Optimally splitting cases for training and testing high dimensional classifiers

Kevin K. Dobbin, Ph.D.
Dept. of Epi. and Biostat.
College of Public Health
University of Georgia
Athens, GA

Richard M. Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis
National Cancer Institute
Rockville, MD

July 6, 2010

Abstract

We consider the problem of designing a study to develop a predictive classifier from high dimensional data. A common study design is to split the sample into a training set and an independent test set, where the former is used to develop the classifier and the latter to evaluate its performance. In this paper we address the question of what proportion of the samples should be devoted to the training set. How does this proportion impact the mean squared error (MSE) of the prediction accuracy estimate? We develop a non-parametric algorithm for determining an optimal splitting proportion that can be applied with a specific dataset. We also perform a broad simulation study for the purpose of better understanding the factors that determine the best split proportions and to develop guidelines for splitting strategies that are robust and close to optimal under a wide variety of conditions.

These methods are based on a decomposition of the MSE into three intuitive component parts. By applying these approaches to a number of synthetic and real microarray datasets we show that the optimal proportion depends on the overall number of samples available and the degree of differential expression between the classes. Over a wide range of settings, it was found that $\frac{2}{3}$-to-$\frac{1}{3}$ training-to-test allocation performs nearly as well as the optimal split, and is more robust than $\frac{1}{2}$-to-$\frac{1}{2}$ allocation. A resampling approach that can be applied to any dataset, using any predictor development method, to determine the best split is presented.

1 Introduction

The split sample approach is a widely used study design in high dimensional settings. This design divides the collection into a training set and a test set as a means of estimating classification accuracy. A classifier is developed on the

training set and applied to each sample in the test set. In practice, statistical prediction models have often been developed without separating the data used for model development from the data used for estimation of prediction accuracy (Dupuy and Simon, 2007). When the number of candidate predictors (p) is larger than the number of cases as in microarray data, such separation is essential to avoid large bias in estimation of prediction accuracy (Simon et al., 2003). This paper addresses the question of how to optimally split a sample into a training set and a test set for a high dimensional gene expression study, that is, how many samples to allocate to each group.

Two approaches to evaluating splits of the data are examined. The first approach is based on simulations designed to understand qualitatively the relationships among dataset characteristics and optimal split proportions. We use these results to develop robust general guidelines for allocation of the data to training and test sets. Our second approach involves development of a nonparametric method that does not rely on distributional assumptions and can be applied directly to any existing dataset without stipulating any parameter values. The nonparametric method can be used with any predictor development method (e.g., nearest neighbor, support vector machine).

This paper addresses the situation in which the accuracy of a predictor will be assessed by its performance on a separate test set. An alternative approach is to apply resampling-based methods to the whole dataset. Because re-sampling strategies have been commonly mis-used, often resulting in highly biased estimates of prediction accuracy (Simon et al., 2003; Ambroise and McLachlan, 2002), many journals and reviewers mis-trust cross-validation and require validation on a sample not used for model development. Another advantage of the split sample method, particularly in large collaborative studies in which multiple groups will be developing predictors, is that the test set can be kept under "lock and key" by a honest broker (Shedden et al., 2008).

The question addressed in this paper has not to our knowledge been addressed before. Sample splitting has been addressed in other contexts, such as comparing different k-fold cross validations (Molinaro et al., 2005) or developing hold out estimation theory (Devroye et al., 1996) and bounds on Bayes error (Fukunaga, 1990). Mukherjee et al. (2003), Fu et al. (2005). Dobbin and Simon (2007) developed methods for planning the size of a training set, but these methods do not address the allocation of cases in an existing dataset to training and test portions. Since many gene expression based classifiers are developed retrospectively, there is often little control of the sample size.

In Section 2 we describe the parametric modeling approach used to generate general recommendations and the nonparametric approach that can be applied to specific datasets. In Section 3, we present the results of application of these methods to synthetic and real world datasets. In Section 4, conclusions and recommendations for dividing a sample into a training set and test set are discussed.

2 Methods

The classifier taken forward from a split-sample study is often the one developed on the full dataset. This full-dataset classifier comes from combining the training and test sets together. The full-dataset classifier has an unknown accuracy which is estimated by applying the classifier derived on the training set to the test set. The optimal split will then be the one that minimizes the mean squared error (MSE) with respect to this full-dataset classifier. The MSE naturally penalizes for bias (from using a training set smaller than n) and variance.

In the supplemental material, it is shown that under mild assumptions this MSE is proportional to

$$MSE \propto A + V + B. \tag{1}$$

Here we have symbols A, V and B to depict the decomposition, and these are used throughout the discussion below. Here is a description of each term in Equation (1). Diagram 1 shows the breakdown visually.

A=Accuracy Variance Term: The first term in Equation (1) reflects the variance in the true accuracy of a classifier developed on a training set $\mathcal{T}$ selected from the full dataset $\mathcal{S}$. Not all training sets $\mathcal{T} \subset \mathcal{S}$ will result in predictors with exactly the same accuracy. The variation in actual (true) accuracy among all these different predictors is the A term.

V=Binomial Variance Term: The second term in Equation (1) is the variance in the estimated accuracy that results from applying the classifier to the test set. This is a binomial variance because the classifier developed for a specific training set has some fixed true accuracy (success probability), and there are $n - t$ independent samples represented in the test set.

B=Squared Bias Term: The third term in Equation (1) is the squared bias that results from using a classifier that was developed on t training samples to estimate the accuracy of a classifier which is developed on n samples.

2.1 Model-based simulations for high dimensional expression profiles

With each sample is associated a p-dimensional vector of log gene expression measurements, say x, which is assumed to follow the multivariate normal distribution with mean vector μ_1 for class 1 and μ_2 for class 2 and common covariance matrix Σ. Of the p genes, m are assumed differentially expressed with difference in mean expression levels between classes of 2δ and the remaining p-m genes are not differentially expressed. Extensive simulations under a variety of conditions indicated that the components of MSE depended on the separation of the classes with regard to gene expression and this is determined by the number of diffeentially expressed genes, the degree of differential expression and the

correlation among the differentially expressed genes. In general none of these quantities are known before analyzing the data. However, we have attempted to utilize extensive simulation results to understand the relationship between sample size, class separation and splitting effectiveness in order to provide robust general recommendations.

Our simulations use the compound covariate predictor (Radmacher et al., 2002), with gene selection performed using the optimal selection cutpoint algorithm described in Dobbin and Simon (2007). We adjusted the method in Dobbin and Simon (2007) for predicting an optimal significance level for gene selection to avoid assuming that the prevalence of the classes is known.

The estimates of MSE as a function of splitting proportion are estimated for each simulated dataset in the following way:

1. Given $2\delta/\sigma$ standardized fold change, m informative genes, dimension p, n samples available, and a covariance matrix Σ, generate a dataset S from the probability model. Randomly select R training sets of size t. A grid of t values are evaluated for each total sample size n.

2. For each t above, calculate the optimal significance level cutoff α to use for gene selection (Dobbin and Simon, 2007).

3. Using the optimal α levels to select genes from pooled variance t-tests, develop compound covariate predictors (CCP) (Radmacher et al., 2002) for each training set.

4. For each classifier developed on a training set of size t, apply the classifier to the corresponding test set of size n-t and estimate the classification accuracy. Average estimates over the R replicates to obtain the mean predicted accuracy estimate.

5. Develop a CCP classifier on the full dataset S of n cases. Using the parameters used to generate S, the true accuracy of the classifier developed on the full dataset was computed from theory (see supplemental materials).

2.2 A method for determining the optimal sample split for a particular dataset, which utilizes a nonparametric data re-sampling approach

The nonparametric bootstrap method of estimating standard error (Efron and Tibshirani, 1993) was used to estimate the variance of the performance of a predictor developed on a training set of size t and applied to a test set of size $n - t$. In our previous notation, this was $A + V$. Splitting was performed prior to resampling in order to avoid overlap between the training and test sets.

In order to estimate the squared bias term B we considered adopting learning curve methods (Duda et al., Section 9.6.7, 2001), as used previously in Mukherjee et al. (2003). Briefly, the 2003 paper uses a parametric nonlinear least squares regression approach that fits a learning curve model to datapoints

of the plot with training set size t on the x-axis and the estimated error rate on the y-axis. Fitting a learning curve of the form $e = a + b/t^{\alpha}$ where e is the expected error and t is the training set size (and $\alpha > 0$), provides an estimate of the asymptotic error rate (a), i.e., when $t = \infty$. However, we found the parametric learning curve model for the data often did not fit our simulated or real data adequately. Also, estimation of the squared bias term B does not require estimation of the asymptotic error (a), but only the mean error rate for limited training sizes $t \leq n$. So instead, we use a nonparametric smoothing spline to fit the plot with the training sample size t on the x-axis and the average error rate on the y-axis. When the learning curve raw data were not monotone (usually because the error rate had stopped decreasing significantly relative to the noise level present), then we used isotonic regression to force monotonicity of the fitted curve.

The squared bias term is estimated as follows:

1. For fixed n, and for $t = 10, 20, ..., n - 10$, randomly divide the dataset into a training set and a test set $1,000$ times.

2. For each t, develop a classifier on each of the $1,000$ training sets and apply the classifier to the corresponding test set. For each t, calculate the mean error rate w of these $1,000$ classifiers.

3. Fit a smoothing spline or isotonic regression of w on t using spline case weights $1/t$. Adjust degrees of freedom visually based on the smoothing spline plot.

4. For $t = 10, 20, ..., n$, calculate $\hat{w}(t)$, the fit-value from the spline or isotonic regression of the error rate on t.

5. Estimate the squared bias using $[\hat{w}(n) - \hat{w}(t)]^2$.

2.3 Software and application to microarray datasets

Computations were carried out in C++ using a Borland 5 compiler and Optivec 5.0 vector and matrix libraries, and R version 2.6.1 (including R "stats" package for smooth.spline and isoreg functions). Gene expression data were obtained from the BRB ArrayTools Data Archive for Human Cancer Gene expression (url: http://linus.nci.nih.gov/BRB-ArrayTools.html), except for (Golub et al., 1999) data which was retrieved from the Broad Institute website (url: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). Data were normalized using MAS 5.0 and median centering each array using either R or BRB ArrayTools (developed by Dr. Richard Simon).

3 Results

We applied the parametric method to high dimensional multivariate normal datasets, while varying the parameter settings and the class prevalences. Results

are shown in Tables 1 and S1 (supplement). We considered total samples of size $n = 200$, $n = 100$ and $n = 50$. For example, when $m = 50$ genes are informative and $n = 200$, then the optimal number of samples for the training set (reading across the first row of Table 1) is 170, 70 or more, 30 or more, and 20 or more for effect sizes of 0.5, 1.0, 1.5 and 2.0, respectively. The "or more" in the last three training set sizes indicate that training set sizes anywhere from the specified size up to 190 result in practically equivalent mean squared error.

Several features are apparent in Table 1: (i) when the achievable accuracy is not much greater than 50%, the optimal split allocates the vast majority of samples to the test set. In this circumstance, no good classifier is possible so additional samples allocated to the training set are wasted and detract from lowering the variance of estimation in the test set; (ii) when the gene expression profiles of the two classes are widely separated, e.g., with a large number of differentially expressed genes and large effect sizes, small training sets are adequate to develop highly effective classifiers. The MSE is flat in this circumstance because the binomial variance is small when the accuracy is close to 1 and hence neither large training set nor large test sets are needed; (iii) when the gene expression profiles are moderately separated for the two classes and the sample size is 100 cases or less, the optimal split has most of the cases allocated to the training set.

Table S1 in the Supplement shows the results when the prevalence is unbalanced, namely, 2/3 from one class and 1/3 from the other class. The results for this imbalanced prevalence setting are very similar to the equal prevalence setting. This suggests that the same general optimal splits apply for a range of class prevalence (33% to 67%).

The relative sizes of the three terms contributing to the mean squared error of Equation (1) for the scenarios of Table 1 and S1 are shown in the Supplementary material. An example is shown in Figure 2. Generally, the A term tends to be relatively small across the range of sample sizes.

The squared bias term B tends to be relatively large for small sample sizes and to dominate the other terms. When development of a good classifier is possible, the actual accuracy of classifiers developed on the training set may initially increase rapidly as the training set size increases. As the sample size increases, the bias term B decreases until no longer dominating. This is because the accuracy of the classifier improves as the size of the training set increases and approaches the maximum accuracy possible for the problem at hand. The rate of decrease of the squared bias term B will depend somewhat on the type of classifier employed and on the separation of the classes. When the classes are not different with regard to gene expression, learning is not possible and B will equal zero for all training set sizes.

The binomial variance term V is generally relatively small unless the test set becomes very small at which point it often dominates. The only exceptions to this general rule are in cases where the prediction accuracy nears 1 for $t < n$, in which case this V term remains near zero even as the test set size becomes small.

Figure 3 is a comparison of the two most common rules of thumb for splitting

a sample into a training set and a test set. The figure compares 50% allotment to the training set versus 67% allotment to the training set for the equal prevalence case. Each scenario represented in Table 1 is also present in Figure 3. The x-axis is the average accuracy (%) for classifiers developed from the full dataset of n samples. The y-axis is the excess error from using a non-optimal split. The discussion is organized around the full dataset accuracy:

- When the full dataset size accuracy is very close to 1, both the 50% allotment and the 67% allotment to the training set result in similar excess error.
- When the full dataset size accuracy is moderate, say between 60% and 99%, then in several cases, assigning 67% to the training set results in noticeably lower excess error, while in other cases the two allotment schemes are roughly equivalent.
- Finally, and not surprisingly, when the full dataset size accuracy is below 60% (shaded area on graph), then allotment of 50% to the training set is preferable.

In sum, this graph shows that allotment of 2/3rds to the training set is more robust than allotment of 1/2 to the training set.

The nonparametric method was applied to simulated datasets and the MSE estimates compared to the parametric approach. Agreement between the two was very good (see Supplement Section 4 for figure).

Table 2 shows the results of the application of the nonparametric method to several real-world datasets.

Note that the rightmost two columns show the excess error when 1/2 and when 2/3rds are allotted to the training set. For the Rosenwald et al. (2002) dataset of diffuse large B-cell lymphoma, we estimated the optimal split for distinguishing between germinal-center B-cell-like lymphoma from all other types. For this dataset of $n = 240$ patient samples, the optimal split was $150 : 90$, with about two-thirds of the samples devoted to the training set. The excess error (root mean square error difference, RMSD) from the 2/3rds to training set rule of thumb is 0.001; as a comparison, the RMSD for a simple binomial random variable (with p=0.96) between a sample size of 236 and 240 is also 0.001. Hence, the excess error at $t = 2n/3$ is very small.

For the Boer et al. (2001) dataset, the optimal split was 80 for the training set and 72 for the test set, so that 53% were used to train the classifier to distinguish normal kidney from renal cell carcinoma. The dramatic difference in gene expression between cancer and normal tissues meant that a smaller training set size was needed to develop a highly accurate classifier (see Supplement for figure). As a result, the 1/2 to training set rule of thumb is a little better than the 2/3rds to training split. That being said, the excess error when 2/3rds ares used for training is only 0.004. For comparison, the RMSD of 0.004 is similar to the RMSD resulting from increasing the sample size from 142 to 152 in simple binomial sampling (when $p = 0.98$).

For the Golub et al. (1999) dataset, the optimal split was 40 for the training set and 32 for the test set, or 56% for the training to distinguish acute lymphoblastic leukemia from acute myologenous leukemia. This is another example of two classes with dramatically different expression profiles. Like the Rosenwald dataset, the 2/3rds to training set rule resulted in smaller excesss error than the 1/2 rule.

To distinguish oligodendroglioma from glioblastoma in the the Sun et al. (2006) dataset required 40 for the training set and 91 for the test set, or 31% for the training set. This optimal training sample size was somewhat smaller than expected. This appeared to be due to the accuracy leveling off after $t = 40$ training samples, while the variance terms increased monotonely for $t > 40$. The multidimensional scaling plot for these data showed a pronounced separation into two groups of cases – but these groups only partly corresponded to the class labels (see MDS plots in Supplement Section 5). The two groups were found easily with n=40 samples, but the corresponding error rate was relatively high because of the imperfect correlation between the class labels and the two clusters in the plots. One is left to speculate whether this pattern was the result of real underlying biology, or artifacts such as batch effects or sample labeling errors. An advantage of our method is that the plotting steps make it easier to identify such anomalies in datasets that might otherwise easily be missed. In this case it did appear that 40 samples in the training set was adequate to achieve accuracy near the best possible with the full n=130 samples. Although this last dataset to some extent violates our rule-of-thumb of 2/3rds to training set, it is also obviously an unusual dataset and perhaps argues for applying our resampling method when possible to identify such anomalies early.

The supplement provides figures related to the fitting on the real datasets. We found that for the application to the real-world microarray datasets it was critical to perform at least 1,000 bootstrap re-samplings and 1,000 sample splits in order to obtain adequately de-noised MSE curves over the range of sample sizes.

4 Discussion

We have presented two methods for determining the optimal split of a set of samples into a training set and a test set in the context of developing a gene expression based classifier. These methods were applied to a range of synthetic and real-world microarray datasets to evaluate the optimal allocation scheme. We discovered that the optimal proportion of cases for the training set tended to be in the range of 40% to 80% for the wide range of conditions studied. In some cases, the MSE function was flat over a wide range of training allocation proportions, indicating the near-optimal MSE performance was easy to obtain. In other cases, the MSE function was less flat, indicating clearer optimal selection. In general, smaller total sample sizes led to a larger proportions devoted to the training set being optimal. Intuitively this is because for a given degree of class separation, developing an effective classifier requires a minimal number

of cases for training and that number is a greater proportion of a dataset with fewer total cases.

The number of cases needed for effective training depends on the"signal strength" or the extent of separation of the classes with regard to gene expression. "Easy" classification problems contain individual genes with large effects or multiple independent genes with moderately large effects. For such problems the potential classification accuracy is high (low Bayes error). The number of training cases required for near optimal classification for such datasets is smaller and hence smaller proportions devoted to the training set could be near optimal (for $n = 100 - 200$).

We found that when the average accuracy of a classifier developed on the full dataset (size n) was $> 60\%$, then a $\frac{2}{3}$-to-$\frac{1}{3}$ training-to-test set split resulted near optimal MSE in all settings considered. Based on careful analysis and interpretation of the extensive simulations in the Appendix, we think a good rule of thumb is to assign $2/3rds$ to the training set and $1/3rd$ to the test set. A separate Section in the Supplemental material describes the reasoning behind this recommendation.

This paper focused on the objective of obtaining a classifier with high accuracy. In some clinical contexts other objectives may be more appropriate, such as estimation of the positive and negative predictive values, or area under the ROC curve. If the prevalence is approximately equal for each class, however, then a high overall accuracy will be highly correlated with high negative and positive predictive values and AUC, so the guidelines here are likely to carry over to these other metrics.

The population prevalence from each class can be an important factor in classifier development. In this paper we looked at equal prevalence from each class, and at the case of 2/3 to 1/3 prevalence split in our simulations. The real datasets had prevalences within this range as well. In cases where there is significant prevalence imbalance between the classes (e.g., 90% versus 10%) there will often be a number of issues outside the scope of this paper. To modify our method for that context, one would need to address whether oversampling from the under-represented class is needed, and whether the cost of misallocation should differ by class.

We looked at a range of sample sizes from $n = 50$ to $n = 200$. In practice, sample sizes of $n = 50$ are probably too small to divide into a training set and a test set, and a better design uses resampling methods to estimate the classification accuracy instead. This study supports the general advice to use resampling methods in small sample settings because in these settings our method indicates that the MSE is generally minimized when most of the samples are devoted to the training set, with a typical allocation of 40 to training and only 10 samples for the test set. This will usually be inadequate except in very preliminary exploratory studies. For example, even if the observed classification accuracy in the test set is $10/10 = 100\%$, the 95% confidence interval for classification accuracy is $69\% - 100\%$.

The data based resampling method presented in this paper can be used with any predictor development method by making minor modifications to the

algorithm outlined in Section 3.4.

5 Acknowledgements

Kevin K. Dobbin's work was partially supported by the Distinguished Cancer Clinicians and Scientists Program of the Georgia Cancer Coalition.

6 Conflict of Interests

The authors declare to conflict of interest.

References

[1] Ambroise C., McLachlan G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences USA***99**, 6562-6.

[2] Boer J.M., Huber W.K., Sultman H., Wilmer F., vonHeydebreck A., Haas, S., Korn B., Gunawan B., Vente A., Fuzesi L., Vingron M., Poustka A. (2001) Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Research***11**, 1861-70.

[3] Devroye L., Gyorfi L., and Lugosi G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[4] Dobbin K.K. and Simon R. (2007) Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8**, 101-7.

[5] Duda R.O., Hart P.E., and Stork D.G. (2001) *Pattern Classification, Second Edition.* Wiley, New York.

[6] Dupuy A. and Simon R.M. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute USA* **99**, 147-57.

[7] Efron B. and Tibshirani R. (1993) *An Introduction to the Bootstrap.* Chapman and Hall, Boca Raton.

[8] Fu W.J., Dougherty E.R., Mallick B. and Carroll R.J. (2005) How many samples are needed to build a classifier: a general sequential approach. *Bioinformatics* **21**, 63-70.

[9] Fukunaga K. (1990) *Introduction to Statistical Pattern Recognition, Second Edition.* Academic Press, San Diego, CA.

[10] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-7.

[11] Molinaro A.M., Simon R., Pfeiffer R.M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301-7.

[12] Mukherjee S., Tamayo P., Rogers S., Rifkin R., Engle A., Campbell C., Golub T.R. and Mesirov J.P. (2003) Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10**, 119-42.

[13] Radmacher M.D., McShane L.M., Simon R. (2002) A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology* **9**, 505-11.

[14] Rosenwald A., Wright G., Chan W.C., Connors J.M., Campo E., Fisher R.I. Gascoyne R.D., Muller-Hermelink H.K., Smeland E.B., Giltname J.M., Hurt E.M., Zhao H., Averett L., Yang L., Wilson W.H., Jaffe E.S., Simon R., Klausner R.D., Powell J., Duffey P.L., Longo D.L., Greiner T.C., Weisenburger D.D., Sanger W.G., Dave B.J., Lynch J.C., Vose J., Armitage J.O., Montserrat E., López-Guillermo A., Grogan T.M., Miller T.P., LeBlanc M., Ott G., Kvaloy S., Delabie J., Holte H., Krajci P., Stokke T., Staudt L.M. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 1937-47.

[15] Shedden K., Taylor J.M., Enkemann S.A., Tsao M.S., Yeatman T.J., Gerald W.L., Eschrich S., Jurisica I., Giordano T.J., Misek D.E., Chang A.C., Zhu C.Q., Strumpf D., Hanash S., Shepherd F.A., Ding K., Seymour L., Naoki K., Pennell N., Weir B., Verhaak R., Ladd-Acosta C., Golub T., Gruidl M., Sharma A., Szoke J., Zakowski M., Rusch V., Kris M., Viale A., Motoi N., Travis W., Conley B., Seshan V.E., Meyerson M., Kuick R., Dobbin K.K., Lively T., Jacobson J.W., Beer D.G. (2008) Gene expression-based survival prediction in lung adenocarcinmoa: a multi-site, blinded validation study. *Nature Medicine* **14(8)**, 822-827.

[16] Simon R., Radmacher M.D., Dobbin K., McShane L.M. (2003) Pitfalls in the use of DNA micoarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute USA* **95**, 14-8.

[17] Sun L., Hui A.M., Su Q., Vortmeyer A., Kotliarov Y., Patorino S., Passanti A., Menon J., Walling J., Bailey R., Rosenblum M., Mikkelsen T., Fine H.A. (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **4**, 287-300.

Figure Legends

Figure 1: Diagram of mean squared error decomposition.

Figure 2: Example figure showing the relative contributions of the three sources of variation to the mean squared error. This is a scenario from one entry in Table 1. Plots for all other scenarios associated with Table 1 and Table S1 appear in the supplement.

Figure 3: Comparison of two common rules-of-thumb: 1/2 the samples to the training set and 2/3rds of the samples to the training set. X-axis is the average accuracy (%) for training sets of size n. "Excess error" on the y-axis is the difference between the root mean squared error (RMSE) and the optimal RMSE. Each point corresponds to a cell in Table 1. Gray shading indicates scenarios where mean accuracy for full dataset size is below 60%.

Tables

Table 1: Table of optimal allocations of the samples to the training sets. Entries in table are $\begin{matrix} t \\ (Acc) \end{matrix}$ where t is the optimal number for the training set and Acc is the average accuracy for a training set of size n. Total sample size is n. "DEG" is the number of independent differentially expressed genes. "Effect" is the standardized fold change for informative genes (difference in mean expression divided by standard deviation). Notation such as "50+" indicates that the MSE was flat, achieving a minimum at t=50 and remaining at that minimum for $t > 50$. (Here, "flat" is defined as having a range of MSE values less than 0.0001.) Data generated with dimension $P = 22,000$. Each table entry based on $1,000$ Monte Carlo simulations. Equal prevalence from each of two classes.

Optimal number to training set				
n = 200				
	Effect = 0.5	**Effect** = 1.0	**Effect = 1.5**	**Effect = 2.0**
DEG=50	170 (86%)	70+ (> 99%)	30+ (> 99%)	20+ (> 99%)
DEG=10	150 (64%)	130 (94%)	100 (99%)	60+ (> 99%)
DEG=1	10 (52%)	150 (69%)	120 (77%)	80 (84%)
n = 100				
DEG=50	70 (64%)	80 (> 99%)	30+ (> 99%)	20+ (> 99%)
DEG=10	10 (55%)	80 (91%)	70 (99%)	40+ (> 99%)
DEG=1	10 (51%)	40 (63%)	80 (77%)	70 (84%)
n = 50				
DEG=50	10 (59%)	40 (99%)	30+ (> 99%)	20+ (> 99%)
DEG=10	10 (52%)	40 (78%)	40 (98%)	40 (> 99%)
DEG=1	10 (50%)	10 (54%)	30 (71%)	40 (83%)

Table 2: Real datasets. Nonparametric bootstrap ($1,000$ bootstrap replicates) with smooth spline (or isotonic regression) learning curve method results (see supplemental material for details). n is the total number of samples from the two classes, and "Prevalence" is the prevalence of the majority class. $t : v$ is the optimal assignment finding, with t the number of samples that should be allocated to the training set, and v the number to the test set ($t + v = n$). $\%t$ is the percent of samples allocated to the training set under optimal allocation, $t/n \cdot 100\%$. is the estimated mean accuracy on the full dataset. "Optimal vs. ..." is the difference between the root mean squared error for an optimal training set allocation and for the "..." allocation. Classes for datasets are: Germinal Center B-cell-like lymphoma versus other (Rosenwald et al., 2002), renal clear cell carcinoma primary tumor versus control normal kidney tissue (Boer et al., 2001), acute myelogenous leukemia versus acute lymphoblastic leukemia (Golub et al., 1999), glioblastoma versus oligodendroglioma (Sun et al., 2006)

Dataset	n	Prevalence	$\%t$	Full dataset Accuracy	Optimal vs. $t = \frac{2n}{3}$ rule	Optimal vs. $t = \frac{n}{2}$ rule
Rosenwald	240	52%	63%	0.96	0.001	0.002
Boer	152	53%	53%	0.98	0.004	2e-4
Golub	72	65%	56%	0.95	0.002	0.004
Sun	131	62%	31%	0.83	0.022	0.008

Figures

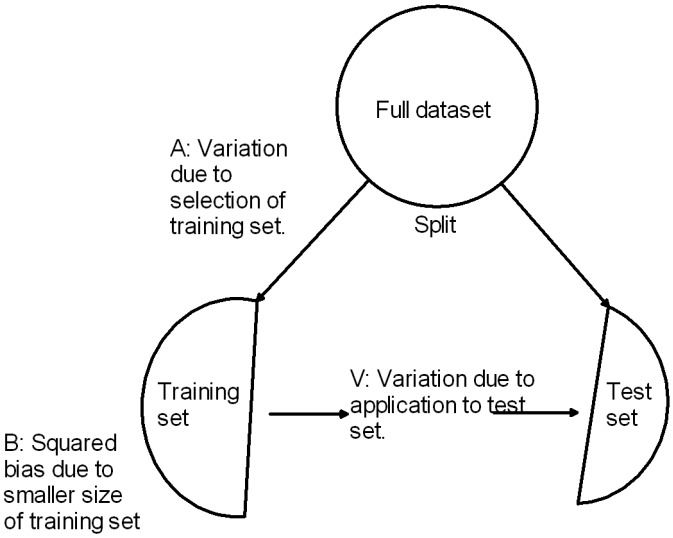

Figure 1: Diagram of mean squared error decomposition.

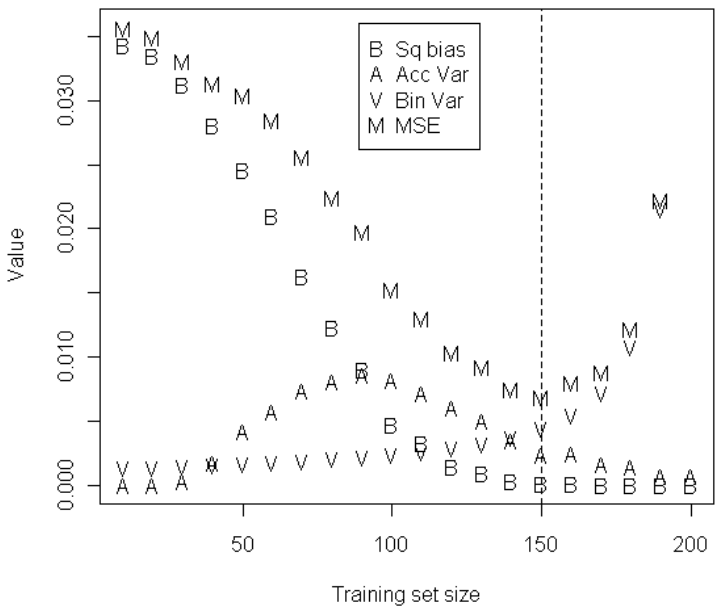

Figure 2: Example figure showing the relative contributions of the three sources of variation to the mean squared error. This is a scenario from one entry in Table 1. Plots for all other scenarios associated with Table 1 and Table S1 appear in the supplement. Here there is m=1 informative gene, n=200 total samples available for study, and the standardized fold change for the informative gene is $2\delta/\sigma = 1.0$.

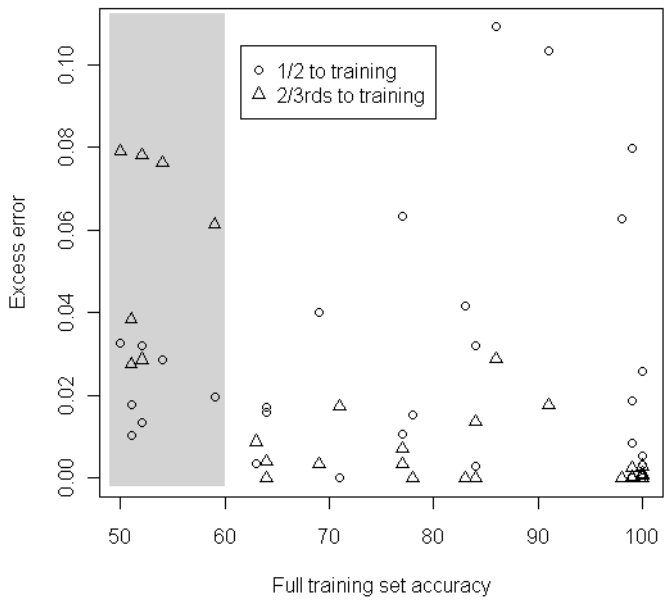

Figure 3: Comparison of two common rules-of-thumb: 1/2 the samples to the training set and 2/3rds of the samples to the training set. X-axis is the average accuracy (%) for training sets of size n. Y-axis "Excess error" on the y-axis is the difference between the root mean squared error (RMSE) and the optimal RMSE. Each point corresponds to a cell in Table 1. Gray shading indicates scenarios where mean accuracy for full dataset size is below 60%.