BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 22 2014, pages 3152–3158
doi:10.1093/bioinformatics/btu520

Genome analysis

Advance Access publication August 1, 2014

Bias correction for selecting the minimal-error classifier from many machine learning models

Ying Ding[1,2,†], Shaowu Tang[2,†], Serena G. Liao[2], Jia Jia[2], Steffi Oesterreich[3], Yan Lin[2] and George C. Tseng[1,2,*]

[1]Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology, [2]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA and [3]Magee-Womens Research Institute, Pittsburgh, PA 15213, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Supervised machine learning is commonly applied in genomic research to construct a classifier from the training data that is generalizable to predict independent testing data. When test datasets are not available, cross-validation is commonly used to estimate the error rate. Many machine learning methods are available, and it is well known that no universally best method exists in general. It has been a common practice to apply many machine learning methods and report the method that produces the smallest cross-validation error rate. Theoretically, such a procedure produces a selection bias. Consequently, many clinical studies with moderate sample sizes (e.g. $n = 30$–60) risk reporting a falsely small cross-validation error rate that could not be validated later in independent cohorts.

Results: In this article, we illustrated the probabilistic framework of the problem and explored the statistical and asymptotic properties. We proposed a new bias correction method based on learning curve fitting by inverse power law (IPL) and compared it with three existing methods: nested cross-validation, weighted mean correction and Tibshirani-Tibshirani procedure. All methods were compared in simulation datasets, five moderate size real datasets and two large breast cancer datasets. The result showed that IPL outperforms the other methods in bias correction with smaller variance, and it has an additional advantage to extrapolate error estimates for larger sample sizes, a practical feature to recommend whether more samples should be recruited to improve the classifier and accuracy. An R package 'MLbias' and all source files are publicly available.

Availability and implementation: tsenglab.biostat.pitt.edu/software.htm.

Contact: ctseng@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 21, 2013; revised on July 10, 2014; accepted on July 26, 2014

1 INTRODUCTION

In the past two decades, fast development in bioinformatics was accompanied by the rapid production of high-throughput genomic data, such as gene expression, genotyping and various types of next-generation sequencing data. Such high-dimensional data usually come with small sample sizes and a large number of genes/features (also known as 'large p, small n' problem) and pose many new challenges in statistical learning and data mining. In the content below, we focus on machine learning of gene expression profile data, but the concept and theoretical issues also apply to other high-throughput genomic (e.g. copy number variation, DNA methylation) or proteomic data. In gene expression profile analysis, it is of great interest to predict or diagnose a disease status (e.g. classify cases versus controls or treatment responders versus non-responders). Because no universally best machine learning method exists in general (Allison *et al.*, 2006), to fulfill this task, multiple models are often constructed with different combinations of features (genes), different machine learning methods as well as different tuning parameters in the methods. To choose among such a large number of classifiers (models), it is common practice to select the model with the smallest cross-validation error rate, called the minimal-error classifier (MEC), and report its associated error rate.

The MEC error rate is, however, generally downward biased and an overly optimistic estimator of the true optimal classification error rate. This is because taking the minimum of cross-validation error rates, where the estimates are random variables, will inevitably yield a downward bias. Such a selection bias has great adverse impact in many biomedical pilot studies with moderate sample sizes (e.g. $n = 30$–60). The problem, however, has often been overlooked in applications. For example, one can examine the small pilot data using ~10 popular machine learning methods, and simultaneously choose among many different numbers of features and tuning parameters in each method. This easily increases the number of tested classifiers to several hundreds and selects the MEC with a falsely small error rate because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. In the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) example that will be demonstrated in Section 3.3, we will show that the MEC bias can mistakenly reduce the error rate from 28.2% to an overly optimistic 19.1% in early- and late-stage classification (i.e. a –9.1% error rate bias). Many researchers have recognized this problem (Bernau *et al.*, 2013; Berrar *et al.*, 2006; Efron, 2009; Fu *et al.*, 2005; Tibshirani and Tibshirani, 2009; Varma and Simon, 2006; Wood *et al.*, 2007). Dupuy and Simon (2007) recommended to 'report the estimates

*To whom correspondence should be addressed.
†The authors wish it be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

for all the classification algorithms not just the minimal error rate'. Some proposed comparing the minimal error rate with the median error rate from the original datasets with permuted class labels (Boulesteix and Strobl, 2009). These suggestions, however, did not provide a real solution. Yousefi *et al*. (2011) proposed an approach to estimate the bias by a multivariate Gaussian distribution assumption between the minimal estimated error rate and the true minimal error rate. The Gaussian assumption is, however, generally questionable, and the method may not be accurate with a small sample size.

In this article, three applicable bias correction methods proposed in the literature will be compared with our proposed inverse power law (IPL) method: nested cross-validation (nestedCV) (Varma and Simon, 2006), Tibshirani-Tibshirani procedure (TT) (Tibshirani and Tibshirani, 2009) and weighted mean correction methods (WMC/WMCS) (Bernau *et al*., 2013). Tibshirani and Tibshirani proposed a simple bias estimation method that is computationally efficient and could be calculated through a traditional K-fold cross-validation. They claimed that the bias is only an issue when p $\gg$ n where p is the number of genes and n is the number of samples. The nestedCV, proposed by Varma and Simon, introduced another outer loop of cross-validation, so that the model selection stage is wrapped in the training samples of the outer loop. This double loop procedure, which amounts to nested double leave-one-out cross-validation (LOOCV), is computationally expensive with complexity of $O(n^2)$. WMC/WMCS (Bernau *et al*., 2013) was proposed as a smooth analytical alternative to nestedCV based on subsampling, which yielded a competitive estimate compared with nestedCV at a much lower computational price. Theoretically, according to Bernau *et al*. (2013), the TT method does not apply subsampling in the bias correction, and its estimation target is the conditional error rate (conditional on the given samples). The nestedCV and WMC/WMCS (and the IPL method we will propose) target on the unconditional error rate by repeated subsampling. In addition, both nestedCV and WMC/WMCS methods target on the error rate of a wrapper algorithm (multiple algorithms and/or rules to decide which one shall be used), which is slightly different from the MEC error rate we discuss in this article. We, however, compare all methods side-by-side because biologically they all conceptually aim to correct MEC bias from many machine learning models.

This article is structured as follows. We first illustrate the MEC bias by a 2D toy example and discuss its asymptotic theory and statistical properties. The performance of the nestedCV, WMC/WMCS and TT will be examined. A subsampling-based IPL method will be proposed for the bias correction and compared with the three existing methods in both simulated and real datasets. In real data evaluation, we will use five Gene Expression Omnibus (GEO) datasets and two large breast cancer datasets [the Cancer Genome Atlas (TCGA) and METABRIC].

2 METHODS

2.1 Problem setting and formulation under simulation scheme

Assume that an observed dataset D with sample size n and number of features p is to be analyzed for machine learning. Assume that D is generated from an underlying data distribution Λ_n. $M \geq 2$ classification methods are used to learn a good classifier for future prediction. M can be large (e.g. several hundred), as different feature selection or different parameter settings under a machine learning method are considered different classifiers. Suppose the unknown true error rate of classification method m for data distribution Λ_n is $P_{n,m}$. The theoretical best machine learning method for data distribution Λ_n is $m^* = argmin_m P_{n,m}$ and the resulting error rate is $P^*_{n,M} = min_{1\leq m\leq M} P_{n,m}$ (Box A of Fig. 1).

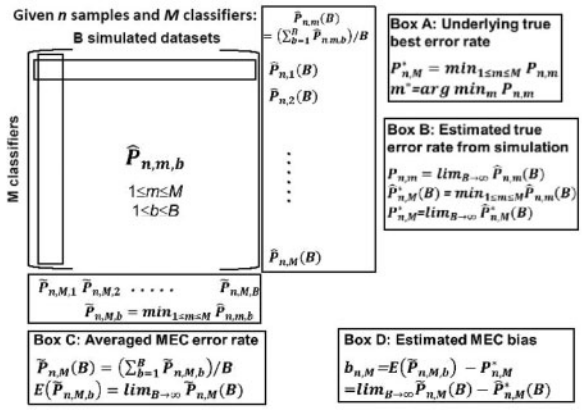

Fig. 1. The framework of estimating the true optimal classification error rate, MEC error rate and its bias from independently simulated datasets

We will illustrate the problem in a simulation framework in Figure 1 because the underlying truth and error rates can be estimated well from repeated simulations. Suppose B datasets $D_b (1 \leq b \leq B)$ are independently generated from Λ_n. We use a cross-validated error rate (from LOOCV) to approximate the error rate for D_b of sample size n using method m, denoted as $\hat{P}_{n,m,b}$. As $\hat{\boldsymbol{P}}_{\boldsymbol{n,m,b}}$ is the LOOCV error rate, it is almost an unbiased error estimator (i.e. $\boldsymbol{E}(\hat{\boldsymbol{P}}_{\boldsymbol{n,m,b}}) \cong \boldsymbol{P}_{\boldsymbol{n,m}}$). We have $P_{n,m} = \lim_{B\to\infty} \hat{P}_{n,m}(B)$, where $\hat{P}_{n,m}(B) = (\sum_{b=1}^{B} \hat{P}_{n,m,b})/B$. Denote by $\hat{P}^*_{n,M}(B) = min_{1\leq m\leq M} \hat{P}_{n,m}(B)$. We will show later that $P^*_{n,M} = min_{1\leq m\leq M} P_{n,m} = \lim_{B\to\infty} \hat{P}^*_{n,M}(B)$ (see Box B of Fig. 1). In other words, the true best classifier error rate $P^*_{n,M}$ can be estimated by $\hat{P}^*_{n,M}(B)$ when many datasets (i.e. B is large) can be repeatedly simulated from Λ_n. In real data analysis, such repeated simulation is, however, not possible. When a single simulated dataset D_b is given, the MEC is chosen by the minimal error rate: $\tilde{m}_b^{(MEC)} = argmin_{1\leq m\leq M} \hat{P}_{n,m,b}$ and $\tilde{P}_{n,M,b} = min_{1\leq m\leq M} \hat{P}_{n,m,b}$. The expected value of the MEC error rate can be estimated as $E(\tilde{P}_{n,M,b}) = \lim_{B\to\infty} \tilde{P}_{n,M}(B)$, where $\tilde{P}_{n,M}(B) = (\sum_{b=1}^{B} \tilde{P}_{n,M,b})/B$ (Box C in Fig. 1). Finally, the bias of the MEC error rate can be estimated as $b_{n,M} = E(\tilde{P}_{n,M,b}) - P^*_{n,M} = \lim_{B\to\infty} \tilde{P}_{n,M}(B) - \hat{P}^*_{n,M}(B)$, where the first term is the estimated expectation of the MEC error rate and the second term is the estimated true best classifier error rate. In Section 2.2, a 2D toy example will be used to demonstrate the issue and properties of the MEC bias $b_{n,M}$. In Section 2.3, we will show that $E(\tilde{P}^*_{n,M,b}) < P^*_{n,M}$ is always true and the MEC error rate is always downward (optimistically) biased (i.e. $b_{n,M} < 0$).

2.2 Illustration by a 2D toy model

Below we present the problem in a 2D toy simulation model. Although the simple simulation model is not intended to mimic a real gene expression profile setting, it illustrates the MEC bias issue with known underlying truth. We simulate $B = 1000$ training sets $D_1, \cdots, D1000$ from Λ_n where Λ_n contains $n = (20, 30, 40, 60, 80, 120, 160, 320, 640, 1280, 2560)$ data points in a 2D Euclidean space. Data points from two equal-size classes are simulated, one (with $n/2$ data points) from $N(\binom{0}{1}, \begin{pmatrix}3 & 1\\ 1 & 1\end{pmatrix})$, and the other from $N(\binom{0}{-2}, \begin{pmatrix}1 & 1\\ 1 & 3\end{pmatrix})$. $M = 10$ classifiers are applied to each dataset: k-nearest neighbors (KNN) with $k = 1, 3, 5$, diagonal linear

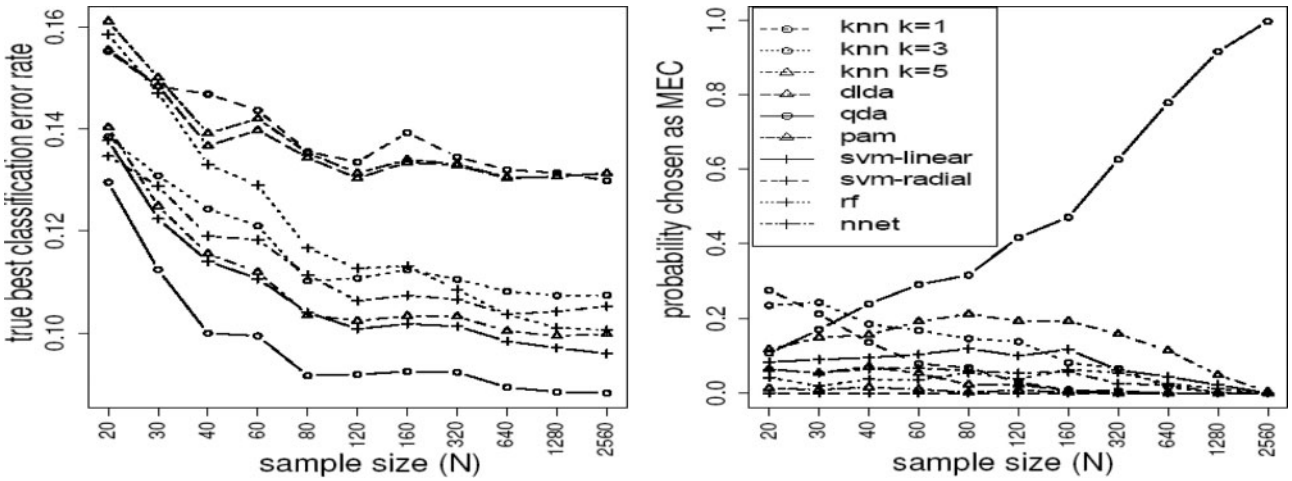

Fig. 2. Left: The estimated classification error rate of each method at different sample sizes from 1000 independent simulation of the 2D toy model in which QDA was the top classifier in this context. Right: Probability for each classifier to be chosen as the minimal error rate classifier in the 1000 simulations

discriminant analysis (DLDA), quadratic discriminant analysis (QDA), shrunken centroids discriminant analysis (SCDA), support vector machines (SVM-linear) with linear kernel, support vector machines (SVM-non-linear) with radial basis kernel, random forest (RF) and neural networks with one hidden layer (NNET). The R package Classification for MicroArrays (*CMA*) (Slawski *et al.*, 2008) is applied to implement all the classification methods. Given the Gaussian assumptions and non-identical covariance matrixes in two classes, QDA is expected to be the optimal Bayes classifier, and the best decision boundary is of a quadratic form. Figure 2, left, shows the averaged error rate for classification method m at sample size n [i.e. $\hat{P}_{n,m}(B)$] estimated from B = 1000 independently simulated datasets. As expected, QDA (solid line) is the optimal Bayes classifier and has the lowest error rates for all different n. However, in real data analysis, we are given only one observed dataset. Figure 2, right, shows the probability of the methods chosen as the error classifier [MEC; i.e. $\tilde{m}_b^{(MEC)}$] for different n. When the sample size is small, MEC may not necessarily select the best method QDA. Particularly, when n = 20 and 30, KNN methods (K = 1 with dashed line and K = 3 with dotted line) generate a smaller error rate than QDA with higher probability. When sample size becomes large ($n > 160$), the dataset contains enough information for QDA to be dominantly (>50% probability) selected by MEC. In Figure 3a, the true minimal error rate $\hat{P}^*_{n,M}(B) \cong P^*_{n,M}$ from QDA (cross) and expectation of MEC error rate $\tilde{P}_{n,M}(\mathrm{B}) \cong E(\tilde{P}_{n,M,b})$ (circle) are shown for different n. In Figure 3b, the estimated MEC biases $\hat{b}_{n,M}(\mathrm{B}) = \tilde{P}_{n,M}(\mathrm{B}) - \hat{P}^*_{n,M}(B)$ are shown for different n. The result clearly demonstrates a downward bias of the MEC error rate, and the bias is greater for small sample sizes and diminishes to zero when sample size is large. It is notable that the bias can be up to 3–5% for n = 20–30. Figure 3c shows the MEC bias for different n when the number of classifiers examined reduces from 10 methods (circle) to four methods (QDA, LDA, SVM-linear and SVM-non-linear) (triangle) or two methods (QDA and LDA) (cross). The result shows that the bias increases as more methods were compared. To our knowledge, only few studies have recognized the increasing trend of MEC bias magnitude when sample size is small or when the searching space of machine learning methods is large. For example, Boulesteix and Strobl's indirectly showed that bias increases with decreasing sample size by showing a non-decreasing trend of the MEC error rate when sample size gets large.

To further study the relationship between the bias and number of classifiers used, we fix n = 20, starting by using the classifier of KNN with k = 1 and keep adding one classifier at a time until all 10 classifiers

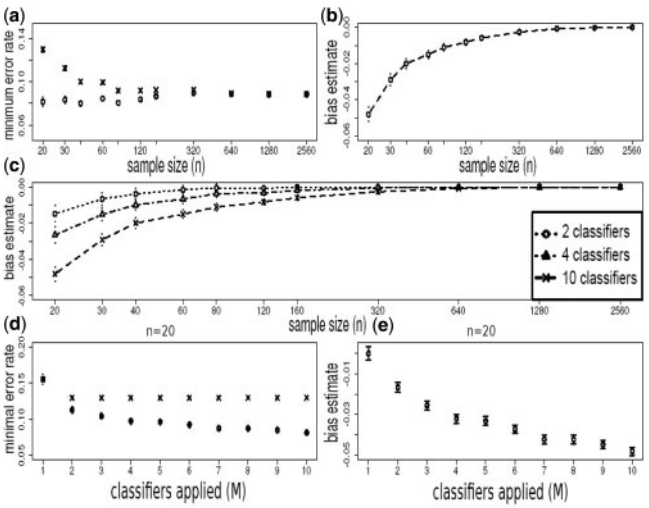

Fig. 3. Illustration of the downward bias of MEC error rate. (a) Trend of true optimal classification error rate (cross) and MEC error rate (circle) as sample size increased. (b) Bias estimate of the MEC error rate diminishes as the sample size increased. (c) Bias estimate when using all 10 classifiers (cross), two classifiers QDA and DLDA (circle) and four classifiers QDA, DLDA, SVM-linear and SVM-non-linear (triangle). (d) For a fixed sample size n = 20, the true minimal error rate (cross) and bias of MEC error rates (circle) as more classifiers were added in the sequence of KNN k = 1, QDA, DLDA, KNN k = 5, PAM, KNN K = 3, SVM-linear, SVM-radial basis, RF and NNET. (e) The corresponding bias as the number of classifiers used increased from 2 to 10

are used. The true optimal classification error rates $\hat{P}^*_{n,M}(B)$ (cross) and MEC error rate $\tilde{P}_{n,M}(\mathrm{B})$(circle) (Fig. 3d) and the MEC error rate biases (Fig. 3e) for different number of classifiers are shown. The result shows that the bias increased from ~2% for two classifiers (KNN K = 1 and QDA) to ~5% for 10 classifiers, while the true optimal error rate $\hat{P}^*_{n,M}(B)$ (cross) no longer decreases once QDA is added as the second classifier. This result indicates that including additional low-performing classifiers

inflates the bias to the extent that may not be compensated by the decrease of the true best classification error rate. In other words, examining too many classifiers and choosing the best is not a good practice, especially if the added classifiers are likely not the top performers. Therefore, caution is called in the small sample size regime when reporting the minimal error rate from multiple classifiers in practice, and it is advantageous if the optimal classifiers can be applied as early as possible without adding more low-performing classifiers. The theorems in the next subsection show that the increasing magnitude of bias for small sample sizes or large numbers of examined classifiers are common statistical properties in data analysis.

2.3 Properties and asymptotic theorems of MEC bias

The proofs of the following theorems are included in the Appendix:

THEOREM 1. Given a smaller set of classifiers, adding more classifiers will decrease the true best classification error rate (i.e. $P^*_{n,M_1} \leq P^*_{n,M_2}$ if $M_1 > M_2$).

THEOREM 2. For a given observed dataset $D^{(n)}$ from Δ_n, $E(\tilde{P}^*_{n,M}) < P^*_{n,M}$, where $\tilde{P}^*_{n,M} = min_{1\leq m\leq M} \hat{P}_{n,m}$ and $\hat{P}_{n,m}$ is the cross-validation error rate of $D^{(n)}$ using classifier m. In other words, the bias of MEC error rate $b_{n,M} = E(\tilde{P}^*_{n,M}) - P^*_{n,M}$ is strictly <0.

THEOREM 3. For observed datasets $D^{(n)}$ from Δ_n of varying n and a fixed number of classifiers $M \geq 2$, it holds that $\lim_{n\to\infty} \tilde{P}^*_{n,M} = lim_{n\to\infty n\to\infty} P^*_{n,M}$. In other words, $b_{n,M} \to 0$as $n \to \infty$ for fixed M.

Theorem 1 shows that when we have two sets of classifiers and the smaller set is a subset of the larger set, the larger set classifiers will yield a smaller true best classification error rate. Theorem 2 shows that the expected MEC error rate $E\left(\tilde{P}^*_{n,M}\right)$ always underestimates the true minimal error rate $P^*_{n,M}$ and the negative bias always strictly holds. This is consistent with the result in Figure 3. In Theorem 3, when the number of classifiers M is fixed, the bias diminishes to zero as the sample size n increases to infinity. This is also consistent with the 10-classifier result in Figure 3b where the bias diminishes to around zero when n is beyond 320. According to Figure 3d, we observe that although the true best classification error rate does not decrease after the QDA is applied, the MEC bias estimate continued to decrease as more classifiers are included. This theoretical result brings clear caution to use MEC without bias correction. In other words, if a researcher runs n = 20–30 samples of pilot study and examines M = 300 classifiers via conventional cross-validation to choose the best, the minimal error rate from the 300 classifiers will likely generate low (or almost zero) error rate, while the underlying true error rate may stay high. The researcher may be misled to expand the study to a larger cohort or a prospective clinical trial, and eventually find it difficult to validate the model and cannot translate into a clinically useful diagnostic tool.

2.4 Three existing bias correction methods

In the literature, several methods have been developed to correct the downward bias of the MEC error rate, and most have focused on correcting the bias of parameter estimation via cross-validation for a given machine learning method. Below, we introduce four bias correction methods that we will compare in this article (Bernau *et al.*, 2013; Tibshirani and Tibshirani, 2009; Varma and Simon, 2006). Bernau *et al.* (2013) assessed the condition with multiple machine learning methods, while the others focused on correcting the bias of parameter tuning via cross-validation (e.g. estimate K for KNN) for a given machine learning method. In practice, if one considers many machine learning models along with feature selection and parameter tuning, the number of classifiers (M) examined can easily reach several hundreds. All three methods considered here can be generalized to this situation.

Nested cross validation (nestedCV): Instead of using a single loop cross-validation to find the minimal error estimate for a particular classifier, nestedCV uses two CV loops (shown in Supplementary Fig. S1). The dataset is initially divided into training and testing sets. Then LOOCV is applied on the training set using all the classifiers, and the classifier with the smallest error rate is selected and used to build the model based on the training set and then evaluate the error rate on the testing set in the end. Therefore, the testing set is independent of the model selection stage, including the selection of MEC. Finally, the process is repeated until each sample acts as the testing set once; thus it is a double LOOCV with two CV loops. The computation therefore scales with the square of the sample size. Instead of LOOCV, it is possible to use 5-fold or 10-fold cross-validation to accelerate the computing when the sample size is large.

Weighted mean correction (WMC/WMCS): The method (Bernau *et al.*, 2013) is proposed to be a smooth analytical alternative to nestedCV, which is a weighted mean of the resampling error rates, obtained using the different machine learning models/parameter values. Instead of using cross-validation, it is based on repeated subsampling. Then it estimates the unconditional error rate as a weighted sum of the error rate of every classifier on all the subsamples. The weights are estimated with two variants, WMC and WMCS. Compared with nestedCV in the original paper, the method is more stable and has a much lower computational demand. We apply the R package CMA to implement this method, and the subsampling fraction is chosen to be 0.8 in this study.

Tibshirani's procedure (TT): TT applies the idea of estimating the bias and adding back to the minimal error rate estimate to correct for the bias in the setting of K-fold cross-validation. It estimates the true best classification error rate as $2\tilde{P}_{n,M} - \frac{1}{K}\sum_{k=1}^{K} \tilde{P}_{n,M,k}$ where $\tilde{P}_{n,M}$ is the biased MEC error rate when sample size is n with M classifiers and $\tilde{P}_{n,M,k}$ is the minimal error rate in the kth-fold among all classifiers (Tibshirani and Tibshirani, 2009). It does not require a significant amount of additional computation as in nestedCV and scales linearly with the number of cross-validation folds. Owing to the calculation of $\tilde{P}_{n,M,k}$, Tibshiranis' method is not suitable for LOOCV or when the size of the left-out test set is too small, and this estimate was shown to over-estimate the bias in some settings (Bernau *et al.*, 2011; http://epub.ub.uni-muenchen.de/12231/). The Tibshiranis' approach targets correcting the optimization bias of the conditional minimal error rate, which heavily depends on the single observed dataset, and the results are more variable. On the contrary, the IPL approach proposed below considers correcting the optimization of the unconditional minimal error rate by using a resampling technique to take into account the sampling variation and, therefore, the results are more reliable and stable.

2.5 The resampling-based IPL method

In this section, we propose a new resampling-based IPL method to correct the MEC error rate bias and estimate the true optimal classification error rate $P^*_{n,M}$. By constructing learning curves for each individual classifier from repeated resampling of the original dataset at different subsample sizes (Mukherjee *et al.*, 2003), we could estimate the error rate of each classifier by fitting a learning curve. Supplement Figure S1 shows the concept of IPL for learning curve fitting using 2D simulated data from Section 2.2. Five simulations in each of the various sample sizes (n) are performed, and the LOOCV error rates (P) from QDA are demonstrated. The trend of decreasing error rates with increasing sample sizes is clear. By fitting an IPL function ($P = a \cdot n^{-\alpha} + b$; a, b, $\alpha > 0$), the learning curve can be well estimated.

Our proposed method applies the IPL concept using repeated subsampling as follows. Consider sample sizes $1 \leq n_1 < n_2 < \cdots < n_L < n$. For a given machine learning method m, assume that the true error rate equals $P_{n_l,m}$ and these true error rates follow an IPL function: $P_{n_l,m} = a_m n_l^{-\alpha_m} + b_m$. Normally, we assume $a_m, b_m, \alpha_m > 0$, as theoretically larger sample size contains more information to produce a lower

prediction error rate. To estimate a_m, b_m and α_m, we first estimate the underlying $P_{n_l,m}$ from subsampling n_l samples from the whole data and repeat the procedure for B times. The resulting observed cross-validated error rate of each of the sub-sampled data is denoted by $\hat{P}^{sub}_{n_l,m,b}$ ($1 \le b \le B$), and the averaged error rate is $\hat{P}^{sub}_{n_l,m} = (\sum_{b=1}^{B} \hat{P}^{sub}_{n_l,m,b})/B$. The least squared error method is then used to estimate a_m, b_m and α_m.

$$\left(\hat{a}_m, \hat{b}_m, \hat{\alpha}_m\right) = arg\ min_{a_m,b_m,\alpha_m} \sum_{l=1}^{L} (\hat{P}^{sub}_{n_l,m} - a_m n_l^{-\alpha_m} - b_m)^2$$

$$s.t.\ a_m, b_m, \alpha_m \ge 0$$

The IPL has been found to fit well in simulation and many real datasets (Mukherjee *et al.*, 2003). It has the advantage to obtain an accurate estimate of $\hat{P}^{IPL}_{n,m} = \hat{a}_m n^{-\hat{\alpha}_m} + \hat{b}_m$ for $P_{n,m}$ for any sample size n.

The bias of MEC error rate can then be estimated by $b^{IPL}_n(B) = \tilde{P}_{n,M} - \hat{P}^{IPL}_n$, where $\tilde{P}_{n,M}$ denotes the MEC error rate for a fixed dataset and $\hat{P}^{IPL}_n = min_m\ \hat{P}^{IPL}_{n,m}$.

The IPL approach has two advantages. First, through subsampling and fitting by constructing learning curves, the IPL method borrows information from neighboring estimates at different sample sizes, which has the potential to reduce the random noise of the true best classification error rate estimate, and the estimator will be more stable and accurate. The second advantage for IPL is its potential to extrapolate the learning curves to estimate the true best classification error rate beyond the current sample size so that it can provide prediction on how the error rate will further decline if more samples are included in future studies. For example, from an existing observed data of $n = 40$ samples, IPL can estimate the expected accuracy at $n = 100$ or $n = 250$ samples and inform researchers whether it is worthwhile to extend the study to a larger cohort.

3 RESULTS

3.1 Bias correction of the simulated 2D example

We evaluated performance of different bias correction methods (TT, nestedCV, WMC/WMCS and IPL) on the 2D toy model. We calculated both bias-corrected estimates with $M = 2$ classifiers (DLDA and QDA) and all $M = 10$ classifiers and compared them with the true best classification error rate at sample size $n = 20$, 40, 80, 160, 640, 1280 with B = 100 simulated datasets. As shown in Figure 4, left, the nestedCV method generally overestimated the true best classification error rate and the overestimated bias was larger when using 10 classifiers compared with two classifiers. Considering each inner cross-validation model selection stage of nestedCV, if the true best classifier (QDA in this example) was not selected frequently (i.e. $\tilde{m}^{(MEC)}_b \ne m^*$ using the notation in Section 2.1), the final minimal error rate was estimated with another suboptimal mix of classifiers. In this respect, on average, without being able to select the true best classifier to estimate the error rate on the test dataset, nestedCV resulted in an overestimate of the true best classification error rate. The downward bias problem could get more severe when more classifiers are included. Figure 4 shows that nestedCV generally overcorrects the bias. TT also overcorrects the bias at $n = 40$ and 80 when $M = 2$ and $n = 80$ and 160 when $M = 10$. The overcorrection of TT is consistent with the results of the original paper (Tibshirani and Tibshirani, 2009) as well as a previous technical report (Bernau *et al.*, 2011). WMC/WMCS perform well at $M = 2$ but fluctuates when $M = 10$. IPL performs overall the best. It slightly

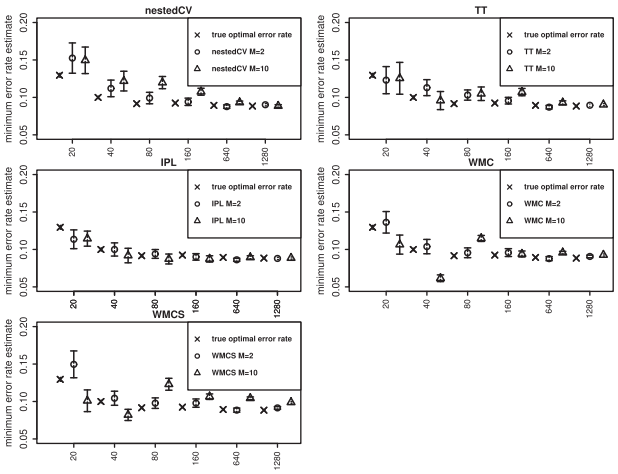

Fig. 4. Comparison of minimal error rate estimates from all four bias correction methods with $M = 2$ classifiers (QDA and DLDA), as well as with $M = 10$ classifiers

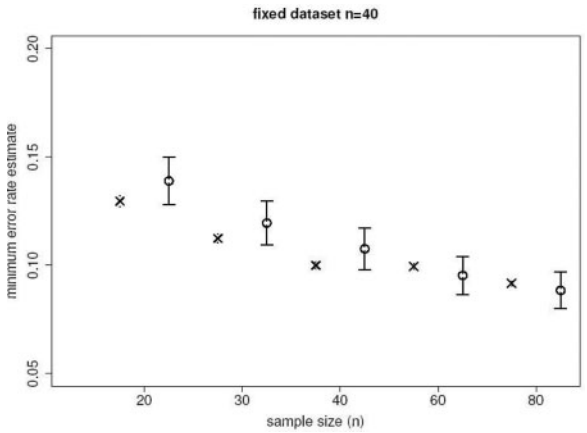

Fig. 5. Estimated minimum error rate (circle) using IPL extrapolation ($\hat{P}^{IPL}_n$) for a fixed dataset with $n = 40$. The bars reflect 95% coverage interval from 100 simulations. The true minimum error rate ($P^*_{n,M}$) is shown (cross) as a reference

underestimates the error rate at $n = 20$ but obtains accurate estimates for $n = 40$–1280.

To illustrate performance of IPL extrapolation for estimating prediction ability in a larger sample size, Figure 5 shows the estimated best error rates for sample sizes $n = 20$, 30, 40, 60 and 80 using IPL (i.e. $\hat{\boldsymbol{P}}^{\boldsymbol{IPL}}_{\boldsymbol{n}}$) from an observed dataset of $n = 40$. The true best error rates ($\boldsymbol{P}^*_{\boldsymbol{n,M}}$) are marked by a 'cross' for reference. The IPL extrapolations generally estimated the truth pretty well. The result shows that increasing sample size from $n = 40$ to $n = 80$ only slightly improved the prediction accuracy, and it is probably not worthwhile to collect an additional 40 samples.

3.2 Application on five GEO datasets

We applied the methods to five randomly selected GEO real datasets (see Supplementary file and Supplementary Table S1

Table 1. MEC error rate and corrected error rates by TT (5-fold and 10-fold cross-validation), nestedCV, WMC, WMCS and IPL

GDS	*n*	MEC	TT(5)	TT(10)	nestedCV	WMC	WMCS	IPL
1627	42	0.000	0.00	0.00	0.02	0.04	0.04	0.003
2190	61	0.323	0.45	0.44	0.39	0.45	0.42	0.40
2362	49	0.000	0.00	0.00	0.06	0.04	0.03	0.002
2415	59	0.320	0.47	0.45	0.49	0.44	0.42	0.38
2520	44	0.022	0.04	0.05	0.05	0.08	0.08	0.03

[a]TT targets on conditional error rate and is not directly comparable with the other methods.

for details on selection criteria). Because TT is not applicable for LOOCV, we applied it to 5-fold cross-validation and 10-fold cross-validation. For WMC/WMCS, B = 30 subsampling of 80% was applied. Ten machine learning methods were applied: KNN (K-nearest neighbor with k = 1, 3 and 5), DLDA, QDA, NNET, SVM with linear kernel, SVM with non-linear kernel (radial), RF and SCDA. Feature selection was done by performing a simple *t*-test and then selecting 2–30 top features by *P*-values in each cross-validation to construct the classifiers; therefore, a total of 290 classifiers were used.

Table 1 shows the best error rate after bias correction by all methods. The result shows that MEC has a significant downward bias compared with the estimates from those correction methods, especially for datasets GDS2190 and GDS2415. IPL generally gives a smaller estimate compared with the other methods, which is consistent with the simulation result that nestedCV and TT usually overcorrect the bias. WMC/WMCS also seems to give a large bias correction in these cases.

3.3 TCGA and METABRIC breast cancer data

Owing to lack of the underlying truth in the real data in Section 3.2, the results could not be conclusive. To circumvent this shortcoming, we applied three methods to two large breast cancer gene expression profiles, one from TCGA and the other from METABRIC. The TCGA breast cancer dataset was downloaded from TCGA Web site (http://tcga-data.nci.nih.gov/tcga) in October 2012. Level 3 RNA-Seq data were extracted from the Illumina HiSeq 2000 platform. We selected the TCGA breast cancer dataset that contained expression data of $n = 406$ tumor samples. We defined two classification problems: one is to classify between ER positive ($n = 391$) and ER negative ($n = 89$), and the second is to classify between early-stage (stages I and II, $n = 292$) and late-stage (stages III and IV, $n = 114$) tumors. The METABRIC gene expression and clinical data are retrieved from Synapse (https://www.synapse.org/#!Synapse:syn2133309). We obtained 1897 samples, consisting of 945 early-stage (stages I and II) and 952 late-stage (stages III and IV) tumors (Curtis *et al.*, 2012). We applied the same set of 290 classifiers in Section 3.2 to both datasets. Because the dataset contained a large sample size, we mimicked the simulation scheme described in Section 2.1 and randomly split the data into equal parts of ~40 samples. Under this setting, we pretended that we obtained B = 10 independent datasets of $n = 40$–41 from an unknown underlying distribution Δ_n in the TCGA dataset. Similarly, we have B = 47 and $n = 40$–41 for the METABRIC

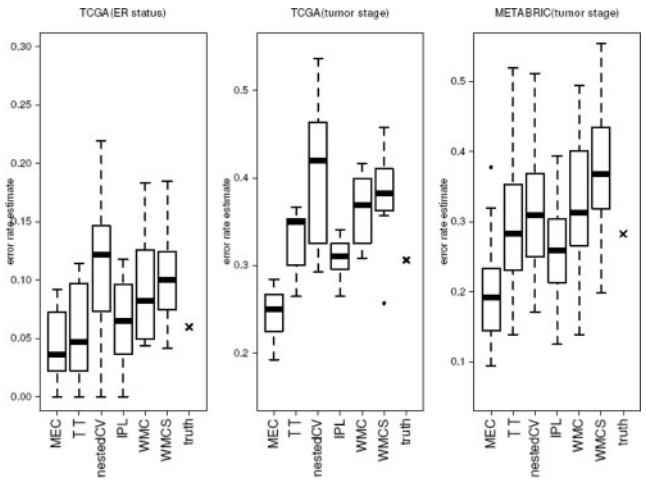

Fig. 6. MEC error rate estimates from all bias correction methods along with the true best optimal classification error rate and MEC error rates on the TCGA and METABRIC datasets. Left: classification between ER positive and ER negative in TCGA. Middle and right: classification between early and late tumor stage in TCGA and METABRIC

dataset. The random partition was also constrained such that sample sizes in two classes are as balanced as possible. By following the workflow of Figure 1, we generated the estimated best error rate $\hat{P}^*_{n,M}(B)$ (denoted as 'truth') and MEC error rates $\tilde{P}_{n,M,b}$ $(1 \le b \le B)$ (denoted as 'MEC') in Figure 6. Five-fold cross-validation was used for TT; leave-one-out was used for nestedCV and IPL. Thirty subsamplings were used for both IPL and WMC/WMCS. For WMC/WMCS, subsampling was at a proportion of 80%. The results showed that nestedCV produced the most variable error estimates spanning a large range, which is confirmed also by Bernau *et al.* (2013). WMC/WMCS yielded more stable estimates than nestedCV but created a large overcorrection in TCGA (tumor stage). TT gave relatively stable estimates but can either overcorrect or undercorrect the bias in terms of an unconditional error rate, although it theoretically estimates the conditional error rates (Bernau *et al.*, 2013). IPL generated the most stable and accurate error rates in all three cases. The averaged MEC bias is as large as 1.69% in TCGA ER status classification, 6.15% in the TCGA tumor stage classification and 9.11% in the METABRIC tumor stage classification. Without bias correction, an overly optimistic conclusion would have been drawn.

4 CONCLUSION AND DISCUSSION

With the advances of high-throughput genomic and proteomic techniques, data are generated in an unprecedentedly increasing pace. Machine learning methods have become a powerful tool in almost all biomedical research of complex diseases to seek new diagnostic or treatment selection tools. In most studies, small sample sizes are encountered ($n = 30$–60), and researchers are tempted to test many classifiers and select the best to report (i.e. applying the MEC). In this article, we illustrated the downward bias of MEC error rate when selecting from many machine learning models in biomedical classification problems. In the application of high-throughput genomic data, this problem is especially magnified because the addition of feature selection easily increases the number of classification models to several hundreds. We first demonstrated the problem using a 2D toy example where QDA is known to be the best classifier. The simulation results and asymptotic theoretical results both illustrated the need of bias correction for MEC, especially when sample size (n) is limited and the number of classifiers examined (M) is large. We discussed three existing methods (nestedCV, WMC/WMCS and TT) and developed a new IPL method from the concept of learning curve fitting. Application of all four methods to the 2D toy example, five selected GEO datasets and two large breast cancer datasets concluded that nestedCV and TT overestimated the error rate, whereas WMC/WMCS produces a fluctuating estimate around the true estimates. IPL provided a stable and accurate solution. The method has an additional advantage to extrapolate and predict the optimal error rate for larger sample sizes, a useful feature to help decide whether it is worthwhile to expand the study to recruit more samples. We note that nestedCV and WMC/WMCS target the error rate of the wrapper algorithm, which is slightly different from the MEC error rate we target in this article. This is consistent with the result that bias corrections by nestedCV and WMC/WMCS are less accurate and more fluctuating. Our proposed IPL method has the advantage of directly targeting the MEC error rate and completely avoid the wrapper algorithm issue.

Our article provides a careful framework and theoretical investigation of the problem, and our result shows that severe bias can be generated for MEC with small sample size (e.g. $n = 30$–60) and a large number of classifiers (e.g. $M = 300$). Without bias correction, one runs the risk of obtaining an overly optimistic error estimate of the classification model, excitedly expanding the investigation to larger independent cohorts and, eventually, failing to validate and translate into a useful clinical tool. The IPL method we proposed in this article not only generates more accurate bias correction but also provides extrapolation estimates to determine whether larger cohorts might warrant improved accuracy. In the era of pursuing translational research and personalized (or precision) medicine, rigorous evaluation and interpretation of the machine learning results are essential to evaluate the clinical potential of a research finding.

There are a few limitations or considerations for our study. First, the IPL method has the modeling assumption that learning curves of each classifier could be fitted well by IPL. Although there is no theoretical proof that this always holds, our simulation and real data showed good fit to the assumption. Second, the curve fitting of the IPL method relies on subsampling at smaller sample sizes. Therefore, if the original sample size is small, IPL will yield an unstable estimate. Third, the IPL methods are more costly compared with WMC/WMCS because of its need to subsample at different sample sizes. However, because all the classifiers are fitted independently, it is easy to parallelize the computation. Last, we sum up all different feature selections, machine learning methods and their associated parameter setting into M classifiers in the investigation. Theoretically different sources of classifiers have different correlated performance. Understanding their correlations may elucidate the contribution of bias from different sources and develop a better solution. In addition, we demonstrated the idea that one should include high-performing machine learning methods in the selection and avoid adding low-performing methods. In practice, one may determine high- and low-performance methods from empirical studies (e.g. comparison of performance in similar studies in large databases, such as GEO). How to systematically integrate the information to decide the set of classifiers for investigation is still an open question. All code and source files are available at http://tsenglab.biostat.pitt.edu/publication.htm to reproduce the results in the article. An R package 'MLbias' is also available.

ACKNOWLEDGEMENT

The authors would like to thank suggestions from the reviewers that have significantly improved this article.

Funding: (NIH R21MH094862).

Conflict of interest: none declared.

REFERENCES

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Bernau,C. *et al.* (2011) Correcting the optimally selected resampling-based error rate: a smooth analytical alternative to nested cross-validation. In: *Technical report*. Department of Statistics, University of Munich.

Bernau,C. *et al.* (2013) Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics*, **69**, 693–702.

Berrar,D. *et al.* (2006) Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*, **22**, 1245–1250.

Boulesteix,A.-L. and Strobl,C. (2009) Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med. Res. Methodol.*, **9**, 85.

Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

Dupuy,A. and Simon,R.M. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl Cancer Inst.*, **99**, 147–157.

Efron,B. (2009) Empirical Bayes estimates for large-scale prediction problems. *J. Am. Stat. Assoc.*, **104**, 1015–1028.

Fu,W.J. *et al.* (2005) Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, **21**, 1979–1986.

Mukherjee,S. *et al.* (2003) Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.*, **10**, 119–142.

Slawski,M. *et al.* (2008) CMA: a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, **9**, 439.

Tibshirani,R.J. and Tibshirani,R. (2009) A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.*, **3**, 822–829.

Varma,S. and Simon,R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91.

Wood,I.A. *et al.* (2007) Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, **23**, 1363–1370.

Yousefi,M.R. *et al.* (2010) Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, **26**, 68–76.