Supplementary materials for
Multiple-Rule Bias in the Comparison of Classification Rules

Mohammadmahdi R. Yousefi 1, Jianping Hua 2, Edward R. Dougherty 1,2,3

1Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843, USA.

2Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, 85004, USA.

3Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

February 17, 2011

1 Real data sets

A set of real data containing four microarray experiments, with sample size larger than 150, aggregated for this study. We tried to maintain the original labeling and also to follow the data preparing directions used in the papers reporting these data sets, but in several cases we re-labeled samples for reasons which would be given accordingly. Here we provide a brief description of these data sets:

1.1 Pediatric acute lymphoblastic leukemia [Yeoh et al., 2002], 248 sample points, 5077 features

Acute lymphoblastic leukemia (ALL) data set has been obtained from a study on pediatric acute lymphoblastic leukemia [Yeoh et al., 2002]. ALL is a complicated disease containing several subtypes. Data points were labeled into six subtypes: T-ALL (43 sample points), E2A-PBX1 (27 sample points), TEL-AML1 (79 sample points), BCR-ABL (15 sample points), MLL (20 sample points), and hyperdiploid with >50 chromosomes (64 sample points). The data have been collected using Affymetrix’s Human Genome HG_U95Av2 array (Santa Clara, CA) and are publicly available at http://www.stjude.org/data/ALL1. This microarray chip contains 12,000 probe sets (features) [Hua et al., 2009].
We have removed the features in which less than 1 percent of the sample points had a present call or more than 10 percent of the sample points had their values missing. This reduced the total feature size to 5077. The missing values were filled by averaging across all sample points. We labeled the data into two classes: one containing T-ALL, E2A-PBX1 and TEL-AML1 subtypes (149 sample points), the other containing BCR-ABL, MLL and hyperdiploid >50 subtypes (99 sample points). Hence, the total sample size is 248 [Hua et al., 2009].

1.2 Hepatocellular carcinoma [Chen et al., 2004], 157 sample points, 10,237 features

The original study was on hepatocellular carcinoma (HCC) which is the most common adult liver malignancy and ranks among the top five causes of cancer death in the world. The liver tissues (tumor and non-tumor) were obtained from surgical resections or transplants performed at Stanford University, CA, USA or Queen Mary Hospital, The University of Hong Kong, Hong Kong, China. Frozen liver samples were used for cDNA microarray study and reverse transcription-polymerase chain reaction (RT-PCR); and paraffin-embedded samples were used for IHC and in situ hybridization (ISH). Total RNA was extracted from tissues using Trizol Reagent (Invitrogen, Carlsbad, CA, USA). For microarray analysis, messenger RNA (mRNA) was isolated from total RNA using FastTract mRNA purification kit (Invitrogen) [Chen et al., 2004].

In our study, the array data were obtained from the Stanford Microarray Database (http://genome-www.stanford.edu/microarray) which had been labeled by 82 primary tumor tissues, and 75 non-tumor tissues. First, we found well-measured genes in each sample (the fluorescent intensity in each channel was greater than 1.5 times the local background). This set might contain flagged genes, also genes which had missing values. We further assumed that flagged genes in this set were also missing values. Therefore, we had three different kinds of missing values: a) actual missing values in; b) bad-measured genes; c) flagged genes. Then, we removed genes which had 25 percent or more missing values across all samples, and imputation process were applied to fill the missing values by the simple average of each gene in the reduced feature set. Doing this process resulted in reduction of the number of features from 24,168 to 10,237.

1.3 Drugs and toxicants response on rats [Natsoulis et al., 2005], 181 sample points, 8491 features

Drugs and toxicants response on rats data set has been obtained from a study characterizing the gene expression of different drugs and toxicants on live rats [Natsoulis et al., 2005]. Altogether 22 drugs and toxicants have been fed to male Sprague-Dawley rats for several durations and up to 12 tissues have been harvested. The data are publicly available at the NIH GEO, under accession number GSE2187. In this paper, authors claimed that there were totally 597 arrays corresponding to 199 triplicate. But the data set at the NIH GEO contains 587 arrays (198 original sample points, some have only duplicates). In 198 sample points, there are 17 points with no label, so they were removed from the data set.

The treatments correspond to four categories: fibrates (36 sample points), statins (31 sample points), azoles (53 sample points) and toxicants (61 sample points). We labeled the data into two classes: one containing 61 toxicants samples, the other containing 120 sample points from three the
remaining categories. Hence, the total sample size is 181 [Hua et al., 2009]. The data have been collected on cRNA microarray chips containing 8565 probes (features). We excluded the features in which more than 10 percent of the sample points had their values missing. This reduced the total feature size to 8491. The missing values were filled by averaging across all sample points.

1.4 Multiple myeloma [Zhan et al., 2006], 234 sample points, 54,613 features

Multiple myeloma (MM) data set has been obtained from a study on MM and monoclonal gammopathy of undetermined significance (MGUS) [Zhan et al., 2006] and [Zhan et al., 2007]. The data were collected using Affymetrix’s Human Genome U133 Plus 2.0 Array (Santa Clara, CA) and are publicly available at the NIH Gene Expression Omnibus (GEO), under accession numbers GSE5900 and GSE2658. This microarray chip contains 54,613 probe sets (features) to cover all kinds of gene transcripts and variants.

The original data set is consisted of four subtypes: MM (559 sample points), MGUS (44 sample points), smoldering MM (SMM, 12 sample points), and healthy donors with normal plasma-cell (NPC, 22 sample points). We labeled sample points based on MM sample points vs. the other containing MGUS, SMM, and NPC sample points (78 sample points). Since the number of MM patients is overwhelming and can have significant effects on the efficiency of feature selection and the accuracy of error estimation, we have randomly selected 156 sample points from among the 559 MM sample points and pair them with the 78 sample points of MGUS/SMM/NPC. Hence, the total sample size is 234 [Hua et al., 2009].
2 Results
2.1 Synthetic data
2.1.1 Joint distributions
Figure s1: The joint distributions between $\varepsilon_{\text{min}}$ and $\varepsilon_{\text{true}}$ with respect to the collection size $m$, for all classifier rule models, 60 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s2: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for all classifier rule models, 60 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s3: The joint distributions between $\varepsilon_{\text{est}}^\text{min}$ and $\varepsilon_{\text{ture}}^\text{min}$ with respect to the collection size $m$, for all classifier rule models, 120 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s4: The joint distributions between $\varepsilon_{est}$ and $\varepsilon_{true}$ with respect to the collection size $m$, for all classifier rule models, 120 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s5: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, 60 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s6: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, 60 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s7: The joint distributions between $\varepsilon_{\text{est}}^\text{min}$ and $\varepsilon_{\text{true}}^\text{min}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, 120 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s8: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, 120 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s9: The joint distributions between $\varepsilon_{est}^{\min}$ and $\varepsilon_{true}^{\min}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, 60 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s10: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, 60 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s11: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, 120 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s12: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, 120 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s13: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, 60 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s14: The joint distributions between $\varepsilon_{\text{est}}$ and $\varepsilon_{\text{true}}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, 60 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s15: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, 120 samples and equal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s16: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, 120 samples and unequal variances. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
2.1.2 Correlation and regression
Figure s17: The correlation and regression between $\varepsilon_{\text{min}}^{\text{est}}$ and $\varepsilon_{\text{min}}^{\text{true}}$ on the synthetic data for all 36 classification models, with respect to the collection size $m$. 
Figure s18: The correlation and regression between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ on the synthetic data for single 5-fold cross-validation error estimation, with respect to the collection size $m$. 
Figure s19: The correlation and regression between $\varepsilon_{\min}^{\text{est}}$ and $\varepsilon_{\min}^{\text{true}}$ on the synthetic data for single 10-fold cross-validation error estimation, with respect to the collection size $m$. 
Figure s20: The correlation and regression between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ on the synthetic data for single LOO cross-validation error estimation, with respect to the collection size $m$. 
2.1.3 Bias, variance, RMS and gain
Figure s21: The expected bias, $\text{Bias}_{\text{Av}}$; the expected variance, $\text{Var}_{\text{Av}}$; the expected RMS, $\text{RMS}_{\text{Av}}$; and the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}}|\Phi_m]$; resulted from the distributions of $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ on the synthetic data for all 36 classification models, with respect to the collection size $m$. 
Figure s22: The expected bias, $\text{Bias}_{AV}$; the expected variance, $\text{Var}_{AV}$; the expected RMS, $\text{RMS}_{AV}$; and the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}} | \Phi_m]$; resulted from the distributions of $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ on the synthetic data for single 5-fold cross-validation error estimation, with respect to the collection size $m$. 
Figure s23: The expected bias, $\text{Bias}_{Av}$; the expected variance, $\text{Var}_{Av}$; the expected RMS, $\text{RMS}_{Av}$; and the expected comparative performance bias, $E_{\epsilon_{m}}[C_{\text{bias}} | \Phi_{m}]$: resulted from the distributions of $\epsilon_{\text{est}}^{\text{min}}$ and $\epsilon_{\text{true}}^{\text{min}}$ on the synthetic data for single 10-fold cross-validation error estimation, with respect to the collection size $m$. 
**Figure s24:** The expected bias, $\text{Bias}_{AV}$; the expected variance, $\text{Var}_{AV}$; the expected RMS, $\text{RMS}_{AV}$; and the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\text{est}}$ and $\varepsilon_{\text{ture}}$ on the synthetic data for single LOO cross-validation error estimation, with respect to the collection size $m$. 
2.2 Real data
2.2.1 Joint distributions
Figure s25: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for all classifier rule models, [Yeoh et al., 2002] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s26: The joint distributions between $\varepsilon_{\text{min}}^{\text{est}}$ and $\varepsilon_{\text{min}}^{\text{true}}$ with respect to the collection size $m$, for all classifier rule models, [Chen et al., 2004] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s27: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for all classifier rule models. [Natsoulis et al., 2005] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s28: The joint distributions between $\varepsilon_{\min}^{\text{est}}$ and $\varepsilon_{\min}^{\text{true}}$ with respect to the collection size $m$, for all classifier rule models. [Zhan et al., 2006] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s29: The joint distributions between $\varepsilon_{\text{min}}^{\text{est}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, [Yeoh et al., 2002] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s30: The joint distributions between $\epsilon_{\text{est}}^{\text{min}}$ and $\epsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator. [Chen et al., 2004] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s31: The joint distributions between $\varepsilon_{\text{min est}}$ and $\varepsilon_{\text{min true}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, [Natsoulis et al., 2005] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
**Figure s32:** The joint distributions between $\varepsilon_{\text{min est}}$ and $\varepsilon_{\text{true min}}$ with respect to the collection size $m$, for single 5-fold cross-validation error estimator, [Zhan et al., 2006] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s33: The joint distributions between $\varepsilon_{\text{est}}^\text{min}$ and $\varepsilon_{\text{true}}^\text{min}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, [Yeoh et al., 2002] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s34: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, [Chen et al., 2004] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s35: The joint distributions between $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{\min}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator. [Natsoulis et al., 2005] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s36: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single 10-fold cross-validation error estimator, [Zhan et al., 2006] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s37: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single LOO cross-validation error estimator. [Yeoh et al., 2002] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s38: The joint distributions between $\xi_{\text{est}}^\text{min}$ and $\xi_{\text{true}}^\text{min}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, [Chen et al., 2004] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s39: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ with respect to the collection size $m$, for single LOO cross-validation error estimator, [Natsoulis et al., 2005] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
Figure s40: The joint distributions between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{i_{\text{min}}}$ with respect to the collection size $m$, for single LOO cross-validation error estimator. [Zhan et al., 2006] data set. The white line shows the regression line and the circle indicates the sample mean of the joint distribution.
2.2.2 Correlation and regression
Figure s41: The correlation and regression between $\varepsilon_{\text{est}}$ and $\varepsilon_{\text{ture}}$ on the real data for all 36 classification models, with respect to the collection size $m$. 
Figure s42: The correlation and regression between $\epsilon_{\text{est}}^{\min}$ and $\epsilon_{\text{true}}^{\min}$ on the real data for single 5-fold cross-validation error estimation, with respect to the collection size $m$. 
Figure S43: The correlation and regression between $\varepsilon_{\text{min}}^\text{est}$ and $\varepsilon_{\text{true}}^\text{true}$ on the real data for single 10-fold cross-validation error estimation, with respect to the collection size $m$. 

Chen
Natsoulis
Yeoh
Zhan
Average
Figure s44: The correlation and regression between $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ on the real data for single LOO cross-validation error estimation, with respect to the collection size $m$. 
2.2.3 Bias, variance, RMS and gain
Figure s45: The expected bias, $\text{Bias}_{\text{Av}}$, the expected variance, $\text{Var}_{\text{Av}}$, the expected RMS, $\text{RMS}_{\text{Av}}$, and the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}} | \Phi_m]$: resulted from the distributions of $\epsilon_{\text{est}}_{\text{min}}$ and $\epsilon_{\text{true}}_{\text{min}}$ on the real data for all 36 classification models, with respect to the collection size $m$. 
Figure s46: The expected bias, $\text{Bias}_{Av}$; the expected variance, $\text{Var}_{Av}$; the expected RMS, $\text{RMS}_{Av}$; and the expected comparative performance bias, $\mathbb{E}_{\Phi_m}[C_{\text{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\text{est}}^{\text{min}}$ and $\varepsilon_{\text{true}}^{\text{min}}$ on the real data for single 5-fold cross-validation error estimation, with respect to the collection size $m$. 

---

38
Figure S47: The expected bias, $\text{Bias}_{\text{Av}}$; the expected variance, $\text{Var}_{\text{Av}}$; the expected RMS, $\text{RMS}_{\text{Av}}$; and the expected comparative performance bias, $E_{\Phi m}[C_{\text{bias}}|\Phi m]$, resulted from the distributions of $\varepsilon_{\text{true}}$ and $\varepsilon_{\text{est}}$ on the real data for single 10-fold cross-validation error estimation, with respect to the collection size $m$. 

Chen Natsoulis Yeoh Zhan
Average

59
Figure S48: The expected bias, $\text{Bias}_{\text{Av}}$; the expected variance, $\text{Var}_{\text{Av}}$; the expected RMS, $\text{RMS}_{\text{Av}}$; and the expected comparative performance bias, $E_{\text{bias}, \Phi_m}$, resulted from the distributions of $\epsilon^{\text{est}}_{\text{true}}$ and $\epsilon^{\text{true}}_{\text{true}}$ on the real data for single LOO cross-validation error estimation, with respect to the collection size $m$. 
References


