High-dimensional Classification Using Features Annealed Independence Rules

Ganchao Wei

November 10, 2021

Overview

Introduction

- Classical methods of classification break down when the dimensionality is extremely large
- The difficulty is intrinsically caused by the existence of many noise features that don't contribute to the reduction of mis-classification rate
- When the dimensionality is high, the aggregated estimation error can be very large.
- In this paper, they proposed feature annealed independent rules (FAIR), which can extract all important features, and overcome both the issues of interpretability and the noise accumulation.

Impact of High Dimensionality

Consider the p-dimensional classification problem between 2 classes C_k, $k = 1, 2$. Each class has n_k observations. Assume the observations follow the model:

$$\boldsymbol{Y}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_{ki}, \quad k = 1, 2, i = 1, \ldots, n_k$$

, where $\boldsymbol{\mu}_k$ is the mean vector of class C_k and $\boldsymbol{\epsilon}_{ki}$ has the distribution $N(0, \boldsymbol{\Sigma}_k)$. Assume

- The 2 classes have compatible sample sizes, i.e., $c_1 \leq n_1/n_2 \leq c_2$.
- 2 have the same covariance matrix $\boldsymbol{\Sigma}$

Consider the independence classification rule: classify as C_1 if

$$\delta(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{D}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$$

, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ and $\boldsymbol{D} = diag(\boldsymbol{\Sigma})$. The parameters can be estimated from samples as:

- $\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n_k} \boldsymbol{Y}_{ki}/n_k$
- $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$
- $\hat{\boldsymbol{D}} = diag\{(S_{1j}^2 + S_{2j}^2)/2, j = 1, \ldots, p\}$

Impact of High Dimensionality

Then the plug-in discriminant function is $\hat{\delta}(\boldsymbol{x})$. Denote the parameter by $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. If a new observation $\boldsymbol{X}$ from class C_1, then the misclassification rate is

$$W(\hat{\delta}, \boldsymbol{\theta}) = P\big(\hat{\delta}(\mathbf{X}) \le 0 | \mathbf{Y}_{ki}, i = 1, \ldots, n_k,\ k = 1, 2\big) = 1 - \Phi(\Psi), \tag{2.2}$$

where

$$\Psi = \frac{(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})' \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{\sqrt{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)' \widehat{\mathbf{D}}^{-1} \boldsymbol{\Sigma} \widehat{\mathbf{D}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}},$$

and $\Phi(\cdot)$ is the standard Gaussian distribution function. The worst case classification error is

$$W(\hat{\delta}) = \max_{\boldsymbol{\theta} \in \Gamma} W(\hat{\delta}, \boldsymbol{\theta}),$$

Impact of High Dimensionality

Let $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\Sigma\boldsymbol{D}^{-1/2}$ be the correlation matrix, $\lambda_{max}(\boldsymbol{R})$ be its largest eigenvalue, and $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Consider the parameter space:

$$\Gamma = \left\{(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) : \boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha} \geq C_p, \lambda_{\max}(\mathbf{R}) \leq b_0, \min_{1\leq j\leq p, k=1,2} \sigma_{kj}^2 > 0\right\}$$

- First term: imposes a lower bound on the strength of signals
- Second term: requires the maximum eigenvalue of $\boldsymbol{R}$ is upper bounded. (But there's no lower bound, the condition number can still diverge)
- Third term: ensures that there are no deterministic features that make classification trivial and $\boldsymbol{D}$is always invertible.

Then consider the asymptotic behavior of $W(\hat{\delta}, \boldsymbol{\theta})$ and $W(\hat{\delta})$.

Impact of High Dimensionality

THEOREM 1. *Suppose that* $\log p = o(n)$, $n = o(p)$ *and* $nC_p \to \infty$. *Then*:

(i) *The classification error* $W(\delta, \boldsymbol{\theta})$ *with* $\boldsymbol{\theta} \in \Gamma$ *is bounded from above as*

$$W(\hat{\delta}, \boldsymbol{\theta}) \le 1 - \Phi\left(\frac{[n_1 n_2/(pn)]^{1/2} \boldsymbol{\alpha}' \mathbf{D}^{-1} \boldsymbol{\alpha}(1 + o_P(1)) + \sqrt{p/(n n_1 n_2)}(n_1 - n_2)}{2\sqrt{\lambda_{\max}(\mathbf{R})}\{1 + n_1 n_2/(pn) \boldsymbol{\alpha}' \mathbf{D}^{-1} \boldsymbol{\alpha}(1 + o_P(1))\}^{1/2}}\right).$$

(ii) *Suppose* $p/(nC_p) \to 0$. *For the worst case classification error* $W(\delta)$, *we have*

$$W(\hat{\delta}) = 1 - \Phi\left(\tfrac{1}{2}[n_1 n_2/(pnb_0)]^{1/2} C_p\{1 + o_P(1)\}\right).$$

Specifically, when $\{\frac{n_1 n_2}{pn}\}^{1/2} C_p \to C_0$ *with* C_0 *a nonnegative constant, then*

$$W(\hat{\delta}) \xrightarrow{P} 1 - \Phi(C_0/(2\sqrt{b_0})).$$

In particular, if $C_0 = 0$, *then* $W(\hat{\delta}) \xrightarrow{P} \frac{1}{2}$.

The independence rule $\hat{\delta}$ would be no better than the random guessing due to noise accumulation, unless the signal levels are extremely high.

Indeed, the discrimination based on linear projections to almost all directions performs nearly the same as random guessing (caused by noise accumulation in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$)

THEOREM 2. *Suppose that* $\mathbf{a}$ *is a* p*-dimensional uniformly distributed unit random vector on a* $(p-1)$*-dimensional sphere. Let* $\lambda_1, \ldots, \lambda_p$ *be the eigenvalues of the covariance matrix* $\boldsymbol{\Sigma}$. *Suppose* $\lim_p \frac{1}{p^2} \sum_{j=1}^p \lambda_j^2 < \infty$ *and* $\lim_p \frac{1}{p} \sum_{j=1}^p \lambda_j = \tau$ *with* τ *a positive constant. Moreover, assume that* $p^{-1}\boldsymbol{\alpha}'\boldsymbol{\alpha} \to 0$. *Then if we project all the observations onto the vector* $\mathbf{a}$ *and use the classifier*

$$\hat{\delta}_{\mathbf{a}}(\mathbf{x}) = (\mathbf{a}'\mathbf{x} - \mathbf{a}'\widehat{\boldsymbol{\mu}})(\mathbf{a}'\widehat{\boldsymbol{\mu}}_1 - \mathbf{a}'\widehat{\boldsymbol{\mu}}_2), \tag{2.3}$$

the misclassification rate of $\hat{\delta}_{\mathbf{a}}$ *satisfies*

$$P(\hat{\delta}_{\mathbf{a}}(\mathbf{X}) \le 0 | \mathbf{Y}_{ki},\ i = 1, \ldots, n_k,\ k = 1, 2) \xrightarrow{P} \frac{1}{2},$$

where the probability is taken with respect to $\mathbf{a}$ *and* $\mathbf{X} \in \mathcal{C}_1$.

Feature Selection by two-sample t-test

The 2 sample t-statistic for feature j is defined as

$$T_j = \frac{\bar{Y}_{1j} - \bar{Y}_{2j}}{\sqrt{S_{1j}^2/n_1 + S_{2j}^2/n_2}}$$

Relax the normality assumption: just assume the noise vector ϵ_{ki} are *i.i.d.* within class with mean 0 and covariance $\mathbf{\Sigma}_k$, and are independent between classes. Consider the following condition:

CONDITION 1.

(a) Assume that the vector $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse and without loss of generality, only the first s entries are nonzero.

(b) Suppose that ϵ_{kij} and $\epsilon_{kij}^2 - 1$ satisfy the Cramér's condition, that is, there exist constants ν_1, ν_2, M_1 and M_2, such that $E|\epsilon_{kij}|^m \leq m! M_1^{m-2} \nu_1/2$ and $E|\epsilon_{kij}^2 - \sigma_{kj}^2|^m \leq m! M_2^{m-2} \nu_2/2$ for all $m = 1, 2, \ldots$.

(c) Assume that the diagonal elements of both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are bounded away from 0.

Feature Selection by two-sample t-test

Under the condition on the previous slide, we can show that the t-test can pick up all important features $w.p.1$. ($c_1 \leq n_1/n_2 \leq c_2$ and $n = n_1 + n_2$)

THEOREM 3. *Let s be a sequence such that $\log(p-s) = o(n^{\gamma})$ and $\log s = o(n^{1/2-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < \frac{1}{3}$. Suppose that $\min_{1\leq j\leq s} \frac{|\alpha_j|}{\sqrt{\sigma_{1j}^2+\sigma_{2j}^2}} = n^{-\gamma}\beta_n$. Then under Condition 1, for $x \sim cn^{\gamma/2}$ with c some positive constant, we have*

$$P\left(\min_{j\leq s}|T_j| \geq x \text{ and } \max_{j>s}|T_j| < x\right) \to 1.$$

Features Annealed Independence Rules

Just apply the independence classifier to the selected features $\rightarrow$ Featured Annealed Independence Rule (FAIR).

In many application, $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are sparse $\rightarrow$ the noise accumulation can exceed the signal accumulation for faint features $\rightarrow$ further single out the most important features to reduce misclassification rate, after using t-test.

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$, and is known, the independence classifier $\hat{\delta}$ becomes the nearest centroids classifier

$$\hat{\delta}_{NC}(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$$

If only the first m dimensions are used in the classification, the corresponding features annealed independence classifier becomes

$$\hat{\delta}^m_{NC}(\boldsymbol{x}) = (\boldsymbol{x}^m - \hat{\boldsymbol{\mu}}^m)'(\hat{\boldsymbol{\mu}}^m_1 - \hat{\boldsymbol{\mu}}^m_2)$$

Features Annealed Independence Rules

The classification error for the truncated classifier is:

THEOREM 4. *Consider the truncated classifier $\hat{\delta}_{\mathrm{NC}}^{m_n}$ for a given sequence m_n. Suppose that $\frac{n}{\sqrt{m_n}} \sum_{j=1}^{m_n} \alpha_j^2 \to \infty$ as $m_n \to \infty$. Then the classification error of $\hat{\delta}_{\mathrm{NC}}^{m_n}$ is*

$$W(\hat{\delta}_{\mathrm{NC}}^{m_n}, \boldsymbol{\theta}) = 1 - \Phi\left(\frac{(1 + o_P(1)) \sum_{j=1}^{m_n} \alpha_j^2 + m_n(n_1 - n_2)/(n_1 n_2)}{2\{(1 + o_P(1)) \sum_{j=1}^{m_n} \alpha_j^2 + n m_n/(n_1 n_2)\}^{1/2}}\right),$$

where $n = n_1 + n_2$ as defined in Section 2.

This theorem tells us that the ideal choice on the number of features is

$$m_0 = \underset{1 \le m \le p}{\arg\max} \frac{[\sum_{j=1}^{m} \alpha_j^2 + m(n_1 - n_2)/(n_1 n_2)]^2}{nm/(n_1 n_2) + \sum_{j=1}^{m} \alpha_j^2}$$

Features Annealed Independence Rules

When $n_1 = n_2$, the expression reduces to $m_0 = \mathrm{argmax}_{1 \le m \le p} \frac{(m^{-1/2}\sum_{j=1}^m \alpha_j^2)^2}{2/n + \sum_{j=1}^m \alpha_j^2/m}$. The term $m^{-1/2}\sum_{j=1}^m \alpha_j^2$ reflects the trade-off between the signal and noise as dimensionality m increases.

An ideal classifier $\hat{\delta}_{NC}$ is to select a subset $A = \{j : |\alpha_j| > a\}$ and use this subset to construct independence classifier. The oracle classifier can be written as

$$\hat{\delta}_{\mathrm{orc}}(\mathbf{x}) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j) 1_{\{|\alpha_j| > a\}}.$$

The misclassification rate is approximately

$$1 - \Phi\left(\frac{\sum_{j \in \mathcal{A}} \alpha_j^2 + m(n_1 - n_2)/(n_1 n_2)}{2\{nm/(n_1 n_2) + \sum_{j \in \mathcal{A}} \alpha_j^2\}^{1/2}}\right) \tag{4.1}$$

when $\frac{n}{\sqrt{m}} \sum_{j \in A} \alpha_j^2 \to \infty$ and $m \to \infty$.

Features Annealed Independence Rules

In practice, we have no such an oracle, and selecting the subset A is difficult. Then just use the estimator-plug-in version: FAIR based on the hard thresholding:

$$\hat{\delta}^b_{\text{FAIR}}(\mathbf{x}) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j) 1_{\{|\hat{\alpha}_j| > b\}}.$$

We study the classification error of FAIR and the impact of the threshold b on the classification result in the following theorem.

THEOREM 5. *Suppose that* $\max_{j \in \mathcal{A}^c} |\alpha_j| < b_n$ *and* $\log(p - m)/[n(b_n - \max_{j \in \mathcal{A}^c} |\alpha_j|)^2] \to 0$ *with* $m = |\mathcal{A}|$. *Moreover, assume that* $\frac{n}{\sqrt{m}} \sum_{j \in \mathcal{A}} \alpha_j^2 \to \infty$ *and* $\sum_{j \in \mathcal{A}} |\alpha_j| / [\sqrt{n} \sum_{j \in \mathcal{A}} \alpha_j^2] \to 0$. *Then*

$$W(\hat{\delta}^{b_n}_{\text{FAIR}}, \boldsymbol{\theta}) \le 1 - \Phi\left(\frac{(1 + o_P(1)) \sum_{j \in \mathcal{A}} \alpha_j^2 + nm(n_1 n_2)^{-1} - m b_n^2}{2\{(1 + o_P(1)) \sum_{j \in \mathcal{A}} \alpha_j^2 + nm(n_1 n_2)^{-1}\}^{1/2}} \right).$$

Features Annealed Independence Rules

Comment: The upper bound of $W(\hat{\delta}^{b_n}_{FAIR}, \boldsymbol{\theta})$ is greater than $W(\hat{\delta}^{m_n}_{NC}, \boldsymbol{\theta})$ (Theorem 4). This is expected as estimating the set A increases the classification error.

When the common covariance matrix is different from identity, FAIR takes a slightly different form:

$$\hat{\delta}_{\mathrm{FAIR}}(\mathbf{x}) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j)/\hat{\sigma}_j^2 1_{\{\sqrt{n/(n_1 n_2)}|T_j|>b\}}$$

, where T_j is the two-sample t-statistic.

Features Annealed Independence Rules

The number of features can be selected by minimizing the upper bound of the classification error. The optimal m is:

$$m_1 = \underset{1 \le m \le p}{\arg\max} \frac{1}{\lambda_{\max}^m} \frac{[\sum_{j=1}^m \alpha_j^2/\sigma_j^2 + m(1/n_2 - 1/n_1)]^2}{nm/(n_1 n_2) + \sum_{j=1}^m \alpha_j^2/\sigma_j^2},$$

where $\lambda_{\max}^m$ is the largest eigenvalue of the correlation matrix $\mathbf{R}^m$ of the truncated observations. It can be estimated from the samples:

$$\hat{m}_1 = \underset{1 \le m \le p}{\arg\max} \frac{1}{\hat{\lambda}_{\max}^m} \frac{[\sum_{j=1}^m \hat{\alpha}_j^2/\hat{\sigma}_j^2 + m(1/n_2 - 1/n_1)]^2}{nm/(n_1 n_2) + \sum_{j=1}^m \hat{\alpha}_j^2/\hat{\sigma}_j^2}$$

$$= \underset{1 \le m \le p}{\arg\max} \frac{1}{\hat{\lambda}_{\max}^m} \frac{n[\sum_{j=1}^m T_j^2 + m(n_1 - n_2)/n]^2}{m n_1 n_2 + n_1 n_2 \sum_{j=1}^m T_j^2}. \quad (4.3)$$

Simulation

The big picture of the simulation:

- The mean vector $\boldsymbol{\mu}_1$ is $(1-c)\delta_0 + \frac{c}{2}\exp(-2|x|)$, while $\boldsymbol{\mu}_2 = 0$.
- $n_1 = 30$ and $n_2 = 30$ for training. Seperate 200 samples are generated from each class as test dataset.
- Set $p = 4500$ and $c = 0.02$: around 90 signal features on an average, many of which are weak signals.

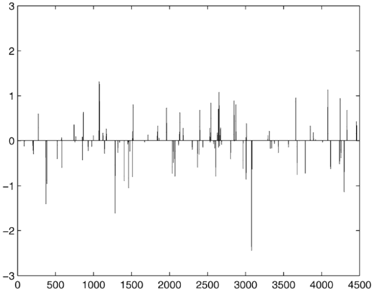
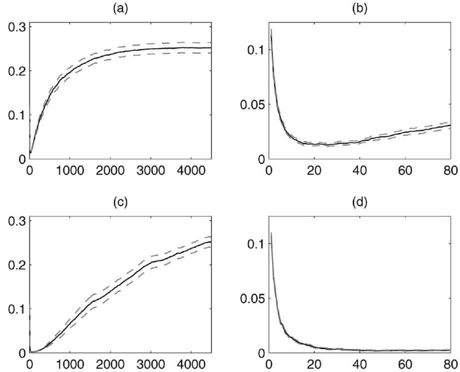


FIG. 1. *True mean difference vector* $\boldsymbol{\alpha}$. *x-axis represents the dimensionality, and y-axis shows the values of corresponding entries of* $\boldsymbol{\alpha}$.

Simulation

Number of features vs. misclassification rates (averages + 2 standard errors) over 100 simulations.

- row 1: ordered by t-statistic (equivalently, $\hat{\alpha}$, estimated mean differernces)
- row 2: ordered by true α



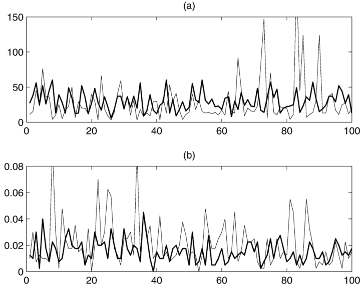
Simulation

Comments:

- Classification results of FAIR are close to those of the oracle -assisted independence classifier.
- m increase increases $\rightarrow$ misclassification rate increases
- min $= 0.0128$ in upper, min $= 0.0020$ in lower.
- when all features are included ($m = 4500$), MR $= 0.2522$.
- When we decrease the signal levels/ increase the dimensionality, MR tend to 0.5.
- Similar results based on projected samples.

Simulation

What about the proposed method for selecting features in FAIR? Compare their method (thick) to nearest shrunken centroids method (NSC, thin). **Upper**: number of chosen features over 100 simulations; **Lower**: Classification error.



FAIR: mean number of feature (29.71), mean MR (0.0154, sd $=$ 0.0085)
NSC: mean number of feature (28.43), mean MR (0.0216, sd $=$ 0.0179)

Application 1: Leukemia data

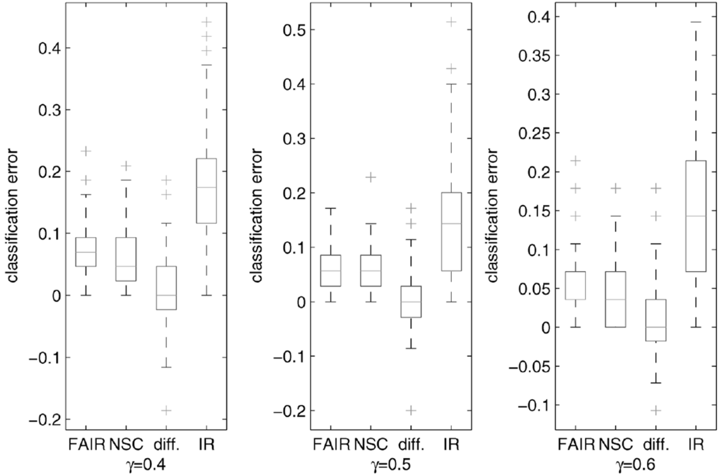
7129 genes (p), 72 samples (47 in ALL and 25 in AML). 27 in ALL and 11 in AML are set to be training.

TABLE 1
Classification errors of Leukemia dataset

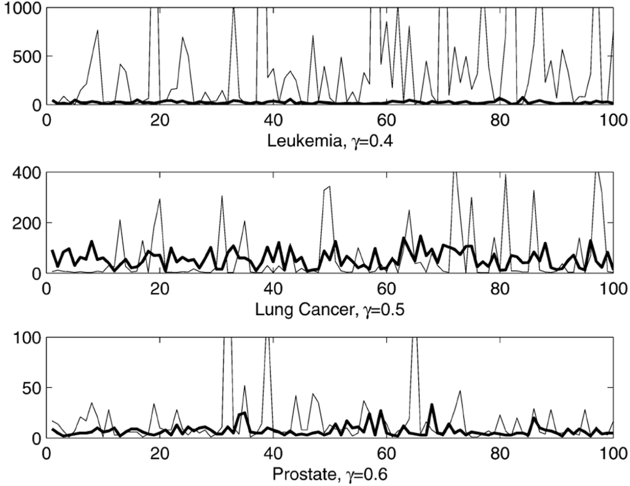
Method	**Training error**	**Test error**	**No. of selected genes**
Nearest shrunken centroids	1/38	3/34	21
FAIR	1/38	1/34	11

Application 1: Leukemia data

Further set different proportion of training ($100\gamma\%$). idff = FARI - NSC; IR = independence rule (all features)



Application 1: Leukemia data

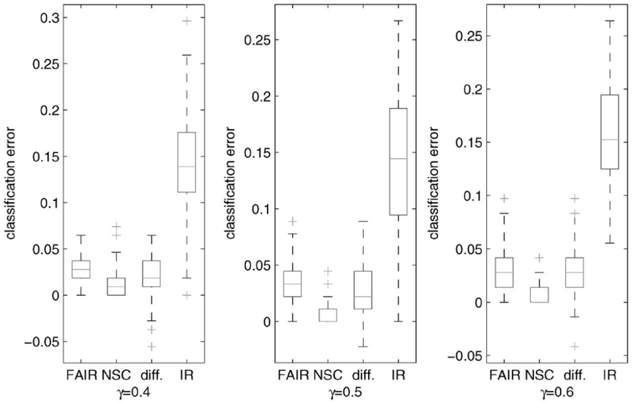


number of features: NSC is not good at feature selection.

Application 2: Lung Cancer data

TABLE 2
Classification errors of Lung cancer data

Method	Training error	Test error	No. of selected genes
Nearest shrunken centroids	0/32	11/149	26
FAIR	0/32	7/149	31



Application 3: Prostate Cancer data

TABLE 3
Classification errors of Prostate cancer dataset

Method	Training error	Test error	No. of selected genes
Nearest shrunken centroids	8/102	9/34	6
FAIR	10/102	9/34	2