Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 16354, 11 pages
doi:10.1155/2007/16354

Research Article

Quantification of the Impact of Feature Selection on the Variance of Cross-Validation Error Estimation

Yufei Xiao,[1] Jianping Hua,[2] and Edward R. Dougherty[1, 2]

[1] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA*
[2] *Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*

Received 7 August 2006; Revised 21 December 2006; Accepted 26 December 2006

Recommended by Paola Sebastiani

Given the relatively small number of microarrays typically used in gene-expression-based classification, all of the data must be used to train a classifier and therefore the same training data is used for error estimation. The key issue regarding the quality of an error estimator in the context of small samples is its accuracy, and this is most directly analyzed via the deviation distribution of the estimator, this being the distribution of the difference between the estimated and true errors. Past studies indicate that given a prior set of features, cross-validation does not perform as well in this regard as some other training-data-based error estimators. The purpose of this study is to quantify the degree to which feature selection increases the variation of the deviation distribution in addition to the variation in the absence of feature selection. To this end, we propose the coefficient of relative increase in deviation dispersion (CRIDD), which gives the relative increase in the deviation-distribution variance using feature selection as opposed to using an optimal feature set without feature selection. The contribution of feature selection to the variance of the deviation distribution can be significant, contributing to over half of the variance in many of the cases studied. We consider linear-discriminant analysis, 3-nearest-neighbor, and linear support vector machines for classification; sequential forward selection, sequential forward floating selection, and the t-test for feature selection; and k-fold and leave-one-out cross-validation for error estimation. We apply these to three feature-label models and patient data from a breast cancer study. In sum, the cross-validation deviation distribution is significantly flatter when there is feature selection, compared with the case when cross-validation is performed on a given feature set. This is reflected by the observed positive values of the CRIDD, which is defined to quantify the contribution of feature selection towards the deviation variance.

1. INTRODUCTION

$\mathbb{R}^2$ P $\mathbb{R}^N$ Given the relatively small number of microarrays typically used in expression-based classification for diagnosis and prognosis, all the data must be used to train a classifier and therefore the same training data is used for error estimation. A classifier is designed according to a classification rule, with the rule being applied to sample data to yield a classifier. Thus, the classifier and its error are functions of the random sample. Regarding features, there are two possibilities: either the features are given prior to the data, in which case the classification rule yields a classifier with the given features constituting its argument, or both the features and classifier are determined by the classification rule. In the latter case, the entire set of possible features constitutes the feature set relative to the classification rule, whereas only the selected features constitute the feature set relative to the designed classifier. Feature selection constrains the space of functions from which a classifier might be chosen, but it does not reduce the number of features involved in designing the classifier. If there are D features from which a classifier based on d features is to be determined, then, absent feature selection, the chosen classifier must come from some function space over D features, whereas with feature selection, the chosen classifier will be a function of some subset consisting of d features out of D. In particular, if cross-validation error estimation is used, then the approximate unbiasedness of cross-validation applies to the classification rule, and since feature selection is part of the classification rule, feature selection must be accounted for within the cross-validation procedure to maintain the approximate unbiasedness [1]. This paper concerns the quality of such a cross-validation estimation procedure.

There are various issues to consider with regard to the quality of an error estimator in the context of small samples.

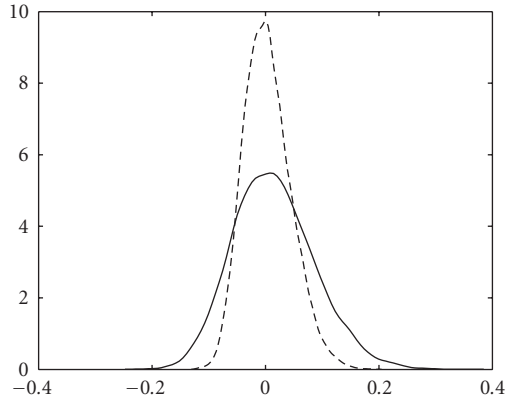

Figure 1: Deviation distributions with feature selection (solid line) and without feature selection (dashed line). The x-axis denotes the deviation, namely, the difference of the estimated error and the true error; the y-axis corresponds to the density.

The most obvious is its accuracy, and this is most directly analyzed via the deviation distribution of the estimator, that is, the distribution of the difference between the estimated and true errors. Model-based simulation studies indicate that, given a prior set of features, cross-validation does not perform as well in this regard as bootstrap or bolstered estimators [2, 3]. Model-based simulation also indicates that, given a prior set of features, cross-validation does not perform well when ranking feature sets of a given size [4]. Moreover, when doing feature selection, similar studies show that cross-validation does not do well in comparison to bootstrap and bolstered estimators when used inside forward search algorithms, such as sequential forward selection and sequential forward floating selection [5].

Here we are concerned with the use of cross-validation to estimate the error of a classifier designed in conjunction with feature selection. This issue is problematic because, owing to the computational burden of bootstrap and the analytic formulation of bolstering, these are not readily amenable to situations where there are thousands of features from which to choose. As in the case of prior-chosen features, the main concern here is the deviation distribution between the cross-validation error estimates and the true errors of the designed classifiers. Owing to the added complexity of feature selection, one might surmise that the situation here would be worse than that for a given feature set, and it is. Even with a given feature set, the deviation distribution for cross-validation tends to have high variance, which is why its performance generally is not good, especially for leave-one-out cross-validation [2]. We observe in the current study that the cross-validation deviation distribution is significantly flatter when there is feature selection, which means that cross-validation estimates are even more unreliable than for given feature sets, and that they are sufficiently unreliable to raise serious concerns when such estimates are reported. Figure 1 shows the typical deviation distributions of cross-validation (i) with feature selection (solid line) and (ii) without feature selection, that is, using the known best features (dashed line). In the simulations to be performed, we choose the models such that the optimal feature set is directly obtainable from the model, and an existing test bed provides the best feature sets for the patient data.

A study comparing several resampling error-estimation methods has recently addressed the inaccuracy of cross-validation in the presence of feature selection [6]. Using four classification rules (linear discriminant analysis, diagonal discriminant analysis, nearest neighbor, and CART), the study compares bias, standard deviation, and mean-squared error. Both simulated and patient data are used, and the t-test is employed for feature selection. Our work differs from [6] in two substantive ways. The major difference is that we employ a comparative quantitative methodology by studying the deviation distributions and defining a measure that isolates as well as assesses the effects of feature selection on the deviation analysis of cross-validation. This is necessary in order to quantify the contribution of feature selection in its role as part of the classification rule. This quantitative approach shows that the negative effects of feature selection depend very much on the underlying classification rule. A second difference is that our study uses three different algorithms, namely, t-test, sequential forward selection (SFS), and the sequential forward floating selection (SFFS) algorithm [7] to select features, whereas [6] relies solely on t-test feature selection. The cost for using SFS and SFFS in a large simulation study is that they are heavily computational and therefore we rely on high-performance computing using a Beowulf cluster.

A preliminary report on our study was presented at the *IEEE International Workshop on Genomic Signal Processing and Statistics* for 2006 [8].

2. SYSTEMS AND METHODS

Our interest is with the deviation distribution of an error estimator, that is being the distribution of difference between the estimated and true errors of a classifier. Three classification rules will be considered: linear discriminant analysis (LDA), 3-nearest-neighbor (3NN), and linear support vector machine (SVM). Our method is to compare the cross-validation (k-fold and leave-one-out) deviation distributions for classification rules used with and without feature selection. For feature selection, we will consider three algorithms: t-test, SFS, and SFFS (see Appendix A). Doing so will allow us to evaluate the degree of deterioration in deviation variance resulting from feature selection. In the case without feature selection, the known best d features among the full feature set will be applied for classification. It is expected that feature selection will result in a larger deviation variance than without feature selection, which is confirmed in this study.

2.1. Coefficient of relative increase in deviation dispersion

Given a sample set $\mathcal{S}$, we use the following notations for classification errors. For the exact mathematical formulae of the cross-validation errors, please refer to Appendix B.

(E) The true error of a classifier in the presence of feature selection, obtained by performing feature selection and designing a classifier on $\mathcal{S}$, and then finding the classification error on a large independent test sample $\mathcal{S}'$.

(E_b) The true error of a classifier using the known best features, obtained by designing a classifier on $\mathcal{S}$ with the known best feature set, and then finding the classification error on a large independent test sample $\mathcal{S}'$.

($\hat{E}$) The (k-fold or leave-one-out) cross-validation error in the presence of feature selection. To obtain the k-fold cross-validation error: divide the sample data into k portions as evenly as possible. During each fold of cross-validation, use one portion as the test sample and the rest as the training sample; perform feature selection and design a classifier on the training sample, and estimate its error on the test sample. Find the average error of k-folds, which is $\hat{E}$. Leave-one-out error is a special case when k equals the sample size.

($\hat{E}_b$) The (k-fold or leave-one-out) cross-validation error with the best features, obtained by performing cross-validation using the known best features.

Based on these errors, we are interested in the following deviations, referring to the difference of the estimated error and the true error:

(ΔE) defined as $\hat{E} - E$;
(ΔE_b) defined as $\hat{E}_b - E_b$.

To quantify the effect of feature selection on cross-validation variance, using the deviation variances we define the *coefficient of relative increase in deviation dispersion* (CRIDD) by

$$\kappa = \frac{\mathrm{Var}(\Delta E) - \mathrm{Var}\left(\Delta E_b\right)}{\mathrm{Var}(\Delta E)}. \tag{1}$$

Notice that κ is a relative measure, which is normalized by $\mathrm{Var}(\Delta E)$, because we are concerned with the relative change of deviation variance in the presence of feature selection. In our experiments, κ is expected to be positive, because ΔE contains two sources of uncertainty: cross-validation and feature selection, while ΔE_b contains none of the latter. When positive, κ will be in the range of $(0, 1]$, which indicates a deterioration in the deviation variance, due to the difference of with and without feature selection, and the larger κ is, the more severe the impact of feature selection.

2.2. Data

The models for simulated data take into account two requirements. First, in genomic applications, classification usually involves a large number of correlated features and the sample size is out-numbered by the features; and second, we need to know from the model the best feature set. We consider the following three models under the assumption of two equiprobable classes (classes 0 and 1).

(a) Equal covariance model: the classes 0 and 1 are drawn from multivariate Gaussian distributions $(\boldsymbol{\mu}_a, \boldsymbol{\Sigma})$ and $(-\boldsymbol{\mu}_a, \boldsymbol{\Sigma})$, respectively, the optimal classifier on the full feature-label distribution being given by LDA.

(b) Unequal covariance model: the classes 0 and 1 are drawn from multivariate Gaussian distributions $(\boldsymbol{\mu}_b, \boldsymbol{\Sigma})$ and $(-\boldsymbol{\mu}_b, 2\boldsymbol{\Sigma})$, respectively, the optimal classifier on the full feature-label distribution being given by quadratic discriminant analysis (QDA).

(c) Bimodal model: class 0 is generated from a multivariate Gaussian distribution $(\mathbf{0}, \boldsymbol{\Sigma})$ and class 1 is generated from a mixture of two equiprobable multivariate Gaussian distributions $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ and $(-\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$.

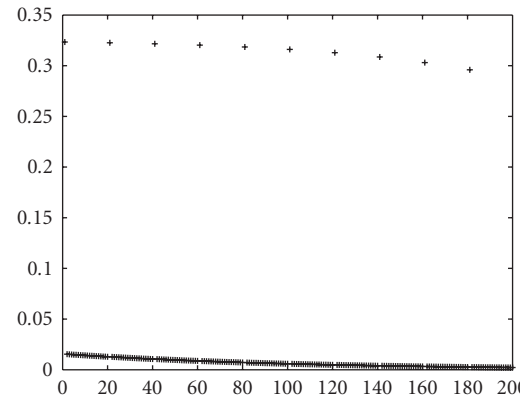

Figure 2: Vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{200})$. The x-axis denotes μ_1, $\mu_2, \ldots, \mu_{200}$, and the y-axis denotes their values.

For the above models, we have chosen $\boldsymbol{\mu}_a = \boldsymbol{\mu}_b = 1.75\boldsymbol{\mu}$ and $\boldsymbol{\mu}_c = 4.0\boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{200})$ is plotted in Figure 2 (for details of generating $\boldsymbol{\mu}$, please go to the companion website http://gsp.tamu.edu/web2/quantify_fscv/generate_mu.pdf). Notice that the scaling factors (1.75 and 4.0) control how far apart the class 0 and class 1 data are, such that classification is possible but not too easy. It can be seen from the figure that $\mu_1, \mu_{21}, \mu_{41}, \ldots, \mu_{181}$ are much larger in magnitude than the others. The covariance matrix $\boldsymbol{\Sigma}$ has a block-diagonal structure, with block size 20. In each of the 10 diagonal blocks, the elements on the main diagonal are 1.0, while all others are equal to ρ. In all of the simulated data experiments, we choose $\rho = 0.1$. Therefore, among the 200 features, the best 10 features are the 1st, 21st, ..., 181st features, which are mutually independent. Each of the best 10 features is weakly correlated with 19 other nonbest features ($\rho = 0.1$).

The experiments on simulated data are designed for two different sizes of sample $\mathcal{S}$, $N = 50$ and $N = 100$. The size of the independent test data set $\mathcal{S}'$ for getting true error is 5000. Each data point is a random vector with dimensionality 200, and 10 features will be selected by the feature selection algorithm. In all the three models, the numbers of sample points from class 0 and class 1 are equal ($N/2$).

The patient data come from 295 breast tumor microarrays, each obtained from one patient [9, 10] and together yielding 295 log-expression profiles. Based on patient survival data and other clinical measures, 180 data points fall into the "good prognosis" class and 115 fall into the "bad

prognosis" class, the two classes to be labeled 0 and 1, respectively. Each data point is a 70-gene expression vector. The 295 70-expression vectors constitute the empirical sample space, with prior probabilities about 0.6 and 0.4, respectively. For error estimation, we will randomly draw a stratified sample of size 35 (i.e., $\mathcal{S}$) from the 295 data points, without replacement. In the sample, 21 data points belong to class 0, and 14 belong to class 1. From the full set of 70 genes, 7 will be selected for classification, where both k-fold ($k = 7$) and leave-one-out cross-validation will be used for error estimation. The key reason for using this data set is that it is incorporated into a feature-set test bed and the 7 best genes are known for 3NN and LDA, these having been derived from a full search among all possible 7-gene feature sets from the full 70 genes [11]. Since the SVM optimal genes are not derived in the test bed, we will use the LDA best genes to obtain the distribution of ΔE_b. To obtain the true classification error, the remaining $260 = 295 - 35$ data points will constitute $\mathcal{S}'$ and be tested on. Since the size of $\mathcal{S}$ is small, compared to the full dataset of 295, the dependence between two random samples will be negligible (see [2] for an analysis of the dependency issue in the context of this data set).

3. IMPLEMENTATION

We consider three commonly employed classification rules: LDA, 3NN, and SVM. All three are used on all data models, with the exception that only 3NN is applicable to the bimodal model. As stated previously, our method is to compare the cross-validation (k-fold and leave-one-out) deviation distributions for classification rules used with and without feature selection. For feature selection, we use t-test, SFS, and SFFS to select d features from the full feature set. To improve feature selection accuracy, within SFS and SFFS, the feature selection criterion is semibolstered resubstitution error with 3NN classifier, or bolstered resubstitution error with LDA and SVM classifiers [5].

To accomplish our goal, we propose the following experiments on simulated and patient data. Draw a random sample $\mathcal{S}$ of size N from the sample space, select d features on $\mathcal{S}$, and denote the feature set by F. Design a classifier $\mathcal{C}_F$ on $\mathcal{S}$, and test it on a large independent sample $\mathcal{S}'$ to get the true error E. Design a classifier $\mathcal{C}_b$ on $\mathcal{S}$ with the known best feature set F_b, and find the true error E_b by testing it on $\mathcal{S}'$. Obtain the (k-fold or leave-one-out) cross-validation errors $\widehat{E}$ and $\widehat{E}_b$. Compute $\Delta E = \widehat{E} - E$ and $\Delta E_b = \widehat{E}_b - E_b$. Finally, repeat the previous sampling and error estimation procedure 10000 times, and plot the empirical distributions of ΔE and ΔE_b.

A step-by-step description that provides the implementation of proposed experiments is shown in Algorithm 1. We use abbreviations CV and LOO for cross-validation and leave-one-out, respectively.

4. RESULTS AND DISCUSSION

Let us first consider the simulated data. Three classifiers, LDA, 3NN, and SVM, are applied to the simulated data with sample sizes $N = 50$ and $N = 100$, and all three model distributions, with the exception that only 3NN is applicable to the bimodal model. Three feature selection algorithms, t-test, SFS, and SFFS, are employed, with the exception that only SFS and SFFS are applicable to the bimodal model. In each case, two kinds of cross-validation error estimation methods, 10-fold cross-validation (CV10) and leave-one-out (LOO), are used. The complete plots of deviation distributions are provided on the companion website (http://gsp.tamu.edu/web2/quantify_fscv/). Here, Figure 3 shows the deviation distributions for the unequal covariance model using CV10. The plots in Figure 3 are fairly typical.

Tables 1, 2, and 3 list the deviation variances and κ for every model, classifier, and feature selection algorithm. From the tables, we observe that κ is always positive, confirming that feature selection worsens error estimation precision. Please note that since no feature selection is involved in obtaining E_b and $\widehat{E}_b$, ΔE_b is independent of feature selection methods. Therefore, in each row of the tables (with fixed classifier and cross-validation method), we combine the ΔE_b's of the three experiments (t-test, SFS, and SFFS) and compute the overall variance $\mathrm{Var}(\Delta E_b)$ (pooled variance).

When interpreting the results, two related issues need to be kept in mind. First, we are interested in measuring the degree to which feature selection degrades cross-validation performance for different feature selection methods, not the performance of the feature selection methods themselves. In particular, two studies have demonstrated the performance of SFFS [12, 13], and for the linear model with weak correlation we can expect good results from the t-test. Second, since the performance of an error estimator depends on its bias and variance, when choosing between feature selection algorithms we prefer a smaller deviation variance $\mathrm{Var}(\Delta E)$. The results show that a smaller variance of ΔE usually corresponds to a smaller κ, but not strictly so, because κ depends on the variance of ΔE_b too. For instance, with the equal covariance model and t-test, when the sample size is 50 and 10-fold CV is used, the 3NN classifier gives a smaller variance of ΔE than the SVM classifier, whereas its κ is larger than SVM. Be that as it may, the sole point of this study is to quantify the increase in variance owing to feature selection, thereby characterizing the manner in which feature selection impacts upon cross-validation error estimation for combinations of feature selection algorithms and classification rules.

Looking at the results, we see that the degradation in deviation variance owing to feature selection can be striking, especially in the bimodal model, where κ exceeds 0.81 for all cases in Table 3. In the unequal covariance model, for sample size 50, κ generally exceeds 0.45. One can observe differences in the effects of feature selection relative to the classification rule and feature selection algorithm by perusing the tables.

An interesting phenomenon to observe is the effect of increasing the sample size from 50 to 100. In all cases, this significantly reduces the variances, as expected; however, while increased sample size reduces κ for the t-test, there is no similar reduction observed for SFS and SFFS with the unequal covariance model. Perhaps here it would be beneficial to emphasize that the performance of the t-test on the simulated data may be due to the nature of the equal covariance and

(1) Specify the following parameters:
$N_{\mathrm{MC}} = 10000$; /∗ number of Monte Carlo experiments∗/
d; /∗number of features to be selected∗/
N_{sample}; /∗sample size∗/
N_{fold}; /∗ $= k$ if k-fold CV; $= N_{\mathrm{sample}}$ if LOO∗/
best feature set F_b; /∗containing d best features; ∗/
(2) $n_{\mathrm{MC}} = 0$; /∗loop count∗/
(3) while ($n_{\mathrm{MC}} < N_{\mathrm{MC}}$) {

(a) Generate a random sample $\mathcal{S}$ of size N_{sample} from the sample space, with $N_{\mathrm{sample}} * p_0$ data points from class 0, and $N_{\mathrm{sample}} * p_1$ data points from class 1, where p_0 and p_1 are the prior probabilities.
(b) Use the best feature set F_b to design a classifier $\mathcal{C}_b$ on $\mathcal{S}$. Perform feature selection on $\mathcal{S}$ to obtain a feature set F of d features. Use F to design a classifier $\mathcal{C}_F$ on $\mathcal{S}$.
(c) To obtain the true classification errors, generate a large sample $\mathcal{S}'$ independent of $\mathcal{S}$ to test $\mathcal{C}_F$ and $\mathcal{C}_b$, then denote their true errors by E and E_b, respectively.
(d) To do N_{fold}-fold cross-validation, divide the data evenly into N_{fold} portions $\mathcal{T}_0, \ldots, \mathcal{T}_{N_{\mathrm{fold}}-1}$, and in each portion, the numbers of class 0 data and class 1 data are roughly proportional to p_0 and p_1, if possible.
(e) For ($i = 0$; $i < N_{\mathrm{fold}}$; $i++$){

(i) Hold out $\mathcal{T}_i$ as the test sample and use $\mathcal{S} \setminus \mathcal{T}_i$ as the training sample.
(ii) Perform feature selection on the training sample, and the resultant feature set is F_i of size d.
(iii) Apply feature set F_i, and use the training sample to design a surrogate classifier $\mathcal{C}_i$, and test $\mathcal{C}_i$ on $\mathcal{T}_i$ to obtain the estimated error $\hat{E}_i$.
(iv) Repeat step (iii), but use feature set F_b instead, to obtain the surrogate classifier $\mathcal{C}_{b,i}$ and error $\hat{E}_{b,i}$.

}
(f) Find the average errors $\hat{E}$ and $\hat{E}_b$ over the N_{fold} folds.
(g) Compute the differences between the estimated and the true errors,

$$\Delta E = \hat{E} - E,$$
$$\Delta E_b = \hat{E}_b - E_b.$$

(h) $n_{\mathrm{MC}}++$.

}
(4) From the N_{MC} Monte Carlo experiments, plot the empirical distributions of ΔE and ΔE_b, respectively.

Algorithm 1: Simulation scheme

Table 1: Results for simulated data: equal covariance model. For easy reading, the variances are in 10^{-4} unit.

N	Classifier	t-test			SFS			SFFS		
		Var(ΔE)	Var(ΔE_b)	κ	Var(ΔE)	Var(ΔE_b)	κ	Var(ΔE)	Var(ΔE_b)	κ
50	3NN,CV10	25.76	16.48	0.3605	62.79	16.48	0.7376	62.26	16.48	0.7354
	3NN,LOO	26.11	17.05	0.3469	65.05	17.05	0.7378	63.05	17.05	0.7295
	LDA,CV10	32.21	17.48	0.4572	50.84	17.48	0.6561	51.76	17.48	0.6622
	LDA,LOO	30.00	16.35	0.4552	52.59	16.35	0.6892	56.79	16.35	0.7121
	SVM,CV10	35.89	25.21	0.2976	54.76	25.21	0.5397	52.47	25.21	0.5195
	SVM,LOO	38.35	26.38	0.3121	51.81	26.38	0.4908	53.71	26.38	0.5088
100	3NN,CV10	7.96	7.42	0.0677	25.53	7.42	0.7094	25.12	7.42	0.7046
	3NN,LOO	7.53	7.38	0.0197	24.55	7.38	0.6993	24.24	7.38	0.6954
	LDA,CV10	6.55	6.00	0.0841	13.18	6.00	0.5448	13.04	6.00	0.5400
	LDA,LOO	6.18	5.74	0.0716	12.90	5.74	0.5555	13.79	5.74	0.5840
	SVM,CV10	10.29	9.74	0.0538	17.16	9.74	0.4326	16.79	9.74	0.4201
	SVM,LOO	11.20	10.52	0.0611	16.20	10.52	0.3508	15.79	10.52	0.3338

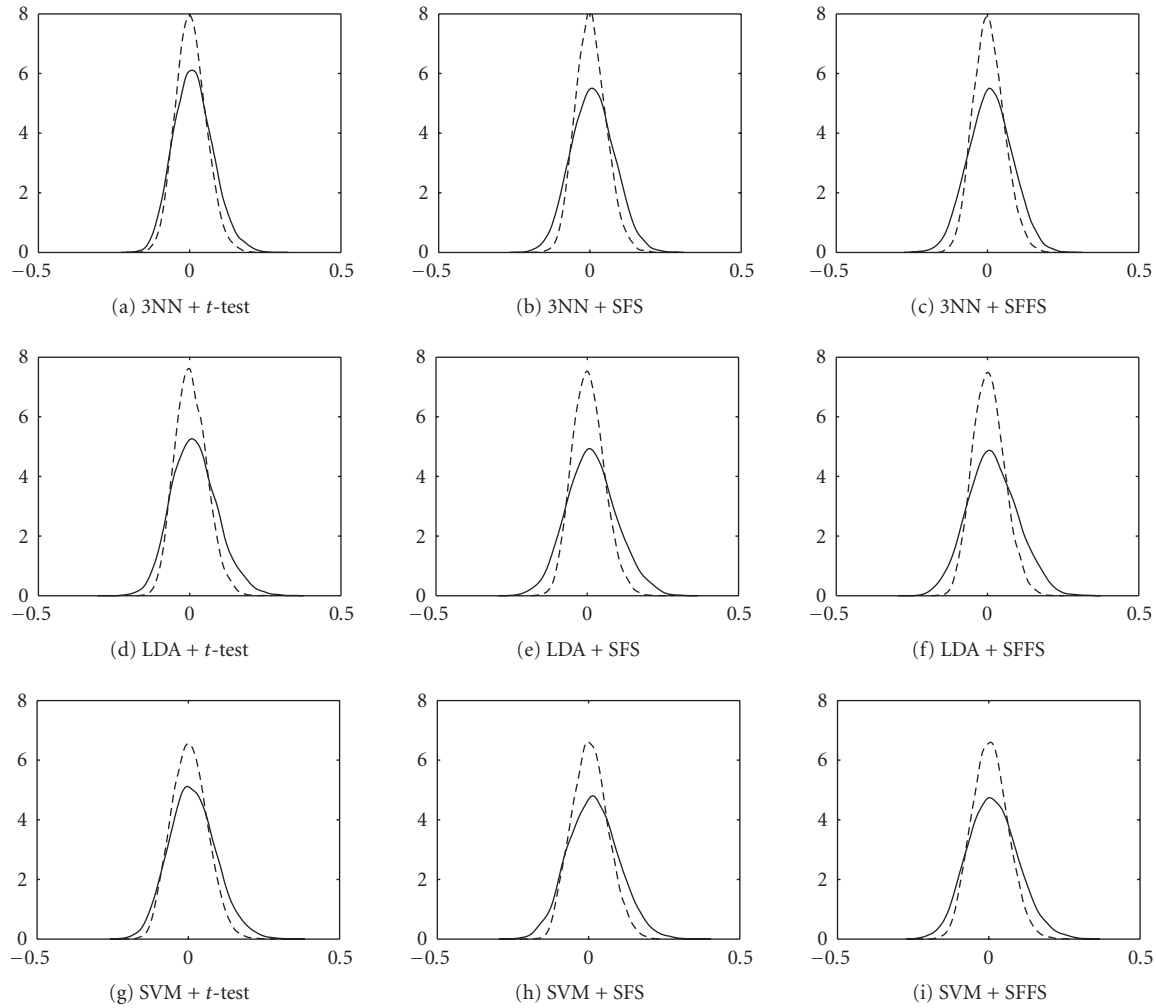

Figure 3: Deviation distributions with feature selection (solid line) and without feature selection (dashed line), unequal covariance model, 10-fold CV with sample size $N = 50$. The x-axis denotes the deviation, and the y-axis corresponds to the density.

Table 2: Results for simulated data: unequal covariance model. For easy reading, the variances are in 10^{-4} unit.

N	Classifier	t-test			SFS			SFFS		
		Var(ΔE)	Var(ΔE_b)	κ	Var(ΔE)	Var(ΔE_b)	κ	Var(ΔE)	Var(ΔE_b)	κ
50	3NN,CV10	41.91	25.61	0.3890	50.24	25.61	0.4904	52.25	25.61	0.5100
	3NN,LOO	46.17	25.69	0.4436	51.93	25.69	0.5054	53.10	25.69	0.5163
	LDA,CV10	57.85	27.16	0.5304	66.44	27.16	0.5912	68.21	27.16	0.6018
	LDA,LOO	61.85	25.33	0.5905	74.46	25.33	0.6598	82.06	25.33	0.6913
	SVM,CV10	60.05	34.85	0.4197	68.71	34.85	0.4929	68.27	34.85	0.4896
	SVM,LOO	70.06	37.23	0.4685	68.67	37.23	0.4578	69.81	37.23	0.4666
100	3NN,CV10	13.75	11.60	0.1562	29.17	11.60	0.6022	28.98	11.60	0.5996
	3NN,LOO	13.97	11.79	0.1560	27.42	11.79	0.5699	28.37	11.79	0.5843
	LDA,CV10	12.67	9.92	0.2170	22.42	9.92	0.5576	22.39	9.92	0.5570
	LDA,LOO	12.77	9.51	0.2556	23.99	9.51	0.6038	25.42	9.51	0.6260
	SVM,CV10	16.88	13.81	0.1816	25.85	13.81	0.4657	25.14	13.81	0.4506
	SVM,LOO	18.74	15.19	0.1895	24.44	15.19	0.3786	23.09	15.19	0.3422

Table 3: Results for simulated data: bimodal model. For easy reading, the variances are in 10^{-4} unit.

Sample size N	Classifier	SFS			SFFS		
		$\mathrm{Var}(\Delta E)$	$\mathrm{Var}(\Delta E_b)$	κ	$\mathrm{Var}(\Delta E)$	$\mathrm{Var}(\Delta E_b)$	κ
50	3NN,CV10	134.80	15.91	0.8820	141.94	15.91	0.8879
	3NN,LOO	116.54	15.72	0.8651	126.08	15.72	0.8754
100	3NN,CV10	47.07	6.77	0.8562	40.94	6.77	0.8346
	3NN,LOO	39.21	6.74	0.8280	36.55	6.74	0.8155

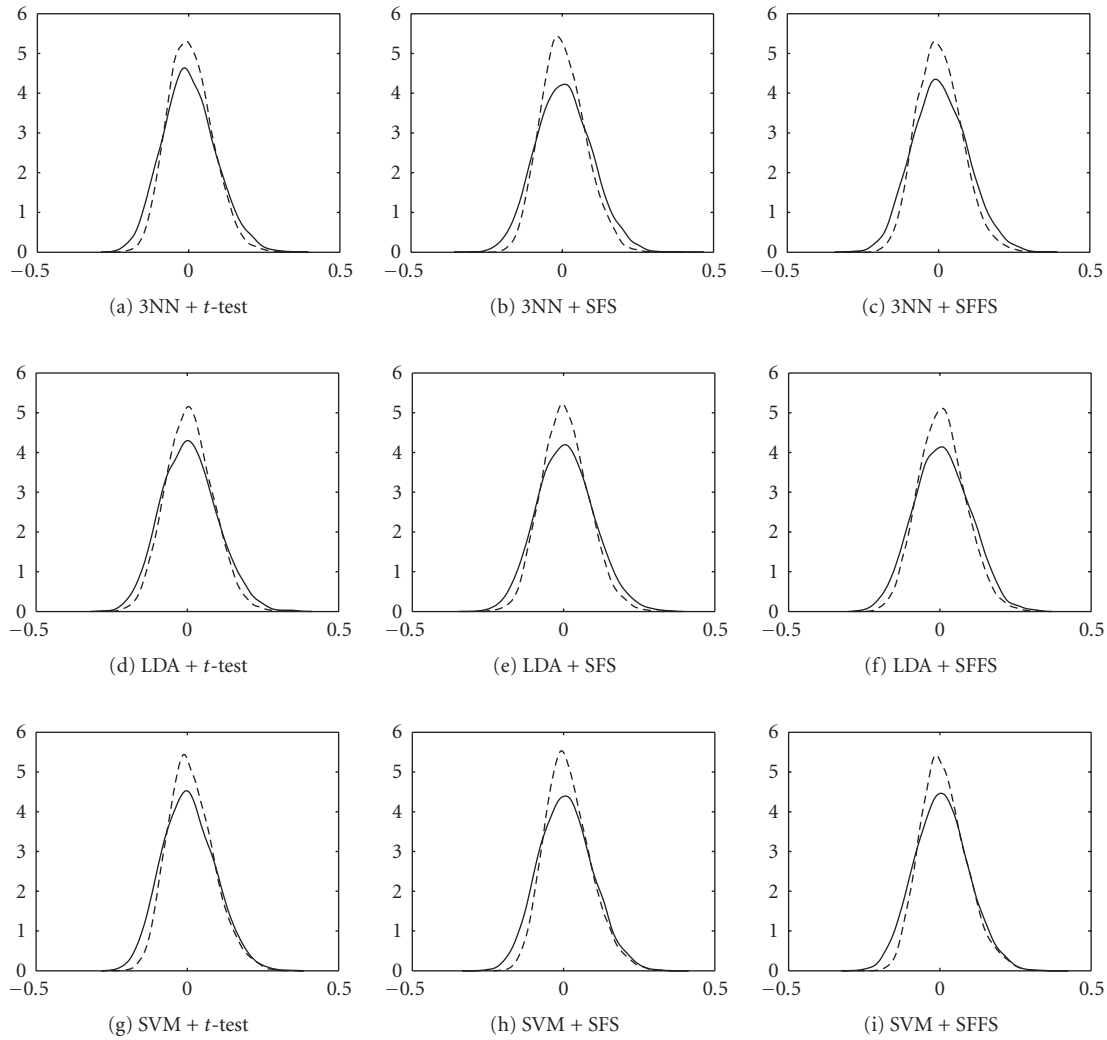

Figure 4: Deviation distributions with feature selection (solid line) and without feature selection (dashed line) for patient data, 7-fold CV. The x-axis denotes the deviation, and the y-axis corresponds to the density.

unequal covariance models: specifically, to obtain the deviation distribution without feature selection, we have to know the optimal feature set from the model, and thus we have chosen the features to be either uncorrelated or weakly correlated, a setting, that is, advantageous for the t-test.

When turning to the patient data (see Table 4, and the pooled variances are used, like in the previous three tables), one is at once struck by the fact that κ is quite consistent across the three-feature selection methods. It differs according to the classification rule and cross-validation procedure, being over 0.4 for all feature selection methods with LDA and LOO, and being below 0.13 for all methods with SVM and LOO; however, the changes between feature selection methods for a given classification rule and cross-validation procedure are very small, as shown clearly in Figure 4. This consistency results in part from the fact that, with the patient

Table 4: Results for patient data. For easy reading, the variances are in 10^{-4} unit.

Classifier	t-test			SFS			SFFS		
	$\mathrm{Var}(\Delta E)$	$\mathrm{Var}(\Delta E_b)$	κ	$\mathrm{Var}(\Delta E)$	$\mathrm{Var}(\Delta E_b)$	κ	$\mathrm{Var}(\Delta E)$	$\mathrm{Var}(\Delta E_b)$	κ
3NN,CV7	77.10	54.74	0.2900	83.88	54.74	0.3474	83.62	54.74	0.3454
3NN,LOO	90.24	56.27	0.3764	93.05	56.27	0.3953	93.39	56.27	0.3975
LDA,CV7	84.85	60.89	0.2824	85.72	60.89	0.2896	86.05	60.89	0.2923
LDA,LOO	99.49	56.98	0.4273	96.89	56.98	0.4120	95.82	56.98	0.4054
SVM,CV7	74.75	57.92	0.2252	78.47	57.92	0.2620	78.12	57.92	0.2586
SVM,LOO	96.10	84.45	0.1212	94.92	84.45	0.1103	95.04	84.45	0.1114

Table 5: Squared biases for simulated data: equal covariance model. The squared biases are in 10^{-4} unit, the same as deviation variances.

N	Classifier	t-test		SFS		SFFS	
		$\mathrm{Mean}^2(\Delta E)$	$\mathrm{Mean}^2(\Delta E_b)$	$\mathrm{Mean}^2(\Delta E)$	$\mathrm{Mean}^2(\Delta E_b)$	$\mathrm{Mean}^2(\Delta E)$	$\mathrm{Mean}^2(\Delta E_b)$
50	3NN,CV10	0.58	0.03	1.72	0.03	1.91	0.03
	3NN,LOO	0.33	0.14	0.73	0.14	0.74	0.14
	LDA,CV10	1.24	0.14	2.10	0.14	2.17	0.14
	LDA,LOO	0.13	0.01	0.13	0.01	0.21	0.01
	SVM,CV10	1.02	0.10	2.19	0.10	1.78	0.10
	SVM,LOO	0.22	0.06	0.47	0.06	0.52	0.06
100	3NN,CV10	0.03	0.01	0.82	0.01	0.69	0.01
	3NN,LOO	0.01	0.01	0.12	0.01	0.12	0.01
	LDA,CV10	0.07	0.02	0.47	0.02	0.42	0.02
	LDA,LOO	0.00	0.00	0.00	0.00	0.01	0.00
	SVM,CV10	0.11	0.04	0.47	0.04	0.51	0.04
	SVM,LOO	0.00	0.00	0.08	0.00	0.06	0.00

data, we are concerned with a single feature-label distribution. On the other hand, the consistency is also due to the similar effects on error estimation of the different feature selection methods with this feature-label distribution, a distribution in which there are strong correlations among some of the features (gene expressions).

Our interest is in quantifying the increase in variance resulting from feature selection; nevertheless, since the mean-squared error of an error estimator equals the sum of the variance and the squared bias, one might ask whether feature selection has a significant impact on the bias. Given that the approximate unbiasedness of cross-validation applies to the classification rule and that feature selection is part of the classification rule, we would not expect a significant effect on the bias. This expectation is supported by the curves in the figures, since the means of the with- and without-feature-selection deviation curves tend to be close. We should, however, not expect these means to be identical, because the exact manner in which the expectation of the error estimate approximates the true error depends upon the classification rule and sample size. To be precise, for k-fold cross-validation with feature selection, the bias is given by

$$\mathrm{Bias}_{N,k}^{FS(D,d)} = E\left[\varepsilon_{N-N/k}^{FS(D,d)}\right] - E\left[\varepsilon_{N}^{FS(D,d)}\right], \tag{2}$$

where $\varepsilon_{N,k}^{FS(D,d)}$ denotes the error for the classification rule when incorporating feature selection to choose d from among D features based on a sample size of N. Without feature selection, the bias is given by

$$\mathrm{Bias}_{N,k}^{(d)} = E\left[\varepsilon_{N-N/k}^{(d)}\right] - E\left[\varepsilon_{N}^{(d)}\right], \tag{3}$$

where $\varepsilon_{N,k}^{(d)}$ denotes the error for the classification rule without feature selection using d features based on a sample size of N. The bias (difference in expectation) depends upon the classification rule, including whether or not feature selection is employed.

To quantify the effect of feature selection on bias, we have computed the squared biases of the estimated errors, both with and without feature selection (namely, the squared means of ΔE and ΔE_b), for the cases considered. Squared biases are computed because they appear in the mean-squared errors. These are given in Tables 5, 6, 7, and 8, corresponding to Tables 1, 2, 3, and 4, respectively. For the model-based data from the equal and unequal covariance models, we see in Tables 5 and 6 that the bias tends to be a bit larger with feature selection, but the squared bias is still negligible in comparison to the variance, the squared biases tending to be very small when $N = 100$. A partial exception occurs for the bimodal model when there is feature selection. In Table 7, we see that, for SFS and SFFS, $\mathrm{mean}^2(\Delta E) > 7 \times 10^{-4}$ for 3NN, CV10, and $N = 50$. Even here, the squared biases are small

Table 6: Squared biases for simulated data: unequal covariance model. The squared biases are in 10^{-4} unit, the same as deviation variances.

N	Classifier	t-test		SFS		SFFS	
		$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$
50	3NN,CV10	1.58	0.19	0.99	0.19	1.02	0.19
	3NN,LOO	0.81	0.33	0.69	0.33	0.83	0.33
	LDA,CV10	2.81	0.21	2.29	0.21	2.96	0.21
	LDA,LOO	0.19	0.02	0.22	0.02	0.35	0.02
	SVM,CV10	2.29	0.20	2.77	0.20	1.96	0.20
	SVM,LOO	0.56	0.07	0.87	0.07	0.99	0.07
100	3NN,CV10	0.35	0.05	0.88	0.05	1.02	0.05
	3NN,LOO	0.07	0.04	0.08	0.04	0.21	0.04
	LDA,CV10	0.31	0.04	0.91	0.04	0.82	0.04
	LDA,LOO	0.00	0.00	0.04	0.00	0.02	0.00
	SVM,CV10	0.48	0.07	0.91	0.07	0.85	0.07
	SVM,LOO	0.02	0.02	0.07	0.02	0.05	0.02

Table 7: Squared biases for simulated data: bimodal model. The squared biases are in 10^{-4} unit, the same as deviation variances.

Sample size N	Classifier	SFS		SFFS	
		$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$
50	3NN,CV10	7.27	0.10	8.31	0.10
	3NN,LOO	1.68	0.12	2.36	0.12
100	3NN,CV10	3.24	0.02	2.26	0.02
	3NN,LOO	0.32	0.02	0.06	0.02

in comparison to the corresponding variances, where we see in Table 3 that $\text{Var}(\Delta E) > 134 \times 10^{-4}$ for both SFS and SFFS. Finally, we note that for the patient data in Table 8 we have omitted SVM because we have used the LDA optimal features from the test bed and therefore the relationship between the bias with and without feature selection is not directly interpretable.

5. CONCLUSION

We have introduced the coefficient of relative increase in deviation dispersion to quantify the effect of feature selection on cross-validation error estimation. The coefficient measures the relative increase in the variance of the deviation distribution due to feature selection. We have computed the coefficient for the LDA, 3NN, and linear SVM classification rules, using three feature selection algorithms, t-test, SFS, and SFFS, and two cross-validation methods, k-fold and leave-one-out. We have applied the coefficient to several feature-label models and patient data from a breast cancer study. The models have been chosen so that the optimal feature set is directly obtainable from the model and the feature-selection test bed provides the best feature sets for the patient data.

Any factor that can influence error estimation and feature selection can influence the CRIDD, and these are numerous: the classification rule, the feature-selection algorithm, the cross-validation procedure, the feature-label distribution, the total number of potential features, the number of useful features among the total number available, the prior class probabilities, and the sample size. Moreover, as is typical in classification, there is interaction among these factors. Our purpose in this paper has been to introduce the CRIDD and, to this end, we have examined a number of combinations of these factors using both model and patient data in order to illustrate how the CRIDD can be utilized in particular situations. Assuming one could overcome the computational impediment, an objective of future work would be to carry out a rigorous study of the factors affecting the manner in which feature-selection impacts cross-validation error estimation, perhaps via an analysis-of-variance approach applied to the factors affecting the CRIDD.

This having been said, we would like to specifically comment on two issues for future study. The first concerns the modest feature-set sizes considered in this study relative to the number of potential features often encountered in practice, such as the thousands of genes on an expression microarray. The reason for choosing the feature-set sizes used in the present paper is because of the extremely long computation times involved in a general study. Even using our Beowulf cluster, computation time is prohibitive when so many cases are being studied. It is reasonable to conjecture that the increased cross-validation variance owing to feature selection that we have observed will hold, or increase, when larger numbers of potential features are observed; however, the exact manner in which this occurs will depend on the

Table 8: Squared biases for patient data. The squared biases are in 10^{-4} unit, the same as deviation variances.

Classifier	t-test		SFS		SFFS	
	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$	$\text{Mean}^2(\Delta E)$	$\text{Mean}^2(\Delta E_b)$
3NN,CV7	0.08	0.19	0.39	0.19	0.34	0.19
3NN,LOO	1.78	0.81	1.10	0.81	1.28	0.81
LDA,CV7	0.47	0.53	0.52	0.53	0.77	0.53
LDA,LOO	0.17	0.04	0.34	0.04	0.08	0.04

proportion of useful features among the potential features and the nature of the feature-label distributions involved. Owing to computational issues, one might have to be contented with considering special cases of interest, rather than looking across a wide spectrum of conditions. As a counterpoint to this cautionary note, one needs only to recognize the recent extraordinary expansion of computational capability in bioinformatics.

A second issue concerns the prior probabilities of the classes. In this study (and common among many classification studies), for both synthetic and patient data, the classes are either equiprobable or close to equiprobable. In the case of small samples, when the prior probabilities are substantially unbalanced, feature selection becomes much harder, and we expect that variation in error estimation will grow and this will be reflected in a larger CRIDD. There are two codicils to this point: (1) the exact nature of the unbalanced effect will depend on the feature-label distributions, feature-selection algorithm, and the other remaining factors, and (2) when there is severe lack of balance between the classes, the overall classification error rate may not be a good way to measure practical classification performance—for instance, with extreme unbalance, good classification results from simply choosing the value of the dominant class no matter the observation—and hence the whole approach discussed in this study may not be appropriate.

APPENDICES

A. FEATURE SELECTION METHODS: SFS AND SFFS

A common approach to suboptimal feature selection is sequential selection, either forward or backward, and their variants. Sequential forward selection (SFS) begins with a small set of features, perhaps one, and iteratively builds the feature set. When there are k features, $x_1, x_2, \ldots, x_k$, in the growing feature set, all feature sets of the form $\{x_1, x_2, \ldots, x_k, w\}$ are compared and the best one is chosen to form the feature set of size $k+1$. A problem with SFS is that there is no way to delete a feature adjoined early in the iteration that may not perform as well in combination as other features. The SFS look-back algorithm aims to mitigate this problem by allowing deletion. For it, when there are k features, $x_1, x_2, \ldots, x_k$, in the growing feature set, all feature sets of the form $\{x_1, x_2, \ldots, x_k, w, z\}$ are compared and the best one is chosen. Then all $(k+1)$-element subsets are checked to allow the possibility of one of the earlier chosen features to be deleted, the result being the $k+1$ features that will form the basis for the next stage of the algorithm. Flexibility is added with the sequential forward floating selection (SFFS) algorithm, where the number of features to be adjoined and deleted is not fixed [7]. Simulation studies support the effectiveness of SFFS [12, 13]; however, with small samples SFFS performance is significantly affected by the choice of error estimator used in the selection process, with bolstered error estimators giving comparatively good results [5].

B. CROSS-VALIDATION ERROR

In two-group statistical pattern recognition, there is a *feature vector* $X \in \mathbb{R}^p$ and a *label* $Y \in \{0,1\}$. The joint probability distribution $\mathbf{F}$ of (X,Y) is unknown in practice. Hence, one has to design classifiers from *training data*, which consists of a set of n independent observations, $S_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, drawn from $\mathbf{F}$. A *classification rule* is a mapping $g : \{\mathbb{R}^p \times \{0,1\}\}^n \times \mathbb{R}^p \to \{0,1\}$. A classification rule maps the training data S_n into the *designed classifier* $g(S_n,\cdot) : \mathbb{R}^p \to \{0,1\}$. The *true error* of a designed classifier is its error rate given the training data set

$$\epsilon_n[g \mid S_n] = P(g(S_n,X) \neq Y) = E_{\mathbf{F}}(\,|\, Y - g(S_n,X)\,|\,), \tag{B.1}$$

where the notation $E_{\mathbf{F}}$ indicates that the expectation is taken with respect to $\mathbf{F}$; in fact, one can think of (X,Y) in the above equation as a random test point (this interpretation being useful in understanding error estimation). The expected error rate over the data is given by

$$\epsilon_n[g] = E_{\mathbf{F}_n}(\epsilon_n[g \mid S_n]) = E_{\mathbf{F}_n} E_{\mathbf{F}}(\,|\, Y - g(S_n,X)\,|\,), \tag{B.2}$$

where $\mathbf{F}_n$ is the joint distribution of the training data S_n. This is sometimes called the *unconditional error* of the classification rule, for sample size n.

In *k-fold cross-validation*, the data set S_n is partitioned into k *folds* $S_{(i)}$, for $i = 1,\ldots,k$ (for simplicity, we assume that k divides n). Each fold is left out of the design process and used as a test set, and the estimate is the overall proportion of error committed on all folds:

$$\hat{\epsilon}_{\text{cvk}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n/k} |\, y_j^{(i)} - g(S_n \setminus S_{(i)}, x_j^{(i)})\,|, \tag{B.3}$$

where $(x_j^{(i)}, y_j^{(i)})$ is a sample in the ith fold. The process may be repeated: several cross-validation estimates are computed using different partitions of the data into folds, and

the results are averaged. A k-fold cross-validation estimator is unbiased as an estimator of $\epsilon_{n-n/k}[g]$. The most well-known cross-validation method is the *leave-one-out estimator*, whereby a single observation is left out each time:

$$\hat{\epsilon}_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - g\left(S_{n-1}^{i}, x_i\right) \right|, \tag{B.4}$$

where S_{n-1}^{i} is the data set resulting from deleting data point i from the original data set S_n. This corresponds to n-fold cross-validation.

ACKNOWLEDGMENT

This research has been supported in part by the National Science Foundation, Grants CCF-0514644 and BES-0536679.

REFERENCES

[1] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, NY, USA, 1996.

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

[3] U. Braga-Neto and E. R. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37, no. 6, pp. 1267–1281, 2004.

[4] C. Sima, U. Braga-Neto, and E. R. Dougherty, "Superior feature-set ranking for small samples using bolstered error estimation," *Bioinformatics*, vol. 21, no. 7, pp. 1046–1054, 2005.

[5] C. Sima, S. Attoor, U. Brag-Neto, J. Lowey, E. Suh, and E. R. Dougherty, "Impact of error estimation on feature selection," *Pattern Recognition*, vol. 38, no. 12, pp. 2472–2482, 2005.

[6] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.

[7] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[8] Y. Xiao, J. Hua, and E. R. Dougherty, "Feature selection increases cross-validation imprecision," in *Proceedings of the 4th IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '06)*, College Station, Tex, USA, May 2006.

[9] L. J. van't Veer, H. Dai, M. J. van de Vijver, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[10] M. J. van de Vijver, Y. D. He, L. J. van't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[11] A. Choudhary, M. Brun, J. Hua, J. Lowey, E. Suh, and E. R. Dougherty, "Genetic test bed for feature selection," *Bioinformatics*, vol. 22, no. 7, pp. 837–842, 2006.

[12] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[13] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, 2000.