Journal of
Proteomics & Bioinformatics

Xu et al., J Proteomics Bioinform 2014, S9
DOI: 10.4172/jpb.S9-006

Research Article Open Access

Feature Selection using Bootstrapped ROC Curves

Ping Xu[1]*, Xiang Liu[1], David Hadley[1#], Shuai Huang[2], Jeffrey Krischer[1] and Craig Beam[3]

[1]Department of Pediatrics, College of Medicine, University of South Florida, 3650 Spectrum Blvd, Suite 100, Tampa, Florida, USA
[2]Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98195, USA
[3]Department of Biomedical Sciences, West Michigan University, Kalamazoo, MI 49008, USA
[#]Current 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 is 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

***Corresponding author:** Ping Xu, Department of Pediatrics, College of Medicine, University of South Florida, 3650 Spectrum Blvd, Suite 100, Tampa, Florida, USA, Tel: (813) 3969552; Fax: (813) 9105952; E-mail: Ping.Xu@epi.usf.edu

Received September 10, 2014; **Accepted** October 21, 2014; **Published** October 24, 2014

Citation: Xu P, Liu X, Hadley D, Huang S, Krischer J, et al. (2014) Feature Selection using Bootstrapped ROC Curves. J Proteomics Bioinform S9: 006. doi:10.4172/jpb.S9-006

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 Rocaucs

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:

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

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:

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

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 Ti and Ci denote survival and censoring times for ith subject, We observe (Zi, δi) where Zi=min(Ti, Ci) and δi=I (Ti ≤ Ci). Denote Di (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:

$$AUC(t_0) = \int_{-\infty}^{+\infty} TPR(c,t_0)d[FPR(c,t_o)]$$

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 monotonic 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 values. 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 Heagerty et al. [6] for survival outcome or the paper for binary outcome by Hanley and McNeil [7].

ROCAUC estimates from bootstrap resampling

Let $S=\{(D_i, Y_i), i=1, 2....n\}$ denote a sample of N independent subjects, who have been measured with m continuous independent variables $y_i=(y_{i1}, y_{i2},....y_{im})$ 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,

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

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

Draw B random samples S*(1), ...,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 >0.80 and <=0.90	Frequency with ROCAUC value >0.70 and <=0.80	Frequency with ROCAUC value >0.60 and <=0.70	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):

$Y_{ROCAUC}^{(1)}, Y_{ROCAUC}^{(2)} \ldots\ldots Y_{ROCAUC}^{(m)}$.

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.

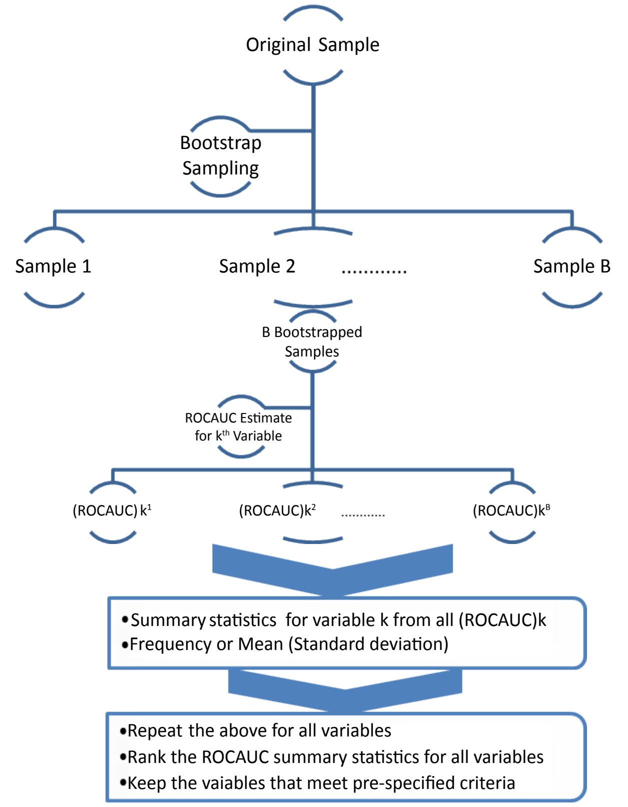

Original Sample
Bootstrap Sampling
Sample 1
Sample 2
....
Sample B
B Bootstrapped Samples
ROCAUC Estimate for k^th Variable
(ROCAUC) k^1
(ROCAUC)k^2
....
(ROCAUC)k^B
•Summary statistics for variable k from all (ROCAUC)k
•Frequency or Mean (Standard deviation)
•Repeat the above for all variables
•Rank the ROCAUC summary statistics for all variables
•Keep the vaiables that meet pre-specified criteria

Figure 1: Flow Chart for Feature Selection Using Bootstrapped ROCAUC Estimate.

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, $Y_{ROCAUC}^{(1)}, Y_{ROCAUC}^{(2)} \ldots\ldots Y_{ROCAUC}^{(p)}$. When we have little information on the disease, it may be appropriate to simply keep any variable which has fair 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 furthered investigated to improve the overall predicative 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. When the setting satisfies 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:

$$ROCAUC = \phi\left\{\frac{a}{\sqrt{1+b^2}}\right\}$$

where Φ is the standard normal cumulative distribution function, $a = \frac{\mu_{(\rho+)} - \mu_{(\rho-)}}{\sigma_{(\rho+)}}$, and $b = \frac{\sigma_{(\rho-)}}{\sigma_{(\rho+)}}$

Substance/gene	**ROCAUC**						
Substance/gene **N**	**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**	**Std Dev**
NM_003600/STK15	2	627	308	52	11	0.7026278	0.0639113
AF108138	7	569	378	39	7	0.6985829	0.0588761
NM_003920	5	568	250	157	20	0.6814975	0.0832337
NM_002497/NEK2	0	253	679	51	17	0.669793	0.0657337
NM_021000/PTTG3	0	169	740	72	19	0.6584086	0.0612181
NM_004553/NDUFS6	0	155	678	148	19	0.6529875	0.0546397
NM_012484/HMMR	0	64	803	63	70	0.6329165	0.0724397
NM_003878/GGH	0	32	773	190	5	0.6284048	0.0487332
NM_006067/NOC4	0	58	687	232	23	0.6269651	0.0634367
NM_014264/STK18	0	52	659	262	27	0.6250741	0.0573268
NM_002180/IGHMBP2	0	23	636	319	22	0.6130325	0.0553191
Contig55216_RC	0	3	584	403	10	0.6016745	0.0475804
NM_007006/CFIM25	0	5	587	351	57	0.597281	0.0629893
NM_005147/TID1	0	4	563	399	34	0.5959746	0.067799
AL117635/DKFZP434G145	0	18	613	271	98	0.5909529	0.067462

Table 1: Result from application data for top 15 substance/genes. Summary table from 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 Boot strapped ROCAUC**	**Standard error of boot strapped ROCAUC**	**True ROCAUC**	**ROCAUC from the original sample**	**Standard errof of ROCAUC from the original sample**
Variable 1	847	153	0	0	0	0	0.9268456	0.0256736	0.967	0.9256	0.0333
Variable 2	1000	0	0	0	0	0	0.9540883	0.019081	0.96432	0.9544	0.0244
Variable 3	1000	0	0	0	0	0	0.9597861	0.0210378	0.96338	0.96	0.0279
Variable 4	957	43	0	0	0	0	0.9398364	0.0224956	0.96291	0.94	0.0288
Variable 5	1000	0	0	0	0	0	0.9857206	0.0090068	0.96262	0.9856	0.0113
Variable 6	257	735	8	0	0	0	0.8791552	0.0330017	0.85824	0.8778	0.043
Variable 7	54	737	209	0	0	0	0.8345433	0.0405639	0.85786	0.8333	0.0527
Variable 8	25	560	410	5	0	0	0.811072	0.0447755	0.85758	0.7942	0.0582
Variable 9	452	547	1	0	0	0	0.89581	0.0317682	0.85736	0.8956	0.0398
Variable 10	231	749	20	0	0	0	0.8727911	0.0360695	0.85718	0.8722	0.0477
Variable 11	1	300	652	47	0	0	0.7756094	0.0450083	0.76224	0.8178	0.0598
Variable 12	8	274	653	65	0	0	0.773865	0.0485094	0.76208	0.7711	0.0628
Variable 13	2	201	695	101	1	0	0.7611366	0.0481824	0.76194	0.7611	0.0611
Variable 14	5	305	627	62	1	0	0.7771825	0.048738	0.76182	0.7756	0.06
Variable 15	2	95	664	234	5	0	0.7366996	0.0500578	0.76171	0.7144	0.0654
Variable 16	0	0	124	601	265	10	0.6335688	0.056518	0.63982	0.6344	0.0732
Variable 17	0	1	230	638	126	5	0.6614763	0.0536198	0.63972	0.7056	0.0709
Variable 18	0	10	261	631	97	1	0.6683102	0.0543015	0.63963	0.6667	0.0707
Variable 19	0	25	399	522	54	0	0.6909663	0.0546546	0.63956	0.6922	0.0693
Variable 20	0	3	259	619	119	0	0.6646633	0.0528927	0.63949	0.6633	0.0706
Variable 21	0	0	23	381	510	86	0.5867278	0.0573819	0.55756	0.5856	0.0752
Variable 22	0	0	14	319	616	51	0.5790827	0.0504566	0.5575	0.5756	0.075
Variable 23	0	0	8	139	710	143	0.5602071	0.0470166	0.55745	0.5622	0.0753
Variable 24	0	1	20	193	725	61	0.5573308	0.0539114	0.55739	0.5756	0.0751
Variable 25	0	0	7	265	663	65	0.5629696	0.0500613	0.55735	0.5656	0.0753

Table 2.1: Result from simulated data for case control study (N=30:30). 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 ≤0.80 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.

Simulation study I

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-variable 5) have true ROCAUC over 0.90, 5 predictors (variable 6-variable 10) have ROCAUC>.80 and ≤ 0.90, 5 predictors (variable11-variable 15) have ROCAUC>0.70 and ≤ 0.80, 5 predictors (variable16-variable20) have ROCAUC>0.60 and ≤ 0.70, 5 predictors have ROCAUC>0.50 and ≤ 0.60 (variable21-variable 25), and rest of predictors (variable 26- variable2000) have ROCAUC ≤ 0.50. The predictor values are simulated from a normal distribution for both cases and controls with **σ**=1 for both cases and controls. The means in controls were set to 0s and the different level of mean in cases was considered, representing a variety of AUC level encountered

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 Boot strapped ROCAUC	Standard error of boot strapped ROCAUC	True ROCAUC	ROCAUC from the original sample	Standard errof of ROCAUC from the original sample
Variable 1	949	51	0	0	0	0	0.9302268	0.0193489	0.967	0.9292	0.0252
Variable 2	1000	0	0	0	0	0	0.9590943	0.0156626	0.96432	0.9592	0.02
Variable 3	1000	0	0	0	0	0	0.9650468	0.0127662	0.96338	0.9648	0.0165
Variable 4	1000	0	0	0	0	0	0.9687368	0.0106768	0.96291	0.9688	0.014
Variable 5	998	2	0	0	0	0	0.9550993	0.0169439	0.96262	0.9452	0.0228
Variable 6	198	799	3	0	0	0	0.8770674	0.0270519	0.85824	0.8776	0.0345
Variable 7	131	867	2	0	0	0	0.8727645	0.0253073	0.85786	0.8724	0.0337
Variable 8	76	898	26	0	0	0	0.8555353	0.029771	0.85758	0.854	0.0374
Variable 9	36	890	74	0	0	0	0.8546053	0.0309861	0.85736	0.8432	0.0402
Variable 10	10	852	138	0	0	0	0.8438458	0.030251	0.85718	0.832	0.0408
Variable 11	0	392	602	6	0	0	0.7909963	0.0336918	0.76224	0.79	0.0451
Variable 12	14	210	776	0	0	0	0.76507	0.0320329	0.76208	0.826	0.0413
Variable 13	0	201	764	35	0	0	0.7691197	0.036349	0.76194	0.7692	0.0468
Variable 14	0	120	838	42	0	0	0.7603243	0.0355622	0.76182	0.7608	0.0475
Variable 15	0	183	784	33	0	0	0.7672161	0.0366547	0.76171	0.7684	0.0477
Variable 16	0	0	116	766	118	0	0.6493394	0.0425009	0.63982	0.648	0.0551
Variable 17	0	0	25	820	150	5	0.6328624	0.0426943	0.63972	0.614	0.0565
Variable 18	0	0	131	864	5	0	0.652602	0.0426586	0.63963	0.6516	0.055
Variable 19	0	0	72	784	144	0	0.6661312	0.0414832	0.63956	0.6668	0.0544
Variable 20	0	2	68	840	90	0	0.6327558	0.0425179	0.63949	0.6428	0.0554
Variable 21	0	0	0	139	827	34	0.55879	0.0375238	0.55756	0.554	0.0579
Variable 22	0	0	1	202	772	25	0.5669316	0.0387936	0.5575	0.5648	0.0578
Variable 23	0	0	3	180	677	140	0.5643574	0.0418478	0.55745	0.572	0.0 579
Variable 24	0	0	1	106	750	143	0.5562991	0.0421073	0.55739	0.5644	0.0578
Variable 25	0	0	1	180	785	34	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 ≤ 0.80 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 Boot strapped ROCAUC	Standard error of boot strapped ROCAUC	True ROCAUC	ROCAUC from the original sample	Standard errof of ROCAUC from the original sample
Variable 1	1000	0	0	0	0	0	0.9610616	0.0063064	0.967	0.9478	0.0153
Variable 2	1000	0	0	0	0	0	0.9516256	0.0062665	0.96432	0.9678	0.0104
Variable 3	1000	0	0	0	0	0	0.9568748	0.008823	0.96338	0.9605	0.0129
Variable 4	1000	0	0	0	0	0	0.9602355	0.0062441	0.96291	0.9669	0.0108
Variable 5	1000	0	0	0	0	0	0.967928	0.008082	0.96262	0.9478	0.0149
Variable 6	70	923	7	0	0	0	0.8640857	0.0124244	0.85824	0.8986	0.0219
Variable 7	26	973	1	0	0	0	0.8897361	0.0128078	0.85786	0.8627	0.025
Variable 8	47	952	1	0	0	0	0.8579662	0.0154984	0.85758	0.8658	0.0254
Variable 9	0	835	165	0	0	0	0.8280122	0.0168573	0.85736	0.7926	0.0291
Variable 10	81	919	0	0	0	0	0.873651	0.018986	0.85718	0.8732	0.025
Variable 11	0	40	935	25	0	0	0.7404198	0.0199555	0.76224	0.7553	0.0337
Variable 12	0	109	881	10	0	0	0.7542329	0.0200538	0.76208	0.7676	0.0335
Variable 13	0	23	944	33	0	0	0.781852	0.0173081	0.76194	0.746	0.0344
Variable 14	0	3	833	164	0	0	0.7241846	0.0221302	0.76182	0.718	0.0351
Variable 15	0	31	954	15	0	0	0.751869	0.025456	0.76171	0.7324	0.0338
Variable 16	0	0	27	897	76	0	0.6566153	0.0220205	0.63982	0.642	0.0391
Variable 17	0	0	16	874	110	0	0.6564672	0.0241737	0.63972	0.6375	0.039
Variable 18	0	0	17	865	118	0	0.6556744	0.0248758	0.63963	0.6334	0.0393
Variable 19	0	0	135	849	16	0	0.6681345	0.0180067	0.63956	0.6669	0.0381
Variable 20	0	0	12	863	125	0	0.634592	0.029759	0.63949	0.6324	0.0392
Variable 21	0	0	0	79	905	16	0.5803688	0.0237889	0.55756	0.5566	0.0407
Variable 22	0	0	0	139	847	14	0.5535859	0.020129	0.5575	0.5691	0.0406
Variable 23	0	0	0	72	813	115	0.5435917	0.0242432	0.55745	0.5476	0.0407
Variable 24	0	0	0	68	825	107	0.538754	0.0189875	0.55739	0.5556	0.0407
Variable 25	0	0	0	84	904	12	0.556578	0.030138	0.55735	0.5557	0.0407

Table 2.3: 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 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 <=0.80 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.

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 λ=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

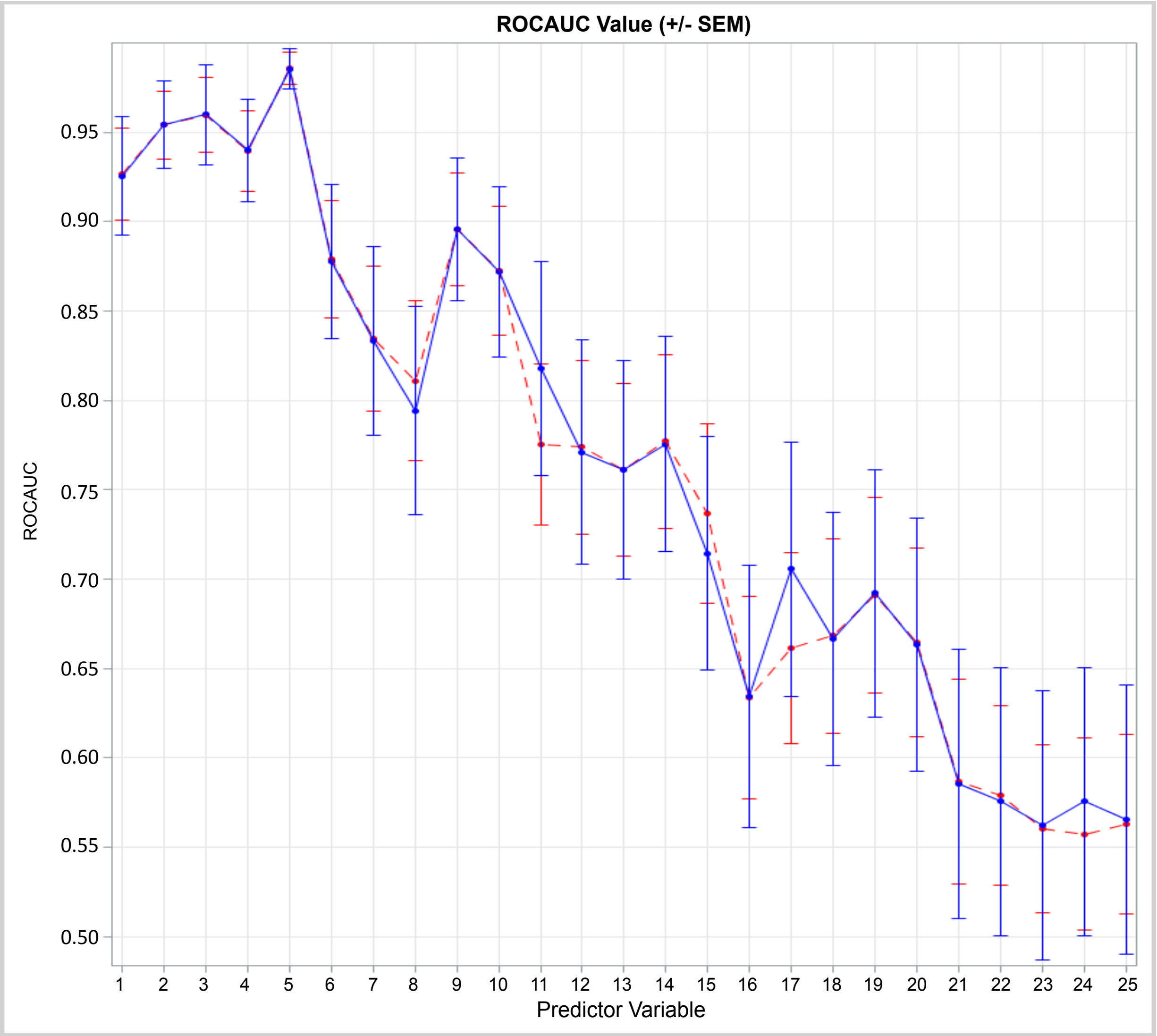
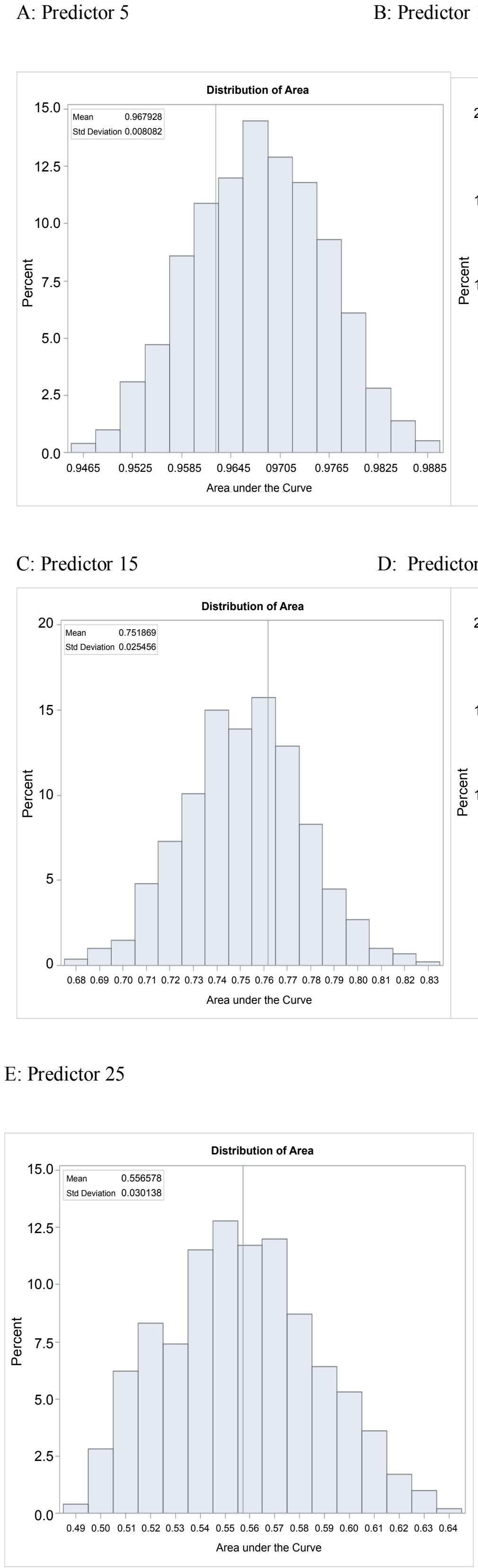


Figure 2: Mean of Bootstrapped ROCAUCs vs. ROCAUC estimate from the original sample. Data from simulation Study I (N=30:30)



A: Predictor 5

B: Predictor 10

C: Predictor 15

D: Predictor 20

E: Predictor 25

Note: The gray reference line is true ROCAUC for the predictor with the assumed bi-normal distribution.

Figure 3: The histogram of ROCAUC from simulation study 2.3 for Predictor 5, 10, 15, 20 and 25.

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	0.9466915	0.0125396
Variable 3	1000	0	0	0	0	0	0.9611843	0.0095612
Variable 4	1000	0	0	0	0	0	0.9593922	0.0101955
Variable 5	1000	0	0	0	0	0	0.9639893	0.0100851
Variable 6	2	980	18	0	0	0	0.8536174	0.0242776
Variable 7	10	912	78	0	0	0	0.8226946	0.0198079
Variable 8	6	939	55	0	0	0	0.9064319	0.0167553
Variable 9	0	960	40	0	0	0	0.8067283	0.0240097
Variable 10	20	970	10	0	0	0	0.8621651	0.020009
Variable 11	0	1	980	19	0	0	0.7645985	0.0266463
Variable 12	0	0	968	32	0	0	0.7434315	0.0264516
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	0	8	920	72	0	0	0.727079	0.0220544
Variable 16	0	0	90	860	50	0	0.6510493	0.0321351
Variable 17	0	0	0	830	150	20	0.5611356	0.0320006
Variable 18	0	0	18	890	92	0	0.6629974	0.0286322
Variable 19	0	0	10	839	151	0	0.6093222	0.0329304
Variable 20	0	0	40	860	100	0	0.6431546	0.0331465
Variable 21	0	0	0	280	720	0	0.5763499	0.0319254
Variable 22	0	0	0	110	829	61	0.5683108	0.0330958
Variable 23	0	0	1	47	752	200	0.5422889	0.0339387
Variable 24	0	0	0	10	610	380	0.5083215	0.0337083
Variable 25	0	0	0	110	844	46	0.5684901	0.0322842

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	999	1	0	0	0	0	0.9533765	0.0184874
Variable 3	1000	0	0	0	0	0	0.9468926	0.0113075
Variable 4	999	1	0	0	0	0	0.9465991	0.0127102
Variable 5	1000	0	0	0	0	0	0.9638535	0.0121829
Variable 6	4	970	26	0	0	0	0.8423335	0.0212838
Variable 7	6	984	10	0	0	0	0.844703	0.0188419
Variable 8	40	960	0	0	0	0	0.865546	0.0168259
Variable 9	90	900	10	0	0	0	0.8888766	0.016785
Variable 10	43	956	0	0	0	0	0.8693859	0.0168756
Variable 11	0	0	950	50	0	0	0.7583765	0.0311731
Variable 12	0	78	921	1	0	0	0.766182	0.0222788
Variable 13	0	84	908	8	0	0	0.7876855	0.0266457
Variable 14	0	0	931	69	0	0	0.7660338	0.0305138
Variable 15	0	70	910	20	0	0	0.7607563	0.0260077
Variable 16	0	0	140	830	30	0	0.6623587	0.0324408
Variable 17	0	0	7	798	195	0	0.6185979	0.0332483
Variable 18	0	0	0	869	110	0	0.6339894	0.031044
Variable 19	0	0	10	835	155	0	0.620362	0.0292556
Variable 20	0	0	114	874	12	0	0.6626244	0.0311567
Variable 21	0	0	1	407	592	0	0.592236	0.0324113
Variable 22	0	0	1	374	625	0	0.5843459	0.0309203
Variable 23	0	0	0	25	848	127	0.5354593	0.0265489
Variable 24	0	0	0	278	718	4	0.5804723	0.0331337
Variable 25	0	0	0	22	840	138	0.5356309	0.0395772

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.

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	0.9522958	0.0132141
Variable 4	970	30	0	0	0	0	0.9175244	0.0157655
Variable 5	1000	0	0	0	0	0	0.9583591	0.0143344
Variable 6	26	970	4	0	0	0	0.8613456	0.0244897
Variable 7	30	960	10	0	0	0	0.864229	0.0207975
Variable 8	26	945	29	0	0	0	0.8548736	0.0215438
Variable 9	0	986	14	0	0	0	0.8569606	0.02288
Variable 10	100	880	20	0	0	0	0.8773767	0.0203976
Variable 11	0	42	870	88	0	0	0.74960253	0.0251298
Variable 12	0	50	930	20	0	0	0.7674741	0.027453
Variable 13	0	100	880	20	0	0	0.776139	0.0258953
Variable 14	0	25	836	138	1	0	0.747269	0.0382842
Variable 15	0	50	940	10	0	0	0.7621768	0.0277766
Variable 16	0	0	0	970	30	0	0.6537729	0.0268335
Variable 17	0	0	100	854	46	0	0.6864764	0.0338212
Variable 18	0	0	20	790	190	0	0.6260076	0.0302598
Variable 19	0	0	104	840	55	1	0.6855863	0.033131
Variable 20	0	0	12	768	220	0	0.6199021	0.033914
Variable 21	0	0	0	200	720	80	0.5817297	0.0334054
Variable 22	0	0	0	27	750	223	0.5449902	0.0352085
Variable 23	0	0	0	164	684	152	0.5643352	0.0343332
Variable 24	0	0	0	30	780	190	0.5302325	0.0363153
Variable 25	0	0	0	74	860	66	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 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 repeated for 1000 bootstrap samples. Tables 3.1-3.3 summarize the results from the simulated data with the sample size of 200 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.

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.

This article was originally published in a special issue, **Computational Intelligence in Bioinformatics** handled by Editor(s). Dr. Jean-Christophe Nebel, Kingston University, London, UK