauc0.25	0.626 (0.299)	0.515 (0.312)	0.798 (0.321)	0.715 (0.311)
	mauc0.5	0.591 (0.300)	0.617 (0.300)	0.723 (0.354)	0.741 (0.356)
	mauc0.75	0.544 (0.302)	0.700 (0.274)	0.657 (0.365)	0.748 (0.385)
	mauc1.0	0.493 (0.296)	0.757 (0.245)	0.623 (0.351)	0.785 (0.378)

3. Simulation study

In simulation studies, we present selection probabilities of informative genes and RCI scores for the top selected genes, using AUC, pAUC, modT, and mAUC as the ranking methods, where the upper bound of FPR to define pAUC is 0.1. Different λs are used for mAUC; for example, *mauc0.5* stands for mAUC with $\lambda = 0.5$. For each ranked gene list, top k genes were used to construct the classifier, $k = 1, \ldots, 20$. For each k, we average the RCI scores over the four classifiers, i.e., SVMl, SVMr, NB, and KNN.

We generate 2,000 genes, of which 99% are non-informative, in the sense that genes from diseased and non-diseased subjects follow the same distribution, say, $N(0, 1)$. The informative genes have two types: 10 genes generated from type A following $N(1, 1)$ for diseased subjects and another 10 genes simulated from type B following $N(1.25, 2^2)$ for diseased subjects. For both types, non-diseased subjects follow the standard normal distribution. Figure 1 shows the ROC curves for genes with types A and B. The $2m$ and $2n$ are sample sizes of the non-diseased and diseased groups, respectively. Note that genes of type B have larger mean differences between diseased and non-diseased groups than

genes of type A, and genes of type B have larger variance, too.

We consider two kinds of data structure, in which different genes are correlated or uncorrelated. In the former structure, all the informative genes with the same type are correlated with the correlation coefficient 0.6. Note that genes from type A have larger AUCs, while genes from type B have higher partial ROC curves when FPR is relatively low. We generate n and m samples for disease and non-diseased groups, respectively, as the training set. Other independent n and m samples are generated in the same way as the test set, upon which the averaged RCI scores are computed over 500 repetitions.

Table 1 summarizes the proportions of the informative genes selected, averaged across 500 repetitions for both of correlated and uncorrelated genes. $P_A(20)$ and $P_B(20)$ are probabilities for genes of type A and genes of type B being ranked in the top 20, respectively. In simulation studies, $P_A(20)$ ($P_B(20)$) is estimated as the average fraction of genes of type A (type B) ranked in top 20 in the test set. AUC is less effective to select type B genes, while pAUC lacks the power to select type A genes. The mAUC and modT methods show better performances than AUC and pAUC methods, in the sense of achieving good balance between selecting type A genes and selecting type B genes. For mAUC approaches, the selection probability of type B genes increases as λ increases, while the selection probability of type A genes decreases. These findings look natural since AUC only considers a gene's global discriminant performance, while pAUC considers the area with low FPR but high TPR. Also, modT has a good balance of selecting both types of genes. Performance of mAUC is good because mAUC keeps the information on the entire ROC curve as well as assigns more importance on the high specificity range.

In Figure 2, pAUC and AUC present weaker prediction performances than other methods. When λ is greater than 0.5, mAUC consistently performs better than others under a different number of top genes, different sample sizes and different correlation structures. Combined with the results from selection probabilities, this implies that type B genes are more predictive than type A genes. However, it fails to have a good prediction performance even though the pAUC approach shows high selection probabilities for type B genes. The reason for this may be because pAUC ignores type A genes too much despite type A genes playing roles to some degree in prediction. The ranking method that achieves good balance between selecting the two types of genes will have good prediction performance; once some level of balance is obtained, selecting more genes of type B is better.

Findings under the correlated scenario are similar to the uncorrelated genes. When the sample size is as small as 10 for each group, the selection probability for each type of genes under the correlated scenario is close to the uncorrelated cases; when the sample size increases, the selection probability under the correlated scenario is lower than that under the uncorrelated scenario (Table 1). The cumulative RCI scores for correlated cases are smaller than those for uncorrelated cases (Figure 2). We also provide the average accuracy of each method in Figure 3 similar findings are observed between different methods found in Figure 2.

4. Real examples

In this section, we use two microarray data sets on colon and lung cancers to demonstrate our method. We randomly divide data into training and test sets; half of the sample comprises the training set and the left half comprises the test set. Gene ranking and training of classifiers were performed on the training sets only. The classification performance of each gene list is evaluated by the RCI score on the test set. The cumulative RCI scores versus selected the top k genes are displayed in Figure 4(a) for $k = 1, \ldots, 50$. RCI scores are averaged over the four classification methods for the test set. This procedure is repeated 500 times while selecting a new test set each time.

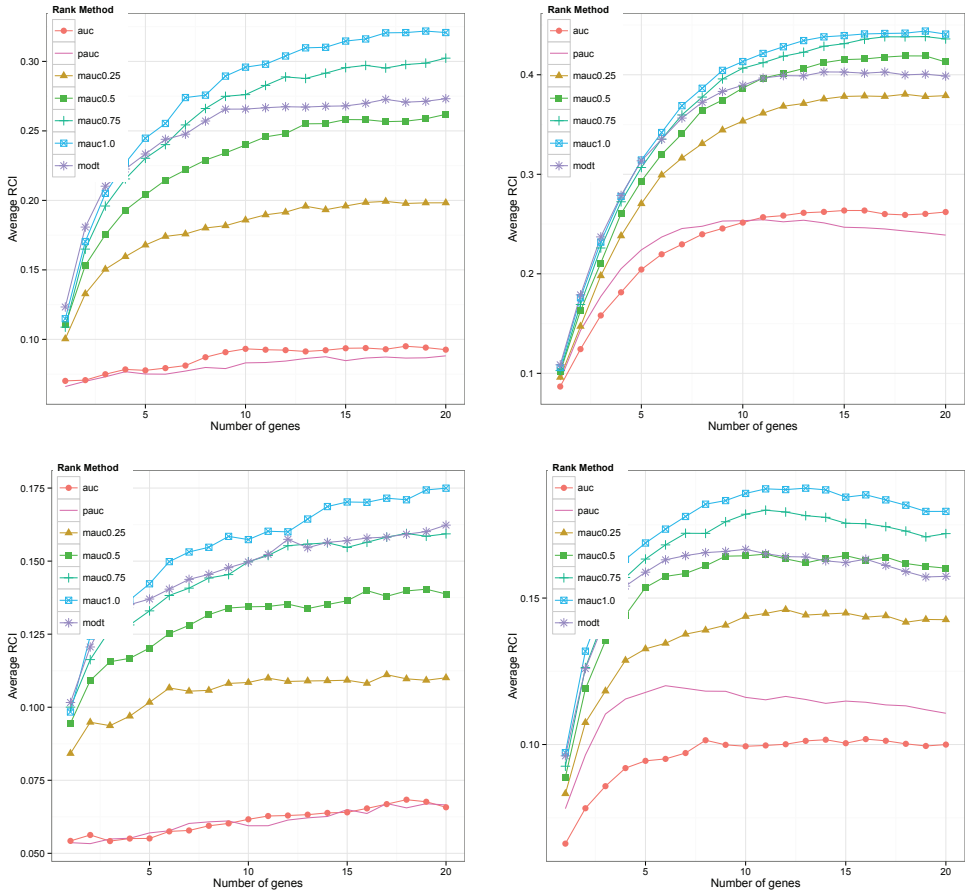


Figure 2: *The average RCI score of each ranked gene list in simulation studies:* $m = n = 10$, $\rho = 0$ *(top left);* $m = n = 20$, $\rho = 0$ *(top right);* $m = n = 10$, $\rho = 0.6$ *(bottom left);* $m = n = 20$, $\rho = 0.6$ *(bottom right). RCI = relative classifier information.*

4.1. Example 1: colon cancer data

Expression levels of 40 tumor and 22 normal colon tissues for 2,000 human genes with the highest minimal intensity were measured from 62 subjects (Alon *et al.*, 1999). The data can be downloaded from http://microarray.princeton.edu/oncology/ or from the *colonCA* package at http://www.bioconductor.org. We preprocessed the data by logarithm transformation and quantile normalization.

In Figure 4(a), mAUC with $\lambda = 1.0$ almost uniformly outperforms others ranged from the top 1 gene selected to the top 50 genes selected. AUC and pAUC approaches cannot achieve good RCI scores as in the simulation study; the modT approach is better than AUC and pAUC, but less powerful than mAUC with $\lambda \geq 0.5$. Note that mAUC with λ equal to 0 (AUC) is not good in the sense of RCI scores. This fact indicates that it is not a good idea to set λ to be extremely small.

Figure 4(b) shows the plot of ASP when a different number of top genes are selected. The mAUC approaches have higher ASP than others; as the λ increases, the ASP increases too. pAUC and AUC have lowest and second lowest ASP values, respectively, the same as the orders of their RCI scores.

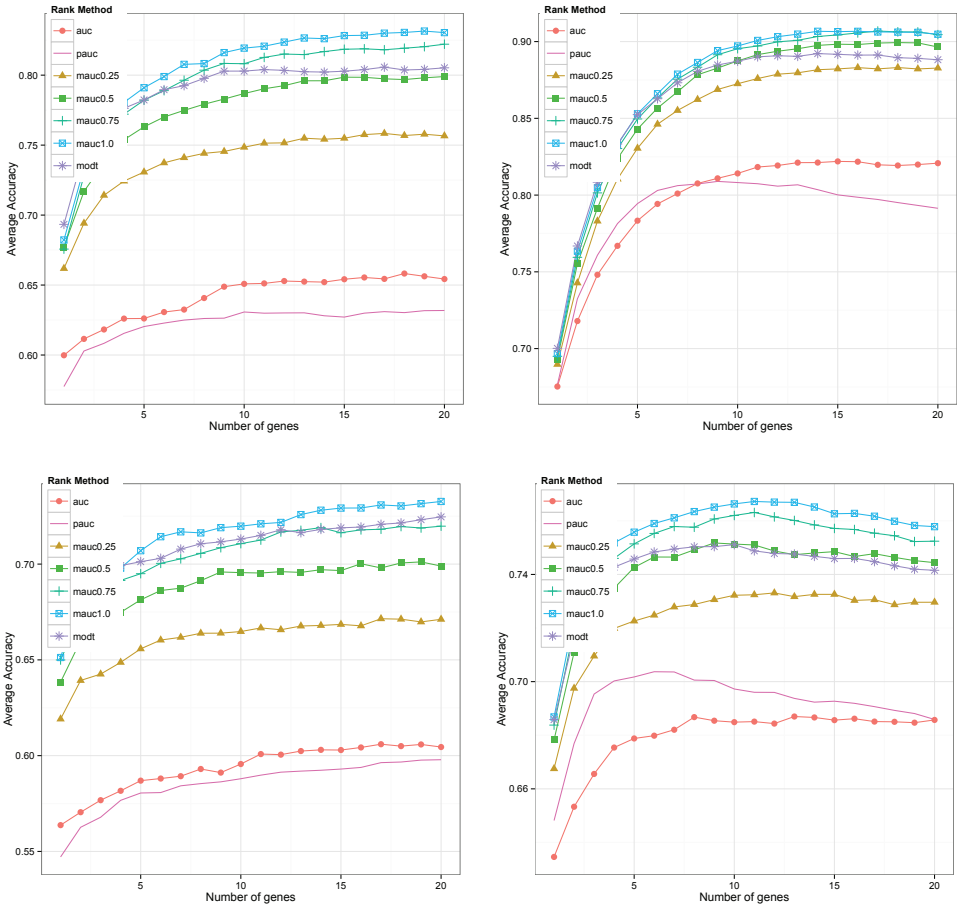


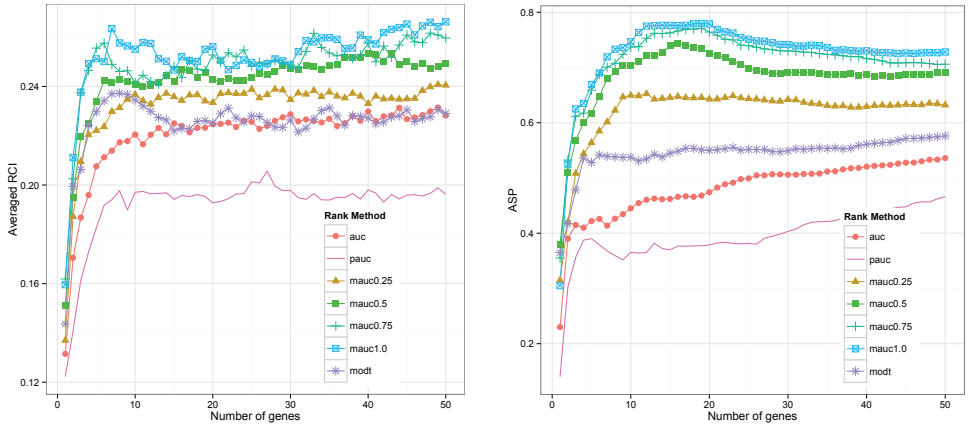
Figure 3: *The average accuracy of each ranked gene list in simulation studies:* $m = n = 10, \rho = 0$ *(top left);* $m = n = 20, \rho = 0$ *(top right);* $m = n = 10, \rho = 0.6$ *(bottom left);* $m = n = 20, \rho = 0.6$ *(bottom right).*

modT approach has a larger ASP than AUC and pAUC, but smaller than the mAUC approaches.

We also apply the *optASP criterion* (2.6) for variable selection; therefore, we use ASP to select genes for each ranking method. Table 2 lists the results; subsequently, we also show the average accuracy of each method. The number of genes selected by the above ranking methods and their corresponding cumulative RCI scores are displayed. mAUC with $\lambda = 1.0$ achieves highest RCI score and accuracy, and only 9 genes were used. The modT method shows a similar sparsity as *mauc1.0*, but has lower prediction performance. To have a good balance between prediction performance and sparsity, *mauc1.0* is the best choice. Note that the *optASP criterion* may also be used to select λ for mAUC approaches. We may select λ for mAUC by choosing the one with the largest RCI score or the most parsimonious model.

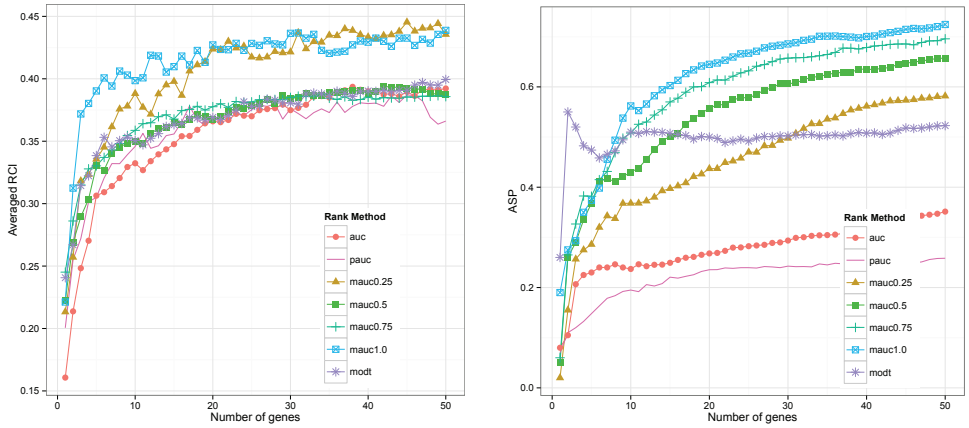
4.2. Example 2: lung cancer data

In this example, we used a lung cancer data set, GSE10245 (Kuner *et al.*, 2009), downloaded from the Gene Expression Omnibus (GEO) data repository. There are two subtypes of lung cancer in



(a) (b)

Figure 4: *The results of Colon cancer study: (a) The average RCI score of each ranked gene list; (b) The plot of ASP. RCI = relative classifier information; ASP = averaged selection probability.*



(a) (b)

Figure 5: *The results of Lung cancer study: (a) The average RCI score of each ranked gene list; (b) The plot of ASP. RCI = relative classifier information; ASP = averaged selection probability.*

these data: subtype adenocarcinoma (AC) and subtype squamous cell carcinoma (SCC). GSE10245 is composed of 40 AC and 18 SCC samples. It includes 54,675 expressions, and they were measured under the sample platform GPL157. The data set was preprocessed separately by rma() function in *affy* package.

In Figure 5(a), mAUCs with $\lambda = 1.0$ or $\lambda = 0.25$ almost uniformly outperform others. The modT approach is better than AUC and pAUC, but less powerful than mAUC with $\lambda = 1.0$ or with $\lambda = 0.25$. Note that in this example, the RCI score of the mAUC approach is not linearly related to λ. Generally, $\lambda = 1.0$ is recommended, which is consistent with the simulation and the colon cancer studies.

Figure 5(b) shows the plot of ASP when a different number of top genes are selected. It shows

Table 2: Number of selected markers, averaged RCI score and accuracy when the *optASP criterion* is applied in real studies

Data	Evaluation	auc	pauc	mauc0.25	mauc0.5	mauc0.75	mauc1.0	modt
Colon	Number of markers	22	32	13	12	10	9	6
	RCI score	0.225	0.194	0.235	0.240	0.241	**0.256**	0.234
	Accuracy	0.830	0.810	0.837	0.839	0.844	**0.850**	0.838
Lung	Number of markers	38	14	21	20	22	23	2
	RCI score	0.389	0.354	0.423	0.368	0.377	**0.428**	0.267
	Accuracy	0.925	0.914	0.935	0.919	0.921	**0.937**	0.868

RCI = relative classifier information; optASP = optimize averaged selection probability.

consistent findings to those of the colon cancer study; mAUC approaches have higher ASP as their λ increases. Note that in this case, the modT method is best when only a very small number of genes are used, for example, less than 5.

After using the *optASP criterion* (2.6) for gene selection (Table 2), *mauc0.25* and *mauc1.0* achieve similarly the best result, the highest RCI score and accuracy, and a similarly modest number of genes; *mauc1.0* is slightly better than *mauc0.25*. The modT method leads to the sparsest model but has the lowest prediction RCI. The AUC method selects more genes than others and achieves a smaller RCI score and accuracy than *mauc0.25* and *0.75*, while pAUC approaches spend less markers and have lower RCI scores and accuracy than mAUC approaches. In the balance of sparsity and prediction, *mauc1.0* is recommended.

5. Conclusion and discussion

In this paper, we proposed using the mAUC to rank genes. We evaluated the ranking methods based on two criteria of the stability and prediction performance of the ranked list (which are correspondingly measured by the selection probability) and the RCI score, respectively. It is shown that with both real examples and simulation studies, the proposed method has a good prediction performance and provides a stable rank list of genes.

From empirical results, we have found that the AUC method selects genes that have small differences between diseased and non-diseased groups and small variance. However, the pAUC method tend to choose genes with large differences between the two groups and also with a large variance. That is, both AUC and pAUC approaches failed to achieve good prediction performances. We have demonstrated that using the mAUC as the ranking metric has a balance between selecting the above two types of genes with better prediction performance. We also show that the mAUC-based method outperforms modT, which is popularly used for screening microarray gene expression.

As pointed out by Pepe *et al.* (2003), ROC statistics are rank-based; this presumably infers their robustness, but at the expense of ignoring the quantitative information of the gene expression. The mAUC metric uses the rank statistic information as well as assigns additional importance to larger differential gene expressions, which may provide to a more sensible ranking list. This may be the reason why mAUC can achieve a ranked list that is both stable and powerful for class prediction. We only illustrated the proposed method through microarray gene expression data; however, mAUC can also be used in other types of data, such as RNA-seq, and DNA methylation data.

Acknowledgement

This study was financially supported by Chonnam National University (2016-2706).

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proceedings of the National Academy of Sciences*, **96**, 6745–6750.

Bamber D (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Boulesteix, AL and Slawski M (2009). Stability and aggregation of ranked gene lists, *Briefings in Bioinformatics*, **10**, 556–568.

Cui X and Churchill GA (2003). Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology*, **4**, 210.

Cui X, Hwang JT, Qiu J, Blades NJ, and Churchill GA (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics*, **6**, 59–75.

De Alava E, Panizo A, Antonescu CR, Huvos AG, Pardo-Mindán FJ, Barr FG, and Ladanyi M (2000). Association of EWS-FLI1 type 1 fusion with lower proliferative rate in Ewing's sarcoma, *The American Journal of Pathology*, **156**, 849–855.

Efron B, Tibshirani R, Storey JD, and Tusher V (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, **96**, 1151–1160.

Furlanello C, Serafini M, Merler S, and Jurman G (2003). Entropy-based gene ranking without selection bias for the predictive classification of microarray data, *BMC bioinformatics*, **4**, 54.

Jeffery IB, Higgins DG, and Culhane AC (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, *BMC Bioinformatics*, **7**, 359.

Joober R, Benkelfat C, Toulouse A, *et al.* (1999). Analysis of 14 CAG repeat-containing genes in schizophrenia, *American Journal of Medical Genetics (Neuropsychiatric Genetics)*, **88**, 694–699.

Kuner R, Muley T, Meister M, *et al.* (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes, *Lung Cancer*, **63**, 32–38.

Newton MA, Noueiry A, Sarkar D, and Ahlquist P (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method, *Biostatistics*, **5**, 155–176.

Noma H and Matsui S (2013). Empirical Bayes ranking and selection methods via semiparametric hierarchical mixture models in microarray studies, *Statistics in Medicine*, **32**, 1904–1916.

Noma H, Matsui S, Omori T, and Sato T (2010). Bayesian ranking and selection methods using hierarchical mixture models in microarray studies, *Biostatistics*, **11**, 281–289.

Pepe MS, Longton G, Anderson GL, and Schummer M (2003). Selecting differentially expressed genes from microarray experiments, *Biometrics*, **59**, 133–142.

Sindhwani V, Bhattacharya P, and Rakshit S (2001). Information theoretic feature crediting in multiclass support vector machines. In *Proceedings of the First SIAM International Conference on Data Mining*, 5–7.

Smyth GK (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, **3**, 3.

Storey JD (2003). The positive false discovery rate: a Bayesian interpretation and the q-value, *Annals of Statistics*, **31**, 2013–2035.

Tusher VG, Tibshirani R, and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Yu W, Chang YCI, and Park E (2014). A modified area under the ROC curve and its application to marker selection and classification, *Journal of the Korean Statistical Society*, **43**, 161–175.

Yu WB, Park E, and Chang YCI (2015). Comparison of paired ROC curves through a two-stage test, *Journal of Biopharmaceutical Statistics*, **25**, 881–902.

Received February 22, 2018; Revised April 10, 2018; Accepted April 16, 2018