Feature Selection using Bootstrapped ROC Curves

Ping Xu1*, Xiang Liu1, David Hadley2*, Shuai Huang3, Jeffrey Krischer1 and Craig Beam3

1Department of Pediatrics, College of Medicine, University of South Florida, 3650 Spectrum Blvd, Suite 100, Tampa, Florida, USA
2Department of Biomedical Sciences, West Michigan University, Kalamazoo, MI 49006, USA
3Current Address: Population Health Research Institute, Division of Population Health Sciences and Education, St George’s University of London, London, United Kingdom

Abstract

Background: In modeling a N by m data matrix, i.e. N samples on a m dimensional space, the issue arises when m is bigger than N. The sample size cannot be increased, especially in medical research, due to the limited number of diseased subjects. Feature selection is often used to select a subset of relevant m variables, often lower than N, for use in model construction.

Method: A multiple step bootstrap method is proposed to quantify relevance of candidate predictors with the outcome based on their areas under the Receiver Operating Characteristic curve (ROCAUCs) from bootstrap resamples and then select only significant variables, which meet pre-specified criteria, as a feature selection process.

Results: Extensive simulation was conducted using thousands of predictor variables and 5 levels of prediction ability between the true predictor and the outcome. The results from the simulation data indicate that the mean of ROCAUCs from bootstrap samples is close to the true ROCAUC. Even with only 30 cases and 30 controls, 25 out of 25 listed predictor variables provide the correct level of classification ability by using mean of bootstrapped ROCAUCs. The proposed bootstrapped ROCAUCs method outperforms the single ROCAUC. The standard error of mean of bootstrapped ROCAUCs was 20% to 50% smaller than the standard error of the single ROCAUC estimate from the original sample. An illustrative example is presented to apply the proposed methodology to identify the gene expressions that could predict clinical survival in breast cancer patients, using the Van’t Veer study’s breast cancer data.

Conclusion: We conclude that the bootstrapped ROCAUCs methodology is intuitive and attractive for use in feature selection problems when the goals of the study are to identify important predictors and to provide insight regarding the discriminative or predictive ability of individual predictor variables. Such goals are common among microarray studies and new biomarker discovery.

Keywords: Feature selection; Receiver Operating Characteristics curve (ROC curve); Bootstrap; Variable importance; Variable selection

Abbreviations: ROC: Receiver Operating Characteristic; ROCAUC: Area Under Receiver Operating Characteristic Curve

Background

Information technology advancement has brought an explosive growth of data in recent decades. We collect data on a diverse and numerous assortments of variables, not knowing which ones will be relevant to the outcome of interest. The sample size of study subjects cannot be increased, especially in medical research, due to the limited number of diseased subjects. In modeling an N by m data matrix, i.e. N subjects on an m dimensional space, the challenge is to identify relevant variables to be included in modeling the outcome. Thus, variable filtering plays an important role in reducing the number of variables before the formal model building. Thereafter, a subset of size k variables (usually k<N) remains as the result of the feature selection process and can be accomplished by a well-understood method for modeling low dimensional data.

The goal of variable filtering is to eliminate the majority of irrelevant variables, while keeping as many of the true predictors as possible, by reducing the size of candidate variables to a smaller number, k. The value of a feature selection procedure largely depends on the accurate assessment of variables in terms of importance to the outcome (variable importance) and the criteria for selecting the candidate variables to be included in the model building (variable selection). When both outcome and predictors are continuous, Fan and Lv [1] proposed Sure Independence Screening (SIS), which ranks all the variables by the absolute value of empirical Pearson’s correlation coefficient between the outcome and each predictor. The method is not applicable when the outcome is dichotomized or survival time, such as case-control or time to disease onset data. Genuer et al. [2] proposed a method for variable selection using random forest score of importance, which limits the predictors as the classifiers. In medical research, the researchers are often interested in the search of biomarkers that could be used to differentiate the disease population from non-diseased population or to predict time to event. Biomarkers are often measured on a continuous scale. In this setting, Pepe et al. [3] proposed to evaluate the predictors based on their individual empirical ROC value, and applied to an ovarian cancer dataset to select a subset of genes that could distinguish between cancerous and normal organ tissues. Jeffrey et al. [4] compared and evaluated a ROC method with other traditional methods, such as t-statistics. They concluded that Pepe’s method performed well with datasets that had low levels of noise and large sample size. However, the method cannot be used when the...
outcome is survival time. Moreover, it did not take into consideration the variability of quantification process due to the finite number of study subjects and large number of predictors in high dimensional data. Boulesteix and Slawski [5] discussed the variability of gene ranking methods and showed ranked gene lists are highly unstable in the sense that a small change of the data set usually affects the obtained gene list. Taking into account the above observations, and considering the importance of variable filtering in high dimensional data, we propose a method using bootstrapped ROCAUCs to quantify each variable’s discriminative or predictive ability. By selecting only relevant variables as a filtering process, which meets pre-specified criteria, the number of variables for model constructions is reduced.

This remaining paper is organized as follows. In Section “Methods: procedure for bootstrapped ROCAUCs”, we describe the procedure to generate bootstrapped ROCAUC estimates for quantifying variable importance and we recommend the various criteria for variable selection. In Section “Application results: breast cancer gene expression and clinical survival”, we evaluate the performance of proposed method, based on simulations of normal models for the case-control study (section “Simulation results”). In Section “Simulation study I”, we present simulations for a prospective follow-up study where the outcome is survival time. In section “Simulation study II”, an illustrative example is presented to apply the proposed methodology to the Van’t Veer dataset, which was screened for gene expression variables that could predict clinical survival in breast cancer patients. We then follow with a discussion on the applicability of the proposed feature selection method and draw our conclusions in Section “Conclusion”.

Methods: Procedure for Bootstrapped Rocaus

The quantification of variable importance is crucial not only for ranking the candidate variables in the screening process but also to interpret and understand the data. It is the initial step of variable screening. When the predictor variables are measured on a continuous scale, the Receiver Operating Characteristic (ROC) curve is one of the best statistical techniques used to characterize their ability in classifying or predicting the disease outcome. The area under ROC curve (ROCAUC) is the summary index of ROC curve and can be interpreted as a measure of distance or, equally, a measure of stochastic dominance.

Area under Receiver Operating Curve (ROCAUC)

We review ROCAUC in this subsection. For a continuous variable $Y$ and a binary outcome $D$, let $D=1$ if diseased and $D=0$ if non-diseased. Using a threshold $c$ to define a binary test from a continuous variable $Y$ as

Positive if $Y \geq c$,

Negative if $Y \leq c$;

Let the corresponding true and false positive rate at the threshold $c$ be $TPR(c)$ and $FPR(c)$ respectively, we define:

$TPR(c)=P[Y \geq c | D=1]$,  
$FPR(c)=P[Y \geq c | D=0]$,

and the ROC curve is plotting the entire set of possible true and false positive fractions obtained by dichotomizing $Y$ with different thresholds. That is:

$ROC(\cdot)=\{ (FPR(c), TPR(c)), c \in (-\infty, +\infty) \}$.

The area under the entire ROC curve (ROCAUC) is a global summary statistic of ROC curve, based on all possible cut-off values of a variable. It is defined as:

$AUC = \int_{-\infty}^{+\infty} ROC(c)dc$

In a prospective cohort study, a binary outcome can change over time. Suppose we have a time-dependent outcome along with continuous biomarkers and we want to see how well marker $Y$ predicts the survival time for the subjects. Let $T_i$ and $C_i$ denote survival and censoring times for ith subject, We observe $(Z_i, \delta_i)$ where $Z_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Denote $D_i(t)$ the time-dependent outcome status for subject i. at time t. For any threshold $c$, the true positive and false positive rates are time-dependent functions, defined as

$TPR(c, t)=P[Y > c | D(t)=1]$

$FPR(c, t)=P[Y > c | D(t)=0]$

The time-dependent ROC curve plots $TPR(c, t)$ vs. $FPR(c, t)$ for any threshold $c$, so that, the area is a time-dependent function:

$AUC(t_i) = \int_{-\infty}^{+\infty} TPR(c, t_i)dc[FPR(c, t_i)]$

This function returns the unique biomarker ROCAUC value corresponding to the time point of interest with taking account of censoring.

The ROCAUC can take on values between 0 and 1. It is a monotonically increasing function. The variable with AUCROC of 1 is a perfect predictor because the true positive is 100% and the false positive is 0%. In contrast, the variable with an area under 50 is useless for classification/prediction. ROC curves are invariant to monotone transformations of the raw data. This property makes them appealing for comparisons across variables and hence for ranking. We suggest using non-parametric estimate of ROCAUC in the sample as it is more robust and do not depend on the distributions of the raw predictor variables. Non-parametric ROCAUC estimates will be utilized to quantify the variable importance. More details on the non-parametric ROCAUC estimates can be found in the paper by Hsegerty et al. [6] for survival outcome or the paper for binary outcome by Hanley and McNeil [7].

ROCAUC estimates from bootstrap resampling

Let $S=\{(D_1, Y_1), \ldots, (D_n, Y_n)\}$ denote a sample of $N$ independent subjects, who have been measured with $m$ continuous independent variables $Y_i=(Y_{i1}, Y_{i2}, \ldots, Y_{im})$ and the binary disease outcome $(D=0$ non-diseased, $D=1$ diseased). We assume that the sampled subjects are independent identical distributed (i.i.d) random variables with the distribution function $F$, i.e.,

$\{ (D_1, Y_1), \ldots, (D_n, Y_n)\} \sim$ i.i.d $F$

The ROCAUCs based on B bootstrap samples are generated as follows:

Draw B random samples $S^*(1), \ldots, S^*(B)$ of size $N$ with replacement from $S$.

1. For the first predictor variable, obtain a non-parametric AUCROC estimate using the bootstrap sample $S^*(b)$;
2. Repeat the steps above for all B bootstrap samples.
3. Iterate step 2 and 3 for all $m$ variables;
Quantification of variable importance based on bootstrapped ROCAUCs

A total of B ROCAUC estimates are obtained for each continuous variable from B bootstrap resamples. We record all ROCAUC values and then summarize them as below for all m variables:

| Variable for classification/prediction | Frequency with ROCAUC value >0.90 | Frequency with ROCAUC value <=0.90 and >0.80 | Frequency with ROCAUC value >0.60 and <=0.80 | Frequency with ROCAUC value >0.50 and <=0.60 | Mean ROC and its Standard Error |
|---------------------------------------|----------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-------------------------------|
| Variable 1                            |                                  |                                                |                                               |                                               |                               |
| Variable 2                            |                                  |                                                |                                               |                                               |                               |
| ...                                   |                                  |                                                |                                               |                                               |                               |
| Variable m-1                          |                                  |                                                |                                               |                                               |                               |
| Variable m                            |                                  |                                                |                                               |                                               |                               |

Once we have quantified the variable importance, we can rank the predictor variables either by the mean or certain frequency of ROCAUC values, that is, we will have a rank for all m variable based on bootstrapped ROCAUCs (Figure 1):

\[ Y_{\text{ROCAUC}}^{(1)} - Y_{\text{ROCAUC}}^{(2)} - \ldots - Y_{\text{ROCAUC}}^{(m)}. \]

Variable selection criteria

It is important to recognize that an appropriate statistical approach depends on the scientific objectives of the study. The decision for selecting candidate variables should be flexible depending on the objective of study and the information you have already known on the disease. We present the various criteria for variable selection.

If the known risk factors are in the data, it is recommended to keep any variable that has the higher rank order than the known risk factors, for example, a variable with mean (ROCAUCs)>-maximum (mean (ROCAUCs)) of known risk factors. Another possibility is that we don’t have known risk factors, but we would like to include a fixed size of p candidate variables. We may keep the top p variables ranked by the frequency or mean ROCAUCs, that is, \( Y_{\text{ROCAUC}}^{(1)}, Y_{\text{ROCAUC}}^{(2)}, \ldots, Y_{\text{ROCAUC}}^{(p)} \).

When we have little information on the disease, it may be appropriate to simply keep any variable which has a high discrimination or prediction ability, such as a frequency of ROCAUC values above 0.70 in over 80% of the re-samplings. The selected candidate variables can then be included in the traditional multivariate modeling for low dimensional data for model selection process. As bootstrapping also provide the variance of ROCAUC estimates, we may consider selecting the variable based on its ROCAUC estimate’s confidence interval, such as the lower limit of confidence interval above 0.60.

Application Results: Breast Cancer Gene Expression and Clinical Survival

We apply our method to the public available dataset of gene expression profiling in predicting clinical survival outcome among breast cancer patients reported by Van de Vijver et al. [8]. The data can be downloaded through R package ‘breastCancerNKI’ (http://www.bioconductor.org/packages/2.13/data/experiment/html/breastCancerNKI.html). Total 295 breast cancer patients who were treated by modified radical mastectomy or breast-conserving surgery, followed by radiotherapy between 1984 and 1995 at the hospital of the Netherlands Cancer Institute were included. In this data set, approximately 25,000 human gene expressions were recorded for each patient. The endpoint of interest was the clinical survival time during the 10-years follow-up. The median follow up time was 7.2 years and the median survival time was 3.8 years. We evaluated the ROCAUCs from 1000 bootstrapped samples for all 24,496 gene expression markers. The results for top 15 substances/genes based on mean ROCAUCs are summarized in Table 1. The best individual gene only had a fair prediction (0.70 ± 0.06) on 10-year’s survival. Thus, the combination of gene expressions may be further investigated to improve the overall predictive ability in the multivariate models.

Simulation results

We conduct extensive simulation studies to evaluate the finite sample performance of the proposed bootstrapped ROCAUCs method under a variety of settings obtained by controlling several critical factors such as the sample size and the type of outcome. By doing so, we can compare the performance across the setting and identify the settings favorable to the method. The advantage of using simulated data is that we know the truth underlying the data and therefore we have a gold standard against which to compare results. The settings satisfy all the assumptions for bi-normal ROC model, that is, data is normally distributed for both diseased population and non-diseased population, it can be shown that true ROCAUC is:

\[ \text{ROCAUC} = \Phi \left( \frac{a}{\sqrt{1+b^2}} \right) \]

where Φ is the standard normal cumulative distribution function, \( a = \frac{\mu_{(1)} - \mu_{(p)}}{\sigma_{(1)}} \), and \( b = \frac{\sigma_{(p)}}{\sigma_{(1)}} \).
In simulation study I, we look at the scenarios from the case-control study where there are total 2000 predictor variables. Among them, 5 predictors (variable1-variable5) have true ROC AUROC over 0.90, 5 predictors (variable 6-variable10) have ROC AUROC >0.80 and ≤ 0.90, 5 predictors (variable11-variable 15) have ROC AUROC >0.70 and ≤0.80, 5 predictors (variable16-variable20) have ROC AUROC >0.60 and ≤ 0.70, 5 predictors have ROC AUROC >0.50 and ≤ 0.60 (variable21-variable25), and rest of predictors (variable26-variable2000) have ROC AUROC ≤0.60.

Simulation study I
true auc <0.60 and ≥0.50, var26-var2000 had the true roc <0.50. Summary table for 1000 bootstrap samples.

Result from simulated data for case control study (N=100:100). Summary table for 1000 bootstrap samples. Variable 1-variable5 has the true ROC AUC above

### Table 2.2: Result from simulated data for case control study (N=50:50)

| Variable | Frequency with boot strapped ROCAUC >0.90 | Frequency with boot strapped ROCAUC >0.80 and ≤ 0.90 | Frequency with boot strapped ROCAUC >0.70 and ≤ 0.80 | Frequency with boot strapped ROCAUC >0.60 and ≤ 0.70 | Frequency with boot strapped ROCAUC >0.50 and ≤ 0.60 | Mean of Boot strapped ROCAUC | Standard error of boot strapped ROCAUC | True ROCAUC | ROCAUC from the original sample | Standard error of ROCAUC from the original sample |
|----------|---------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|-----------------------------------------------|---------------------------------------------|----------------|-----------------------------------|---------------------------------------------|
| Variable 1 | 949 | 51 | 0 | 0 | 0 | 0 | 0 | 0.9302268 | 0.0193489 | 0.967 | 0.9292 | 0.0252 |
| Variable 2 | 1000 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9590943 | 0.0156626 | 0.96432 | 0.9592 | 0.02 |
| Variable 3 | 1000 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9650468 | 0.0127662 | 0.96338 | 0.9648 | 0.0165 |
| Variable 4 | 1000 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9697368 | 0.0106768 | 0.96291 | 0.9688 | 0.014 |
| Variable 5 | 998 | 2 | 0 | 0 | 0 | 0 | 0 | 0.9550993 | 0.0169439 | 0.96262 | 0.9542 | 0.0228 |
| Variable 6 | 198 | 799 | 3 | 0 | 0 | 0 | 0 | 0.8776744 | 0.0227051 | 0.85824 | 0.8776 | 0.0345 |
| Variable 7 | 131 | 867 | 2 | 0 | 0 | 0 | 0 | 0.8727645 | 0.0253073 | 0.85786 | 0.8724 | 0.0327 |
| Variable 8 | 76 | 898 | 26 | 0 | 0 | 0 | 0 | 0.8555353 | 0.0297711 | 0.85758 | 0.854 | 0.0374 |
| Variable 9 | 36 | 890 | 74 | 0 | 0 | 0 | 0 | 0.8546053 | 0.0309861 | 0.85736 | 0.8432 | 0.0402 |
| Variable 10 | 10 | 852 | 138 | 0 | 0 | 0 | 0 | 0.8438458 | 0.030251 | 0.85718 | 0.832 | 0.0408 |
| Variable 11 | 0 | 392 | 602 | 6 | 0 | 0 | 0 | 0.7909693 | 0.0336918 | 0.76224 | 0.79 | 0.0451 |
| Variable 12 | 14 | 210 | 776 | 0 | 0 | 0 | 0 | 0.765057 | 0.030239 | 0.76208 | 0.826 | 0.0413 |
| Variable 13 | 0 | 201 | 764 | 35 | 0 | 0 | 0 | 0.7691197 | 0.036349 | 0.76194 | 0.7692 | 0.0468 |
| Variable 14 | 0 | 120 | 838 | 42 | 0 | 0 | 0 | 0.7603243 | 0.0356522 | 0.76182 | 0.7608 | 0.0475 |
| Variable 15 | 0 | 183 | 784 | 33 | 0 | 0 | 0 | 0.7672161 | 0.0366547 | 0.76171 | 0.7684 | 0.0477 |
| Variable 16 | 0 | 0 | 116 | 766 | 118 | 0 | 0 | 0.6493394 | 0.0425009 | 0.63982 | 0.648 | 0.0551 |
| Variable 17 | 0 | 0 | 25 | 620 | 150 | 5 | 0 | 0.6328624 | 0.0428943 | 0.63972 | 0.614 | 0.0565 |
| Variable 18 | 0 | 0 | 131 | 684 | 84 | 5 | 0 | 0.652602 | 0.0426586 | 0.63693 | 0.6516 | 0.055 |
| Variable 19 | 0 | 0 | 72 | 784 | 144 | 0 | 0 | 0.6661312 | 0.0441832 | 0.65766 | 0.6564 | 0.0527 |
| Variable 20 | 0 | 2 | 68 | 840 | 90 | 0 | 0 | 0.6327558 | 0.0425179 | 0.63949 | 0.642 | 0.0554 |
| Variable 21 | 0 | 0 | 139 | 827 | 140 | 34 | 0.55879 | 0.0375238 | 0.55756 | 0.554 | 0.0579 |
| Variable 22 | 0 | 0 | 202 | 772 | 25 | 0 | 0 | 0.5669312 | 0.041073 | 0.56739 | 0.5646 | 0.0578 |
| Variable 23 | 0 | 0 | 180 | 677 | 140 | 0 | 0 | 0.5643574 | 0.0418478 | 0.55745 | 0.572 | 0.0579 |
| Variable 24 | 0 | 0 | 1 | 106 | 750 | 143 | 0 | 0.5562991 | 0.0429173 | 0.55739 | 0.554 | 0.0578 |
| Variable 25 | 0 | 0 | 1 | 180 | 758 | 34 | 0 | 0.5615382 | 0.040217 | 0.55735 | 0.558 | 0.0579 |

Table 2.2: Result from simulated data for case control study (N=50:50). Summary table for 1000 bootstrap samples. Variable 1-variable5 has the true ROC AUC above 90, var6-var10 had the true auc >0.80 and <=0.90, var 11-var15 had the true auc >0.70 and <=0.80, var16-20 had the true auc >0.60 and <=0.70, var21-var25 had the true auc >0.50 and <=0.60, var26-var2000 had the true roc <0.50. Summary table for 1000 bootstrap samples.
in real medical research. In this simulation study, the predictors were statistically independent across predictor variables. A non-parametric ROCAUC was obtained for each predictor and for 1000 bootstrap samples. As our goal is screening the individual predictor, the correlation among the predictors would not change the ROCAUCs obtained for each predictor variable. We simulated the data for the sample size of 30 cases and 30 controls (Table 2.1), 50 cases and 50 controls (Table 2.2) and then for 100 cases and 100 controls (Table 2.3).

The results from the simulation data indicate that the mean of ROCAUCs from 1000 bootstrap samples is close to the true ROCAUC. Even with only 30 cases and 30 controls, 25 out of 25 listed predictor variables provide the correct level of classification ability by using mean of ROCAUCs from bootstrap samples. Using the single ROCAUC estimate from the original sample, 3 (more than 10%) predictor variables did not provide the correct level of classification ability. ROCAUC estimate varies across the bootstrap samples, even when the sample size is moderate (Figure 2). The quantification of variable importance based on a single ROCAUC estimate from the original sample thus is not stable and sensitive to the small change of data. Figure 3 illustrates the comparison of variability of a single ROCAUC and variability of mean of bootstrapped ROCAUCs. The standard error of mean of bootstrapped ROCAUCs was 20% to 50% smaller than the standard error of the single ROCAUC estimate from the original sample.

Simulation study II

In simulation study II, we look at the scenarios from the prospective cohort study when the disease outcome is time-dependent. Assuming a cohort of subjects at risk for certain disease, we collect biomarker values at baseline then follow the subjects for 5-year or until disease onset. The outcome of interest here is time to disease onset. We simulate the outcome variable following the exponential distribution with rate parameter \( \lambda = 0.10, 0.20 \) or \( 0.25 \). With this set up, approximately 40%, 60% and 70% of subjects will have disease onset by the end of study.

![Figure 2: Mean of Bootstrapped ROCAUCs vs. ROCAUC estimate from the original sample. Data from simulation Study I (N=30:30)](image-url)
Figure 3: The histogram of ROCAUC from simulation study 2.3 for Predictor 5, 10, 15, 20 and 25.

Note: The gray reference line is true ROCAUC for the predictor with the assumed bi-normal distribution.
Variable | Frequency with boot strapped ROCAUC >0.90 | Frequency with boot strapped ROCAUC >0.80 and ≤ 0.90 | Frequency with boot strapped ROCAUC >0.70 and ≤ 0.80 | Frequency with boot strapped ROCAUC >0.60 and ≤ 0.70 | Frequency with boot strapped ROCAUC >0.50 and ≤ 0.60 | Frequency with boot strapped ROCAUC <0.50 | Mean of Bootstrapped ROCAUC | Standard error of bootstrapped ROCAUC
---|---|---|---|---|---|---|---|---
Variable 1 | 999 | 1 | 0 | 0 | 0 | 0 | 0.9408271 | 0.0130895
Variable 2 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9469615 | 0.0125396
Variable 3 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9611843 | 0.0095612
Variable 4 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9593922 | 0.0101995
Variable 5 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9638983 | 0.0108851
Variable 6 | 2 | 980 | 18 | 0 | 0 | 0 | 0.8536174 | 0.0247766
Variable 7 | 10 | 912 | 78 | 0 | 0 | 0 | 0.8229696 | 0.0198079
Variable 8 | 6 | 939 | 55 | 0 | 0 | 0 | 0.9064319 | 0.0175536
Variable 9 | 0 | 960 | 40 | 0 | 0 | 0 | 0.8067283 | 0.0240097
Variable 10 | 20 | 970 | 10 | 0 | 0 | 0 | 0.8621651 | 0.0200097
Variable 11 | 0 | 1 | 980 | 19 | 0 | 0 | 0.7645985 | 0.0264636
Variable 12 | 0 | 0 | 968 | 32 | 0 | 0 | 0.7434315 | 0.0265167
Variable 13 | 0 | 37 | 961 | 2 | 0 | 0 | 0.7735275 | 0.0229168
Variable 14 | 0 | 18 | 982 | 0 | 0 | 0 | 0.7610192 | 0.0226087
Variable 15 | 8 | 920 | 72 | 0 | 0 | 0 | 0.7270790 | 0.0220545
Variable 16 | 0 | 9 | 960 | 50 | 0 | 0 | 0.6510493 | 0.0322842
Variable 17 | 0 | 0 | 930 | 844 | 46 | 0 | 0.5684901 | 0.0328242

Table 3.1: Result from simulated data for a prospective cohort study (N=200). The outcome follows the exponential distribution with lambda =0.1. Variable 1-variable5 has the true ROC AUC above 90, var6-var10 had the true auc >0.80 and ≤0.90. var11-var15 had the true auc >0.70 and ≤0.80, var16-20 had the true auc >0.60 and ≤ 0.80 and var21-var25 had the true auc <0.60 and ≥0.50, var26-var2000 had the true roc < 0.50. Summary table for 1000 bootstrap samples.

Variable | Frequency with boot strapped ROCAUC >0.90 | Frequency with boot strapped ROCAUC >0.80 and ≤ 0.90 | Frequency with boot strapped ROCAUC >0.70 and ≤ 0.80 | Frequency with boot strapped ROCAUC >0.60 and ≤ 0.70 | Frequency with boot strapped ROCAUC >0.50 and ≤ 0.60 | Frequency with boot strapped ROCAUC <0.50 | Mean of Bootstrapped ROCAUC | Standard error of bootstrapped ROCAUC
---|---|---|---|---|---|---|---|---
Variable 1 | 9998 | 2 | 0 | 0 | 0 | 0 | 0.9388571 | 0.0110693
Variable 2 | 9999 | 1 | 0 | 0 | 0 | 0 | 0.9533765 | 0.0184874
Variable 3 | 10000 | 0 | 0 | 0 | 0 | 0 | 0.9468926 | 0.0113075
Variable 4 | 9999 | 1 | 0 | 0 | 0 | 0 | 0.9465991 | 0.0127102
Variable 5 | 10000 | 0 | 0 | 0 | 0 | 0 | 0.9638535 | 0.0121829
Variable 6 | 4 | 970 | 26 | 0 | 0 | 0 | 0.8423335 | 0.0312943
Variable 7 | 6 | 984 | 10 | 0 | 0 | 0 | 0.8447030 | 0.0188419
Variable 8 | 0 | 0 | 980 | 844 | 46 | 0 | 0.5684901 | 0.0328242

Table 3.2: Result from simulated data for a prospective cohort study (N=200). The outcome follows the exponential distribution with lambda=0.20. Variable 1-variable5 has the true ROC AUC above 90, var6-var10 had the true auc >0.80 and <=0.90. var11-var15 had the true auc >0.70 and <=0.80, var16-20 had the true auc >0.60 and <=0.80 and var21-var25 had the true auc <0.60 and >0.50, var26-var2000 had the true roc < 0.50. Summary table for 1000 bootstrap samples.
at year 5. For predictor variables, we use the settings that are similar to those in simulation study I for diseased and non-diseased subjects. A time-dependent ROCAUC was obtained for each predictor and compared to those in simulation study I for diseased and non-diseased subjects.

Conclusion

We have demonstrated the procedure of feature selection based on the discriminative or predictive ability of variables via bootstrapped ROCAUCs. Filtering the variables can eliminate noise and alleviate the effect of the curse of dimensionality. The pre-selected variables can then be entered in the traditional low dimensional multivariate model for model building.

The proposed method has the unique advantage of quantifying the importance of candidate variables and is computationally much faster than any penalized or stepwise regression methods for variable selections. The evaluation of variables is based on the whole ROC curve rather than a single accuracy index. ROC curves have become ubiquitous in many application areas and the various advances have been discussed across published articles. Moreover, the uncertainty of variable importance is taken into consideration by computing the frequencies or mean of bootstrapped ROCAUC estimate values. Comparing to a single estimate of ROCAUC from the actual sample, the bootstrapped ROCAUC estimate is more robust as the sample variance or sample standard deviation, which are non-robust, can be greatly influenced by outliers. This type of variable selection is flexible and the criteria for selection may depend on the scientific objectives of the study. Another advantage of the proposed method is that we can select the variables by evaluating their predicting power at different time point when the outcome variable is time dependent.

The shortcoming of the proposed method is the high computational cost, which is a problem for other current feature selection methods as well. However, with the advancement in computer technology, computational time should not constitute a substantial drawback relative to other approaches to feature selections.

The above method can be readily used to many applications such as new biomarker discovery when the outcome is time dependent, and/or microarray studies that are aimed to explore a large pool of genes and select a subset of genes that are differentially expressed. This information may help provide insight regarding the discriminative ability and/or predictive ability of individual predictor variables and develop more specific treatment strategies for patients. It is particular useful for screening the variables in the bioinformatics studies when the amount of variables is extremely large and the number of subjects is comparatively very small.

References

1. Zhang HH (2008) Discussion of “Sure Independence Screening for Ultra-High Dimensional Feature Space. J R Stat Soc Series B Stat Methodol 70: 849-911.
2. Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognition Letters 31: 2225-2236.
3. Pepe MS, Longton G, Anderson GL, Schummer M (2003) Selecting differentially expressed genes from microarray experiments. Biometrics 59: 133-142.
4. Jeffery IB, Higgins DG, Culhane AC (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 7: 359.
5. Boulesteix AL, Slawski M (2009) Stability and aggregation of ranked gene lists. Brief Bioinform 10: 556-568.

| Variable | Frequency with boot strapped ROCAUC >0.90 | Frequency with boot strapped ROCAUC 0.80 and ≤ 0.90 | Frequency with boot strapped ROCAUC >0.70 and ≤ 0.80 | Frequency with boot strapped ROCAUC >0.60 and ≤ 0.70 | Frequency with boot strapped ROCAUC >0.50 and ≤ 0.60 | Frequency with boot strapped ROCAUC <0.50 | Mean of bootstrapped ROCAUC | Standard error of bootstrapped ROCAUC |
|----------|------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| Variable 1 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9378468 | 0.0110584 |
| Variable 2 | 999 | 1 | 0 | 0 | 0 | 0 | 0.9640356 | 0.0158319 |
| Variable 3 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.9583591 | 0.0143344 |
| Variable 4 | 970 | 30 | 0 | 0 | 0 | 0 | 0.8613456 | 0.0244897 |
| Variable 5 | 1000 | 0 | 0 | 0 | 0 | 0 | 0.8954376 | 0.0215438 |
| Variable 6 | 26 | 970 | 4 | 0 | 0 | 0 | 0.7496025 | 0.0251298 |
| Variable 7 | 30 | 945 | 29 | 0 | 0 | 0 | 0.7054376 | 0.0202797 |
| Variable 8 | 0 | 986 | 14 | 0 | 0 | 0 | 0.8569606 | 0.02288 |
| Variable 9 | 100 | 880 | 20 | 0 | 0 | 0 | 0.8773767 | 0.0203976 |
| Variable 10 | 42 | 870 | 88 | 0 | 0 | 0 | 0.7496025 | 0.0251298 |
| Variable 11 | 0 | 50 | 30 | 0 | 0 | 0 | 0.7761398 | 0.0258953 |
| Variable 12 | 0 | 100 | 80 | 0 | 0 | 0 | 0.8773767 | 0.0203976 |
| Variable 13 | 0 | 25 | 866 | 138 | 1 | 0 | 0.747269 | 0.0382842 |
| Variable 14 | 0 | 50 | 940 | 10 | 0 | 0 | 0.7621768 | 0.027776 |
| Variable 15 | 0 | 970 | 30 | 0 | 0 | 0 | 0.6537729 | 0.0263353 |
| Variable 16 | 0 | 100 | 854 | 46 | 0 | 0 | 0.6864764 | 0.0338212 |
| Variable 17 | 0 | 20 | 790 | 190 | 0 | 0 | 0.620076 | 0.0302598 |
| Variable 18 | 0 | 104 | 840 | 55 | 1 | 0 | 0.6858863 | 0.033131 |
| Variable 19 | 0 | 12 | 768 | 220 | 0 | 0 | 0.6199021 | 0.0395194 |
| Variable 20 | 0 | 0 | 680 | 20 | 0 | 0 | 0.5817297 | 0.0334554 |
| Variable 21 | 0 | 27 | 750 | 233 | 223 | 0 | 0.5449902 | 0.0352085 |
| Variable 22 | 0 | 164 | 684 | 152 | 0 | 0 | 0.5643352 | 0.034332 |
| Variable 23 | 0 | 30 | 780 | 190 | 0 | 0 | 0.5302325 | 0.0363153 |
| Variable 24 | 0 | 74 | 866 | 66 | 0 | 0 | 0.5668155 | 0.0369776 |

Table 3.3: Result from simulated data for a prospective cohort study (N=200). The outcome follows the exponential distribution with lambda=0.25. Variable 1-variable5 has the true ROCAUC above 90, var6-var10 had the true auc >0.80 and <=0.90, var11-var15 had the true auc >0.70 and ≤ 0.80, var16-20 had the true auc >0.60 and ≤ 0.80 and var21-var25 had the true auc <0.60 and ≥0.50, var26-var200 had the true roc < 0.50. Frequency with boot strapped ROCAUC >0.80 and ≤ 0.90, Frequency with boot strapped ROCAUC >0.70 and ≤ 0.80, Frequency with boot strapped ROCAUC >0.60 and ≤ 0.70, Frequency with boot strapped ROCAUC >0.50 and ≤ 0.60, Frequency with boot strapped ROCAUC <0.50, Mean of bootstrapped ROCAUC, Standard error of bootstrapped ROCAUC.
6. Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56: 337-344.

7. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29-36.

8. van de Vijver MJ, He YD, van’t Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.