CLASSIFICATION WITH MINIMUM EXPECTED ERROR OVER AN UNCERTAINTY CLASS OF GAUSSIAN DISTRIBUTIONS

Lori A. Dalton

The Ohio State University
Department of Electrical and Computer Engineering
Columbus, OH 43210 USA

ABSTRACT

In biomedicine, it is typical to find studies discriminating between types of pathology or stages of disease based on as few as 30 sample points and tens of thousands of genes. Unfortunately, out-of-the-box classification and error estimation rules come with no small-sample performance guarantees, which has greatly contributed to the crisis in biomarker reproducibility. Recent work addresses this by supplementing the data with expert biological knowledge via a prior distribution over an uncertainty class of feature-label distributions, and uses the resulting probabilistic framework to define minimum mean-square-error (MMSE) estimators for the misclassification rate of any fixed classifier, as well as the sample-conditioned MSE itself for arbitrary error estimators. Here, we use the same framework to also define minimum expected error (MEE) classifiers, completing a Bayesian optimized theory of classification. We also present examples on real genomic data resulting in classifiers that greatly outperform popular rules.

Index Terms— Bayesian modeling, classification, error estimation, genomics, small samples

1. INTRODUCTION

Given a labeled training sample, the usual procedure is to apply a classification rule, which may involve feature selection, and then to estimate the misclassification rate of the designed classifier using an error estimation rule. The main epistemological issue here is model validity, which is addressed by error estimation [1]. When large amounts of data are available, one can split the data into a training set for classifier design and an independent test set used to estimate the error of the classifier by r/m, where r is the number of incorrectly classified points and m is the total number of test points. A distribution-free bound for the root-mean-square (RMS) of this holdout estimator is given by $\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{holdout}}|S_{n-m},F) \leq 1/\sqrt{4m}$, where S_{n-m} is any training sample and F is any feature-label distribution [2]. Clearly, accurate error estimation is assured when m is large enough.

That being said, when samples are expensive or difficult to acquire, training-data error estimation methods like bootstrap and cross-validation are often used to avoid sacrificing performance in the classifier. Not only is the situation complicated by the use of training-data error estimation rules typically lacking theoretical performance bounds, but absent prior knowledge these bounds are useless in the range of sample sizes where these methods are needed [3].

With large samples, one may also appeal to Vapnik-Chervonenkis theory [4] to find distribution-free bounds, as a function of sample size, on the tail probability of the difference between the true and apparent error for any classifier, as well as the tail probability of the difference between the error of the best classifier in a family of classifiers and the error of the designed classifier. However, just as with RMS bounds, VC bounds are too loose to be useful for small samples.

We are thus left with no distribution-free guarantees regarding the performance of a designed classifier in a small sample setting. In fact, studies have shown that error estimation is indeed problematic even for fixed distributions, for instance cross-validation and other re-sampling methods often have large RMS due to high variance [5, 6, 7]. Small-sample problems must be treated in their own right, without appealing to limiting theorems or bounds that require large sample sizes. As long ago as 1925, R. A. Fisher wrote, "Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data" [8].

Small sample performance can be good when the Bayes error is small enough [9, 10], and we may achieve performance guarantees by assuming the feature-label distribution is a member of some uncertainty class of states where the error estimator is known to perform well. Recent work goes a step further to assign a prior distribution to an uncertainty class and update the prior to a posterior given the data. The prior can contain expert knowledge about the state of the system, which is critical in practical problems where the sample alone does not contain enough information for validation. Given the posterior and a classifier, we may then find a sample-conditioned RMS for any error estimator [11, 12]; for the first time we achieve a practical measure of the performance of an error estimator relative to a Bayesian model,

observed sample and designed classifier.

Not only that, but one may utilize the proposed framework throughout the entire classification procedure to obtain optimal MMSE estimators for the misclassification rate of any classifier [13, 14], and herein we present new MEE classifiers designed in the same framework [15, 16]. Optimization is especially important in the small-sample setting because it is here that classification and error estimation are most difficult.

We begin by reviewing the general theory behind MMSE error estimation, the sample-conditioned MSE, and new MEE classification rules. We then find MEE classifiers under a Gaussian model with conjugate priors and demonstrate good performance on real gene-expression microarray data.

2. THE BAYESIAN FRAMEWORK

Consider a binary classification problem with classes 0 and 1, and let c be the *a priori* probability that a sample point is from class 0. Let $\mathbf{\Theta}_y$ be an uncertainty class of possible distributions for class $y \in \{0,1\}$ and let $\theta_y \in \mathbf{\Theta}_y$ parameterize the class-conditional distribution for class y, denoted by f_{θ_y}. Our Bayesian framework assigns priors, $\pi(c)$, $\pi(\theta_0)$ and $\pi(\theta_1)$, to the parameters, c, θ_0 and θ_1, respectively. We assume that θ_0 and θ_1 are independent from c prior to observing the data.

Given n_y independent and identically distributed (i.i.d.) sample points from class y, these priors are updated to posteriors denoted by π^*, which combine prior knowledge with observed data to quantify uncertainty in our knowledge about the true state of nature. Independence is preserved after observing the data. Assuming a beta(α,β) prior for c (a uniform prior corresponds to $\alpha=\beta=1$) and random sampling where the class of each point is an independent Bernoulli trial, the posterior of c is also beta with updated hyperparameters $\alpha+n_0$ and $\beta+n_1$. In this case, the expectation of c is

$$E_{\pi^*}[c] = \tfrac{n_0+\alpha}{n+\alpha+\beta}, \tag{1}$$

where $n = n_0 + n_1$ and E_{π^*} denotes an expectation relative to the posterior (conditioned on the sample). Alternatively, c may be known quite accurately *a priori*, for instance it may represent the probability that an individual has a certain type of cancer, in which case we set $E_{\pi^*}[c]$ to the known value of c. The posterior of θ_y is found by normalizing the product of the prior and a likelihood function on sample points observed from class y. That is,

$$\pi^*(\theta_y) \propto \pi(\theta_y) \textstyle\prod_{i=1}^{n_y} f_{\theta_y}(\mathbf{x}_i^y),$$

where $\mathbf{x}_i^y$ is the i^{th} sample point in class y. This follows from Bayes rule if the prior is proper (normalizeable) and may be taken as a definition otherwise, but in all cases the posterior must be proper.

Letting $\theta = [c, \theta_0, \theta_1]$, the misclassification rate for any fixed classifier ψ_n is of the form

$$\varepsilon(\theta, \psi_n) = c\varepsilon^0(\theta_0, \psi_n) + (1-c)\varepsilon^1(\theta_1, \psi_n), \tag{2}$$

where $\varepsilon^y(\theta_y, \psi_n)$ is the probability that ψ_n wrongly classifies a point from class y having true parameter θ_y. Since c, θ_0 and θ_1, are unknown, we must estimate the error from data. The MMSE estimator is equivalent to the expected true error conditioned on the sample, i.e.,

$$\begin{aligned}\widehat{\varepsilon}_{\text{MMSE}}(S_n, \psi_n) &= \mathrm{E}_{\pi^*}[\varepsilon(\theta, \psi_n)] \\ &= \mathrm{E}_{\pi^*}[c]\widehat{\varepsilon}^0(S_n, \psi_n) + (1 - \mathrm{E}_{\pi^*}[c])\widehat{\varepsilon}^1(S_n, \psi_n),\end{aligned} \tag{3}$$

where S_n is the sample and $\widehat{\varepsilon}^y(S_n, \psi_n) = \mathrm{E}_{\pi^*}[\varepsilon^y(\theta_y, \psi_n)]$. The expectation of c depends on our prior model for c, and $\widehat{\varepsilon}^y(S_n, \psi_n)$ can be found using the following theorem [15].

Theorem 1 *Let ψ be a fixed classifier given by $\psi(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where R_0 and R_1 are measurable sets partitioning the sample space. Then the MMSE error estimator is given by* (3)*, where*

$$\widehat{\varepsilon}^y(S_n, \psi_n) = \int_{R_{1-y}} f(\mathbf{x}|y)\, d\mathbf{x}, \tag{4}$$

$$f(\mathbf{x}|y) = \int_{\mathbf{\Theta}_y} f_{\theta_y}(\mathbf{x}|y)\, \pi^*(\theta_y)\, d\theta_y. \tag{5}$$

The sample-conditioned MSE of $\widehat{\varepsilon}_{\text{MMSE}}(S_n, \psi_n)$ is

$$\text{MSE}(\widehat{\varepsilon}_{\text{MMSE}}|S_n) = \mathrm{E}_{\pi^*}[(\varepsilon(\theta, \psi_n) - \widehat{\varepsilon}_{\text{MMSE}}(S_n, \psi_n))^2].$$

By the orthogonality principle from MMSE estimation theory, this is equivalent to the variance of the true error, $\text{Var}_{\pi^*}(\varepsilon(\theta, \psi_n))$. Similarly, we have $\text{MSE}(\widehat{\varepsilon}^y|S_n) = \text{Var}_{\pi^*}(\varepsilon^y(\theta_y, \psi_n))$, and one can show:

$$\begin{aligned}\text{MSE}(\widehat{\varepsilon}_{\text{MMSE}}|S_n) &= \text{Var}_{\pi^*}(c)\left(\widehat{\varepsilon}^0(S_n, \psi_n) - \widehat{\varepsilon}^1(S_n, \psi_n)\right)^2 \\ &\quad + \mathrm{E}_{\pi^*}\left[c^2\right]\text{MSE}(\widehat{\varepsilon}^0|S_n) + \mathrm{E}_{\pi^*}\left[(1-c)^2\right]\text{MSE}(\widehat{\varepsilon}^1|S_n).\end{aligned}$$

The sample-conditioned MSE for an arbitrary error estimate $\widehat{\varepsilon}$ also falls out naturally:

$$\text{MSE}(\widehat{\varepsilon}|S_n) = \text{MSE}(\widehat{\varepsilon}_{\text{MMSE}}|S_n) + (\widehat{\varepsilon}_{\text{MMSE}} - \widehat{\varepsilon})^2.$$

This quantifies the accuracy of $\widehat{\varepsilon}$ as an estimator for $\varepsilon(\theta, \psi_n)$ conditioned on the actual sample in hand. There are numerous applications, for instance we may employ it as a stopping criterion in censored sampling, where sample points are collected one at a time until the sample-conditioned MSE and expected true error reach desired levels, either for the classes individually or combined [12].

Closed form solutions for the MMSE error estimator and the sample conditioned MSE are availalbe in both Gaussian models with with conjugate priors and linear classification and multinomial models with Dirichlet priors [13, 14, 11, 12].

3. THE MEE CLASSIFIER

The MEE classifier minimizes the expected true error:

$$\psi_{\text{MEE}} = \arg\inf_{\psi \in \mathcal{C}} \mathrm{E}_{\pi^*}[\varepsilon(\theta, \psi)], \tag{6}$$

where $\mathcal{C}$ is an arbitrary family of classifiers [15, 16]. To motivate this definition, note that under the Bayesian framework,

$$\begin{aligned} \mathrm{P}\left(\psi\left(\mathbf{X}\right) \neq Y | S_n\right) &= \mathrm{E}_{\pi^*}\left[P\left(\psi\left(\mathbf{X}\right) \neq Y | \theta, S_n\right)\right] \\ &= \mathrm{E}_{\pi^*}\left[\varepsilon(\theta, \psi)\right]. \end{aligned} \tag{7}$$

Thus, MEE classifiers minimize the misclassification probability relative to the assumed model, given the sample. If $\mathcal{C}$ is the set of all classifiers with measurable decision regions, the MEE classifier is solved in the following theorem [15].

Theorem 2 *An MEE classifier, ψ_{MEE}, satisfying* (6)*, where $\mathcal{C}$ is the set of all classifiers with measurable decision regions, exists and is given pointwise by*

$$\psi_{\mathrm{MEE}}\left(\mathbf{x}\right) = \begin{cases} 0 & \text{if } \mathrm{E}_{\pi^*}[c] f\left(\mathbf{x}|0\right) \geq \left(1 - \mathrm{E}_{\pi^*}[c]\right) f\left(\mathbf{x}|1\right), \\ 1 & \text{otherwise}. \end{cases}$$

The MEE classifier is equivalent to the Bayes classifier for fixed class-conditional distributions $f\left(\mathbf{x}|y\right)$, $y \in \{0, 1\}$, and class-0 probability $\mathrm{E}_{\pi^*}[c]$. This is like a plug-in rule, only $f\left(\mathbf{x}|y\right)$ is not necessarily a member of the class-conditional densities in our uncertainty class, but possibly some other kind of density that happens to result in the optimal classifier. We thus refer to $f\left(\mathbf{x}|y\right)$ as the *effective class-conditional density* with respect to the posterior. Further, the optimal classifier is defined pointwise, labeling each point in the sample space with the class having maximum a posteriori probability.

By Theorem 1, the MMSE error estimator for an arbitrary classifier is equivalent to the Bayes error for the same effective feature-label distribution. Hence, $f\left(\mathbf{x}|y\right)$ contains all of the necessary information to find the optimal classifier, the optimal classifier error, and the error for arbitrary classifiers, and we do not have to deal with the priors directly. Rather, upon defining a model we find $f\left(\mathbf{x}|y\right)$, which depends on the sample because it depends on π^*, and the problem is solved by treating $f\left(\mathbf{x}|y\right)$ as the true feature-label distribution.

4. THE GAUSSIAN MODEL

Suppose each sample point is a column vector of D features, where the class-y conditional distribution is Gaussian with parameter $\theta_y = [\mu_y, \Sigma_y]$, where μ_y is the mean and Σ_y is the covariance. Although in [15] we consider three types of structure in the covariance, here we only consider the arbitrary covariance model where the parameter space of Σ_y consists of all positive definite (valid covariance) matrices. We also consider only the independent covariance model, where θ_0 and θ_1 are independent prior to observing the data, although a homoscedastic covariance model has also been treated in [15].

We assume a conjugate prior where Σ_y is invertible with probability 1, and for invertible Σ_y we have

$$\pi(\theta_y) = \pi(\mu_y|\Sigma_y)\pi(\Sigma_y), \tag{8}$$

where

$$\pi(\mu_y|\Sigma_y) \propto |\Sigma_y|^{-\frac{1}{2}} \exp\left(-\tfrac{\nu_y}{2}(\mu_y - \mathbf{m}_y)^T \Sigma_y^{-1}(\mu_y - \mathbf{m}_y)\right)$$
$$\pi(\Sigma_y) \propto |\Sigma_y|^{-\frac{\kappa_y + D + 1}{2}} \exp\left(-\tfrac{1}{2}\mathrm{trace}\left(S_y \Sigma_y^{-1}\right)\right).$$

We allow for proper or improper priors where $\nu_y \geq 0$, $\mathbf{m}_y$ is a length D real vector, κ_y is a real number, and S_y is a non-negative definite $D \times D$ matrix. If $\nu_y > 0$, $\kappa_y > D - 1$ and S_y is positive definite, then this is a proper prior [17, 18]. In this case $\pi(\mu_y|\Sigma_y)$ is Gaussian with mean $\mathbf{m}_y$ and covariance Σ_y/ν_y and $\pi(\Sigma_y)$ is an inverse-Wishart distribution where $\mathrm{E}_\pi[\Sigma_y] = S_y/(\kappa_y - D - 1)$. Hyperparameter S_y controls the expected covariance, and if S_y is scaled appropriately then the larger κ_y is the tighter the prior is about this mean.

It can be shown that the posterior, $\pi^*(\theta_y)$, has the same form as the prior with updated hyperparameters $\nu_y^* = \nu_y + n_y$, $\kappa_y^* = \kappa_y + n_y$ and

$$\mathbf{m}_y^* = (\nu_y \mathbf{m}_y + n_y \widehat{\mu}_y)(\nu_y + n_y),$$
$$S_y^* = S_y + (n_y - 1)\widehat{\Sigma}_y + \tfrac{\nu_y n_y}{\nu_y + n_y}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T,$$

where $\widehat{\mu}_y$ and $\widehat{\Sigma}_y$ are the usual sample mean and covariance, repectively, of the n_y points in class y. The priors are proper if $\nu_y^* > 0$, $\kappa_y^* > D - 1$ and S_y^* positive definite.

The effective density is a multivariate student's t distribution having location vector $\mathbf{m}_y^*$, scale matrix $\Psi_y = \frac{\nu^* + 1}{(\kappa^* - D + 1)\nu^*} S^*$ and $k_y = \kappa^* - D + 1$ degrees of freedom:

$$f\left(\mathbf{x}|y\right) = \tfrac{1}{k_y^{D/2} \pi^{D/2} |\Psi_y|^{1/2}} \times \tfrac{\Gamma\left(\frac{k_y + D}{2}\right)}{\Gamma\left(\frac{k_y}{2}\right)} \times \left(1 + \tfrac{1}{k_y}\left(\mathbf{x} - \mathbf{m}_y^*\right)^T \Psi_y^{-1}\left(\mathbf{x} - \mathbf{m}_y^*\right)\right)^{-\frac{k_y + D}{2}}.$$

As long as π^* is proper, the effective density is also proper. Also, if $\kappa^* > D$ the mean of this distribution is $\mathbf{m}^*$, and if $\kappa^* > D + 1$ the variance is $\frac{\nu^* + 1}{(\kappa^* - D - 1)\nu^*} S^*$.

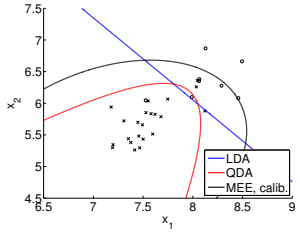
The MEE classifier can be expressed as $\psi_{\mathrm{MEE}}(\mathbf{x}) = 0$ if $g_{\mathrm{MEE}}(\mathbf{x}) \leq 0$ and $\psi_{\mathrm{MEE}}(\mathbf{x}) = 1$ if $g_{\mathrm{MEE}}(\mathbf{x}) > 0$, where

$$g_{\mathrm{MEE}}(\mathbf{x}) = K\left(1 + \tfrac{1}{k_0}\left(\mathbf{x} - \mathbf{m}_0^*\right)^T \Psi_0^{-1}\left(\mathbf{x} - \mathbf{m}_0^*\right)\right)^{k_0 + D} - \left(1 + \tfrac{1}{k_1}\left(\mathbf{x} - \mathbf{m}_1^*\right)^T \Psi_1^{-1}\left(\mathbf{x} - \mathbf{m}_1^*\right)\right)^{k_1 + D}$$

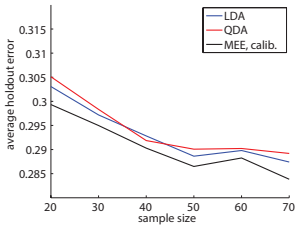
and

$$K = \left(\tfrac{1 - \mathrm{E}_{\pi^*}[c]}{\mathrm{E}_{\pi^*}[c]}\right)^2 \left(\tfrac{k_0}{k_1}\right)^D \tfrac{|\Psi_0|}{|\Psi_1|} \left(\tfrac{\Gamma(k_0/2)\Gamma((k_1 + D)/2)}{\Gamma((k_0 + D)/2)\Gamma(k_1/2)}\right)^2.$$

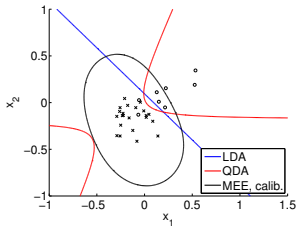
This classifier has a polynomial decision boundary whenever κ_0 and κ_1 are integers. Although we only consider Gaussian distributions in our model, the form of the MEE classifier is not necessarily linear or quadratic, although it is easy to evaluate at a fixed point $\mathbf{x}$. The expected error of the classifier may be found via Monte-Carlo integral approximation by



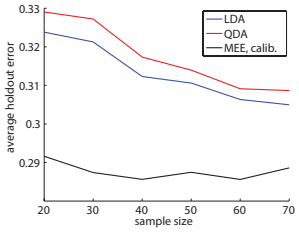
(a) holdout errors 0.28571, 0.30655 and 0.28274 for LDA, QDA and MEE, respectively, from [19]



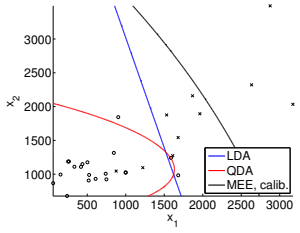
(b) average holdout error from [19]



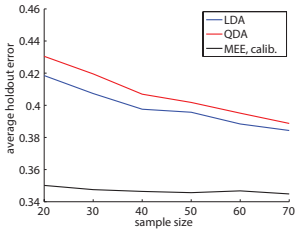
(c) holdout errors 0.34958, 0.34534 and 0.31356 for LDA, QDA and MEE, respectively, from [20]



(d) average holdout error from [20]



(e) holdout errors 0.36992, 0.39228 and 0.33333 for LDA, QDA and MEE, respectively, from [21]



(f) average holdout error from [21]

Fig. 1. (a), (c) and (e): Example training sample of size $n = 30$ and classifiers from a real dataset with $D = 2$ selected features. Class 0 points are marked with o's and class 1 points with x's. (b), (d) and (f): Average holdout errors on a real dataset with $D = 2$ selected features versus sample size.

drawing points from the effective densities $f(\mathbf{x}|y)$ and evaluating the proportion of misclassified points. The sample-conditioned MSE can also be found via Monte-Carlo approximation, though not via the effective densities.

5. PERFORMANCE ON REAL GENOMIC DATA

We consider three real microarray gene-expression datasets from [22]. The first is a non-small-cell lung cancer (NSCLC) data set with 198 sample points and 22,215 features, where class 0 contains 54 points associated with 5-year disease-free survival and class 1 contains 144 points associated with death within 2.5 years [19]. The second is a primary breast carcinoma dataset with 266 sample points and 5003 features, where class 0 contains 70 points associated with distant metastases within five years and class 1 contains 196 points associated with disease-free outcome for more than five years [20]. The final set is a lymph-node-negative breast cancer dataset with 276 sample points and 22,215 features, where class 0 contains 183 points associated with a metastasis-free outcome for 5 years and class 1 contains 93 points associated with relapse in less than 5 years [21].

In all cases, we assume that $c = N_0/(N_0+N_1)$ is the true class-0 probability, where N_y is the class-y sample size in the full dataset. For a training sample of size n, we draw the appropriate number of points for each class from the full dataset, keeping the proportion of points in class 0 as close as possible to c. We eliminate features rejected by a Shapiro-Wilk hypothesis test at a 95% significance level on either class, and select 2 of the remaining features using a t-test.

After feature selection, we design three classifiers: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and an MEE classifier from priors calibrated using the method-of-moments approach proposed in [23] from the 5000 discarded features having largest t-test statistic. An example training sample and designed classifier is shown in Fig. 1(a), (c) and (e), each corresponding to a different dataset. Observe that the MEE classifier is not necessarily linear or quadratic.

We approximate the true error for each designed classifier using the holdout error on points not used in training. This process is repeated 1,000 times for each dataset and sample size. Once this iteration is complete, for each classifier we find the average holdout error, which is plotted with respect to sample size in Fig. 1(b), (d) and (f). Although MEE classification is not optimal for a fixed (empirical) distribution, but only optimal when averaged over the uncertainty class, for all datasets shown here the MEE classifier consistently performs at least as well as LDA and QDA, often with substantial gain. In addition, there may be much room for improvement since we have achieved this using only a very simple Gaussian model and a purely data driven method of devising priors.

6. CONCLUSION

MEE classification, along with MMSE error estimation and the sample-conditioned MSE, constitute a new Bayesian theory of optimal classification. This theory facilitates the addition of expert prior knowledge into the model and optimizes classifier and error estimator design to improve performance. More importantly epistemologically, we can validate findings using the sample-conditioned MSE. Much is known beyond this: invariance of MEE classification under invertible transformations of the sample space, consistency of both optimal classification and error estimation in the discrete and Gaussian models under mild regularity conditions, and a connection between MEE classification with optimal Bayesian robust classification [15, 16]. Robustness to incorrect modeling assumptions has also been investigated [14, 16].

7. REFERENCES

[1] E. R. Dougherty and U. Braga-Neto, "Epistemology of computational biology: Mathematical models and experimental prediction as the basis of their validity," *Biological Systems*, vol. 14, no. 1, pp. 65–90, 2006.

[2] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

[3] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, "The illusion of distribution-free small-sample classification in genomics," *Current Genomics*, vol. 12, no. 5, pp. 333–341, 2011.

[4] V. N. Vapnik and A. Chervronenkis, *Theory of Pattern Recognition*, Nauka, Moscow, 1974.

[5] N. Glick, "Additive estimators for probabilites of correct classification," *Pattern Recognition*, vol. 10, no. 3, pp. 211–222, 1978.

[6] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[7] B. Hanczar, J. Hua, and E. R. Dougherty, "Decorrelation of the true and estimated classifier errors in high-dimensional settings," *EURASIP J. Bioinf. Sys. Bio.*, vol. 2007, January 2007, Article ID 38473, 12 pages.

[8] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.

[9] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4238–4255, 2011.

[10] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate heteroskedastic Gaussian model," *Pattern Recognition*, vol. 45, no. 2, pp. 908–917, 2012.

[11] L. A. Dalton and E. R. Dougherty, "Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part I: Representation," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2575–2587, 2012.

[12] L. A. Dalton and E. R. Dougherty, "Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part II: Consistency and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2588–2603, 2012.

[13] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error–Part I: Definition and the Bayesian MMSE error estimator for discrete classification," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 115–129, 2011.

[14] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error–Part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 130–144, 2011.

[15] L. A. Dalton and E. R. Dougherty, "Optimal classifiers with minimum expected error within a Bayesian framework–Part I: Discrete and Gaussian models," *Pattern Recognition*, in press, 2012.

[16] L. A. Dalton and E. R. Dougherty, "Optimal classifiers with minimum expected error within a Bayesian framework–Part II: Properties and performance analysis," *Pattern Recognition*, in press, 2012.

[17] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

[18] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*, MIT Press, Cambridge, MA, 1961.

[19] A. Potti, S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild *et al*., "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer," *N. Eng. J. Med.*, vol. 355, pp. 570–80, 2006.

[20] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart *et al*., "A gene-expression signature as a predictor of survival in breast cancer," *N. Eng. J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002.

[21] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look *et al*., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671–679, 2005.

[22] M. R. Yousefi, J. Hua, C. Sima, and E. R. Dougherty, "Reporting bias when using real data sets to analyze classification performance," *Bioinformatics*, vol. 26, no. 1, pp. 68–76, 2010.

[23] L. A. Dalton and E. R. Dougherty, "Application of the Bayesian MMSE estimator for classification error to gene expression microarray data," *Bioinformatics*, vol. 27, no. 13, pp. 1822–1831, 2011.