Estimating Dataset Size Requirements for Classifying DNA Microarray Data

Sayan Mukherjee*[+#1], Pablo Tamayo[+1], Simon Rogers[o1], Ryan Rifkin[+#], Anna Engle[#], Colin Campbell[o], Todd R. Golub[+†] and Jill P. Mesirov[+].

[+] Whitehead Institute / Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139; [†] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115;[#] McGovern Institute and CBCL, Massachusetts Institute of Technology, Cambridge, MA 02139; [o] Department of Engineering Mathematics, University of Bristol, UK.

* Corresponding author

Phone: (617) 258-0263

Fax: (617) 253-2964

Key words: Gene expression profiling, molecular pattern recognition, DNA microarrays, microarray analysis, sample size estimation.

[1] These authors contributed equally to this paper.

Abstract

A statistical methodology for estimating dataset size requirements for classifying microarray data using learning curves is introduced. The goal is to use existing classification results to estimate dataset size requirements for future classification experiments and to evaluate the gain in accuracy and significance of classifiers built with additional data. The method is based on fitting inverse power-law models to construct empirical learning curves. It also includes a permutation test procedure to assess the statistical significance of classification performance for a given dataset size. This procedure is applied to several molecular classification problems representing a broad spectrum of levels of complexity.

1. Introduction

Over the last few years the routine use of DNA microarrays has made possible the creation of large datasets of molecular information characterizing complex biological systems. Molecular classification approaches based on machine learning algorithms applied to DNA microarray data have been shown to have statistical and clinical relevance for a variety of tumor types: Leukemia [Golub et al. 1999], Lymphoma [Shipp et al. 2001], Brain cancer [Pomeroy et al. 2002], Lung cancer [Bhattacharjee et al. 2001] and the classification of multiple primary tumors [Ramaswamy et al. 2001, 2002, Yeang et al. 2001]. In this context, after having obtained initial or preliminary classification results for a given biological system, one is often left pondering the possibility of embarking on a larger and more systematic study using additional samples. This is usually the case when one tries to improve the accuracy of the original classifier or to provide a more rigorous statistical validation of the existing prediction results. As the process of obtaining additional biological samples is often expensive, involved, and time consuming it is desirable to be able to estimate the performance of a classifier for yet unseen larger dataset sizes. In this situation one has to address two sets of questions:

1. For a given number of samples, how significant is the performance of a classifier, i.e. are the results better than what one would expect by chance?

2. If we know the answers to (1) for a range of dataset sizes, can we predict the performance of the classifier when trained with additional samples? Will the

accuracy of the classifier improve significantly? Is the effort to collect additional samples worthwhile?

These two questions arise in other classification tasks with high dimensional data and few samples such as classifying and functional MRI images of patients with neural dysfunction [Golland et al. 2002]. In this paper we develop a methodology for assessing the significance of a classifier's performance via a permutation test. We then fit an inverse power law model to construct a learning curve with error rates estimated from an existing dataset and use this learning curve to extrapolate error statistics for larger datasets. Power calculations [Adcock 1997] are a standard approach to estimate the number of data samples required. However, these approaches do not address our data set size estimation problem for two reasons. First, the assumptions that the underlying data comes from a Gaussian distribution and independence of variables do not hold. Second, the question addressed by power calculations is: given a particular data set size how confident are we of our empirical error estimate. This is very different from asking how the error rate might decrease given more data.

A non-trivial classifier changes its structure as more training data become available and therefore determining how the error rate might decrease becomes a problem of function extrapolation rather than convergence estimation. In this regard it is important not to confuse this problem with the more standard problem of estimating the confidence of an error estimate as a function of training set size: i.e. estimating the variance in an observed quantity, the error estimate, as a function of the number of measurements. In

general, this latter problem is addressed using power calculations or deviation bounds [Adcock 1997, Guyon et al. 1998]. These methods compute bounds or estimates of a given quantity's deviation from its expected value as a function of the number of observations, or in this case, samples. Other methods study the variation produced by technical factors that can be addressed by experimental design or replicating sample preparation or array hybridization [Cheng and Wong 2001, Tseng et al. 2001, Kerr and Churchill 2001a,b]. There are also methods to model differential expression across experiments [Lee and Whitmore 2002] that assess the effect of replication and sample size in increasing the statistical power of ANOVA models. In the context of our problem, these approaches can only help to find bounds on the deviation between the misclassification error rate and its expected value as a function of the number of measurements, i.e., the realizations of the classifier for a given fixed classification dataset size. These standard error estimation methods are therefore not particularly useful in estimating the future performance of a classifier as a function of increasing dataset size with yet unseen additional data. We test our methodology on eight data sets which represent a range of difficulty or complexity of classification. In some cases the distinction is quite dramatic, while in others it is more subtle. The examples are drawn from existing cancer classification data sets where discriminating the morphology of a sample (five sets) represents the "easier" end of the range, and predicting treatment outcome (three sets) lies at the other extreme. This hierarchy of difficulty will be reflected by the increase in the data set size requirements we estimate for these prediction problems. Our results give an indication of the minimal number of samples

that are needed to obtain significant performance data and to extrapolate the improvement one might get by building the classifier on a larger data set.

In the next section, we will give some background on general approaches addressing the problem of estimating classifier performance and learning rates. In Section 3, we describe our methodology in more detail. The results of applying our methodology to molecular classification problems are contained in Section 4. Section 5 summarizes the results of our tests. The proofs and technical details have been collected in the Appendices.

Section 2. Background and General Approach

The problem of estimating performance of a classifier for larger yet unseen sets of examples is a difficult analytical problem. It amounts to developing a model to compute how fast a given classifier "learns" or improves its "fitting" to the data as a function of dataset size.

In machine learning, a natural way to study classification accuracy as a function of training set size is by building empirical scaling models called learning curves [Cortes et al. 1994]. Learning curves estimate the empirical error rate as a function of training set size for a given classifier and dataset. The advantage of this approach is that one avoids making assumptions about the distribution generating the dataset or the distribution of the classification errors. These learning curves are usually well characterized by inverse power-laws:

$$\boldsymbol{e(n) = an^{-\alpha} + b}. \qquad (1)$$

The variables are the expected error rate $\boldsymbol{e(n)}$ given $\boldsymbol{n}$ training samples, the learning rate $\boldsymbol{a}$, the decay rate α, and the Bayes error $\boldsymbol{b}$ which is the minimum error rate achievable [Devroye et al. 1997, Duda et al. 2000]. Notice that the value of the constants $\boldsymbol{a}, \boldsymbol{b}, \text{and } \alpha$ will change according to the classifier and dataset being studied. Based on this scaling model, as the size of the dataset increases the misclassification error of a classifier will asymptotically approach $\boldsymbol{b}$. This inverse power-law "learning"

behavior appears to be universal and is observed for many classifiers and types of datasets [Cortes et al. 1994, Shrager et al. 1988]. It is in fact observed not only in machine learning but also in human and animal learning [Anderson et al. 1983]. It is common to find empirical α values around or less than 1. Besides its empirical prevalence, the power-law model can be motivated analytically and in some cases derived within the statistical mechanics approach to learning. The basic idea behind this approach is to formulate the average error as a set of equations which are then solved via a statistical mechanics replica approach [Hertz et al. 1991] involving integration over the parameters of the classifier. This approach has been applied to various classification algorithms such as Support Vector Machines [Dietrich et al. 2000], Large Margin Perceptrons [Opper et al. 1995], Adaline and other classifiers based upon Hebbian rules [Opper et al. 1990]. The resulting analysis of the classification errors for all of the above algorithms results in inverse power-laws of the form (1).

Using this power-law scaling model as a basis, one can use the empirical error rates of a classifier over a range of training set sizes drawn from a dataset to fit an inverse-power law model and then use this model to extrapolate the error rate to larger datasets. In order to make this a practical approach one also needs a statistical test for classifier significance as a function of training set size. The reason for this is that the inverse power-law model usually breaks down for small training set sizes where the model lacks enough data give accurate predictions. In this case, the error rates are large and not significant. If a given classifier's results are not significant, then it is better to exclude them when fitting the learning curve. To directly address this problem we

have included a permutation test for the significance of a classifier as part of our methodology. This test compares the performance of the actual classifier with the performance of random classifiers trained to predict data whose target labels are permuted (randomized). A classifier that is able to “find” structure in the data and produce significant results should outperform its random counterparts most of the time. By fixing a significance level (0.05) we can produce an effective test to eliminate classifiers that are not significant from the fitting of the learning curve. Since the classifier performance usually improves with increasing training set size, this significance test also allows us to find the minimum number of samples that produced significant classifiers.

Section 3. Methodology for estimating error rates as a function of dataset size

Given an arbitrary input dataset and classification algorithm the methodology we will describe provides the following,

1) A measure of the statistical significance of the classifier built at each training set size. Based on this one can find the minimum training set size for which the classification performance of the classifiers is statistically significant.
2) An analytic expression (power-law) of the error rate as a function of the increasing dataset size as well as similar expressions for the 25^{th} and 75^{th} error rate quantiles. These provide a means to extrapolate the error bar "envelope" for the error rate for larger yet unseen data sets.

As described in Section 2, a significance test is needed to know at which training set size error rates are reliable enough to accurately extrapolate the error rate as a function of dataset size. The 25^{th} and 75^{th} quantiles are used to compute the analog of error bars for the estimated error rates as a function of dataset size. Figure 1 shows a pictorial summary of the method. The procedure can be broken down into two main computational tasks, the first involving random sub-sampling (train/test) and a significance permutation test to evaluate the classifiers, and the second consisting of fitting learning curves to the error rates that passed the significance test.

We first describe the rationale behind fitting the learning curves. We want to fit the inverse power curves to the true average error and the true 25th and 75th quantile error of a classifier trained with various training set sizes. Our first step is to estimate the true average error rate and 25th and 75th error quantiles for a range of training set sizes. For a dataset of size ℓ with fixed training set size, n, and test set size, $\ell - n$, T_1 train/test realizations were constructed by sub-sampling the dataset. For each of these T_1 realizations an error rate $e_{n,i}$ is computed and the average of the sequence $\{e_{n,1},...,e_{n,T_1}\}$, $\bar{e}_n = \frac{1}{T_1}\sum_{i=1}^{T_1} e_{n,i}$, is used as an estimate of the true average error e_n. The average error rate is an unbiased estimator of the true error rate of a classifier trained with n samples,

$$E_{D_\ell}\, \tfrac{1}{T_1}\sum_{i=1}^{T_1} e_{n,i} = E_{D_n}\, e_n,$$

where $E_{D_n}\, e_n$ is the probability of classifying a new sample incorrectly when the classifier was trained with n samples (see Appendix 1a for proof and details). The 25th and 75th percentile of the sequence $\{e_{n,1},...,e_{n,T_1}\}$ were also fitted to an inverse power law so that we could also estimate the variation in error rates as a function of dataset size. The 25th and 75th percentile of the sequence $\{e_{n,1},...,e_{n,T_1}\}$ are good approximations of the 25th and 75th quantiles of error rates of classifiers trained with n samples (see Appendix 1c for proof and details). We did not use the variance of the error rates $e_{n,i}$ because this statistic is not an unbiased estimator of the variance of the error rate of classifiers trained with n samples and tested on a new sample. Indeed one can prove

that the variance of this statistic is in general optimistic: the variance of the error rates $e_{n,i}$ is less than the variance of classifiers trained with n samples and tested on a new sample (see Appendix 1b for proof and details).

As described in the introduction, theoretical justifications for the use of inverse power laws can be made using analyses of classification accuracy based upon techniques from statistical mechanics [Opper et al. 1990, Opper et al. 1995] and approximation theory [Niyogi et al. 1996] as described in more detail in Appendix 2.

Fitting the parameters of the learning curve by minimizing

$$\min_{\alpha,a,b} \sum_{l=1}^{M} \left(a n_l^{-\alpha} + b - \bar{e}_{n_l} \right)^2 \quad \text{subject to } \alpha, a, b \geq 0$$

is a convex optimization problem when b is fixed. For a fixed b, one can estimate α and a by taking logarithms and solving the following equivalent linear minimization problem

$$\min_{\alpha,a,b} \sum_{l=1}^{M} \left(\ln(a) - \alpha n_l + \ln(b - \bar{e}_{n_l}) \right)^2 \quad \text{subject to } \alpha, a, b \geq 0$$

Solving this linear problem for various values of b followed by line search gives us our estimate of $\alpha, a, \text{and } b$.

As described in Section 2, the fitted learning curve does not extrapolate accurately when error rates are large and not statistically significant. This motivates a procedure to determine at what training set size the error rate is statistically significant when compared to the null hypothesis of the error rate of a random classifier

$$\boldsymbol{H}_0: \qquad \boldsymbol{p}\left(\boldsymbol{y}=1 \mid \boldsymbol{x}, \left\{\boldsymbol{x}_1, \boldsymbol{y}_1, ..., \boldsymbol{x}_{\boldsymbol{n}}, \boldsymbol{y}_{\boldsymbol{n}}\right\}\right) = \boldsymbol{p}\left(\boldsymbol{y}=-1 \mid \boldsymbol{x}, \left\{\boldsymbol{x}_1, \boldsymbol{y}_1, ..., \boldsymbol{x}_{\boldsymbol{n}}, \boldsymbol{y}_{\boldsymbol{n}}\right\}\right),$$

the null hypothesis states that given a particular training set the conditional probability of a label being 1 or -1 is equal. A random classifier is built from the same input data with the class labels of the data randomly permuted. This addresses the question of how well the classifier can learn the mapping $\boldsymbol{f}: \boldsymbol{x} \to \boldsymbol{y}$ when the $\boldsymbol{y}$ values are random, $\boldsymbol{y} = \{1, -1\}$. In essence we ask how well a classifier trained on randomly labeled data can classify correctly labeled data. The permutation procedure outlined above helps to answer this question. For each train/test realization for which an error rate $\boldsymbol{e}_{\boldsymbol{n},\boldsymbol{i}}$ was computed we construct $\boldsymbol{T}_2$ randomized realizations where the labels of the training set are randomly permuted. We build classifiers on these randomized training sets and test on the corresponding test set. This results in a set of error rates $\boldsymbol{e}_{\boldsymbol{n},\boldsymbol{i},\boldsymbol{j}}$ for training set size $\boldsymbol{n}$. From these error rates we construct an empirical distribution function for the random classifier,

$$\boldsymbol{P}_{\boldsymbol{n}}^{\boldsymbol{ran}}\left(\boldsymbol{x}\right) = \frac{1}{\boldsymbol{T}_1 \times \boldsymbol{T}_2} \sum_{\boldsymbol{i}=1}^{\boldsymbol{T}_1} \sum_{\boldsymbol{j}=1}^{\boldsymbol{T}_2} \boldsymbol{\theta}\left(\boldsymbol{x} - \boldsymbol{e}_{\boldsymbol{n},\boldsymbol{i},\boldsymbol{j}}\right),$$

where $\boldsymbol{\theta}(\boldsymbol{z}) = 1 \text{ if } \boldsymbol{z} \geq 0 \text{ and } 0$ otherwise. The significance of the classifier is $\boldsymbol{P}_{\boldsymbol{n}}^{\boldsymbol{ran}}\left(\bar{\boldsymbol{e}}_{\boldsymbol{n}}\right)$ which is the percentage of random classifiers with error rate smaller than $\bar{\boldsymbol{e}}_{\boldsymbol{n}}$. The procedure is illustrated with the following two examples for a fictitious dataset:

Example 1: $\bar{e}_{15} = .37$ (Error rate of the classifier with 15 samples). $e_{15,i,j} = \{.215,.260,.290,.320,.366,.388,.395,.408,.420,.495\}$ (Error rates of the random classifiers with 15 samples). $P_{15}^{ran}(\bar{e}_{15}) = .50$. Since the p-value is greater than .05 the error rate of the classifier is not statistically significant (see figure 2a).

Example 2: $\bar{e}_{45} = .1$ (Error rate of the classifier with 45 samples). $e_{45,i,j} = \{.205,.270,.333,.337,.370,.392,.399,.406,.425,.499\}$ (Error rates of the random classifiers with 45 samples), $P_{45}^{ran}(\bar{e}_{45}) = 0$. Since the p-value is less than .05 the error rate of the classifier is statistically significant (see Figure 2b).

The detailed description of the entire methodology for a two-class problem is as follows:

1) Sub-sampling and significance permutation test
    a. Sub-sampling procedure
        i. Given ℓ_{c_1} samples from class 1 and ℓ_{c_2} samples from class 2, the total number of samples is $\ell = \ell_{c_1} + \ell_{c_2}$, where $\ell \geq 10$.
        ii. Select 10 training set sizes $(n_1,..,n_j,..n_{10})$ over the interval $[10, \ell - 10]$.

1. For each training set size $\boldsymbol{n}_j$ run the following sub-sampling procedure $T_1 = 50$ times, indexed by $\boldsymbol{i} = 1,...,\boldsymbol{T}_1$
    a. Randomly split the dataset into a training set with $\boldsymbol{n}_j$ samples and a test set with $\ell - \boldsymbol{n}_j$ samples subject to the requirement that $\frac{\boldsymbol{n}_{c_2}}{\boldsymbol{n}_{c_1}} \approx \frac{\ell_{c_2}}{\ell_{c_1}}$ where $\boldsymbol{n}_{c_2}$ and $\boldsymbol{n}_{c_1}$ are the number of samples from class 1 and class 2 in the training set. Call the two datasets generated $\boldsymbol{S}_{n,i}$
    b. Train a classifier on each of the training sets and measure its error rate on its corresponding test set, call each of these error rates $\boldsymbol{e}_{n,i}$

b. Permutation test
    i. For each sub-sampled train/test split $\boldsymbol{S}_{n,i}$ run the following permutation procedure $\boldsymbol{T}_2 > 50$ times, indexed by $\boldsymbol{j} = 1,...,\boldsymbol{T}_2$
        1. Randomly permute the labels of the samples in the training set (leave the test set al.one), call the dataset generated $\boldsymbol{S}^{ran}_{n,i,j}$
        2. Train a classifier on the training set and measure its error on the test set, call this error rate $\boldsymbol{e}^{ran}_{n,i,j}$

c. Significance calculation

1. For each training set size $\boldsymbol{n}$ construct an empirical distribution function from the error rates of the permuted datasets $\boldsymbol{P_n^{ran}(x) = \frac{1}{T_1 \times T_2} \sum_{i=1}^{T_1} \sum_{j=1}^{T_2} \theta(x - e_{n,i,j}^{ran})}$, where $\boldsymbol{\theta(z) = 1}$ if $z \geq 0$ and 0 otherwise.
2. Given the above empirical distribution function compute for each $\bar{e}_n$ the value $t_n = P_n^{ran}(\bar{e}_n)$, statistical significance with respect to an α-value of p is achieved for n_0, the smallest n for which $t_n < p$

2) Learning curves and training set size estimation

a. Assume the sub-sampling procedure was run for M different sample sizes n, indexed by $l = 1,...M$, take the sequence of error rates and compute the following quantities for each training set size $n > n_0$ for which the classifier passed the significance test ($t_n < p$): he mean error rate $\bar{e}_n = \frac{1}{T_1} \sum_{i=1}^{T_1} e_{n,i}$, the 25th, and 75th quantiles of the vector of error rates $\{e_{n,1},...,e_{n,T_1}\}$.

b. Use the above quantities to fit the following learning curves:

i. Given training set sizes n_l and mean error rates $\bar{e}_{n_l}$ compute α, a, b via the following minimization procedure: $\min_{\alpha,a,b} \sum_{l=1}^{M} \left(a n_l^{-\alpha} + b - \bar{e}_{n_l}\right)^2$ subject to $\alpha, a, b \geq 0$, designate the values

$\alpha, \boldsymbol{a}, \boldsymbol{b}$ as $\alpha_{\boldsymbol{m}}, \boldsymbol{a}_{\boldsymbol{m}}, \boldsymbol{b}_{\boldsymbol{m}}$. The resulting curve estimates the error rate as a function of training set size

$$\boldsymbol{L}_{\boldsymbol{m}} : \boldsymbol{e}(\boldsymbol{n}) = \boldsymbol{a}_{\boldsymbol{m}} \boldsymbol{n}^{-\alpha_{\boldsymbol{m}}} + \boldsymbol{b}_{\boldsymbol{m}}.$$

ii. Repeat the above procedure for the 25th and 75th quantiles of the vector of error rates $\left\{\boldsymbol{e}_{\boldsymbol{n},1}, \ldots, \boldsymbol{e}_{\boldsymbol{n},\boldsymbol{T}_1}\right\}$

Section 4. Methodology applied to several cancer classification problems

The procedure outlined in the previous section has been applied to eight binary DNA microarray cancer classification problems representing a broad range of level of complexity of classification. The classification problems are to discriminate between tumor morphologies. (including disease vs. normal and submorphologies) or treatment outcome. A more detailed analysis of the methodology will be given for the largest dataset (cancer vs. normal tissue classification). For the seven other datasets we will present only the final results.

The set of examples falls into two cases. The first case consists of classification problems that are relatively easy and where statistical significance for the classifiers is achieved at a low number of training set samples (e.g. between 10-20 samples) and where the dataset is sufficiently large (e.g., 40 samples) to accurately fit a learning curve. The second case consists of classification problems that are more difficult and where statistical significance is achieved at between 40-60 training samples while the total number of samples in the dataset is barely larger (e.g. between 50-70). For these more difficult problems we cannot strictly follow the methodology since we do not have enough training set sizes at which significance is reached to make an accurate fit of the learning curve. However, we can still fit the curves and use the results as indicative and exploratory. A possible third case is when significance is never reached for any available training set size for a dataset. In this case it is difficult to draw any conclusion but it is possible that either adding more samples will not help (e.g. because there is not sufficient molecular information to classify this dataset) or the problem is very hard and

substantial numbers of samples are needed before one sees an statistically significant results. Our first five morphology datasets are examples of the first case. The final three treatment outcome datasets are examples of the second case. Table 1 summarizes the learning curve parameters and extrapolated error rate estimates at 400 training samples for all the data sets. Table 2 summarizes and comments on the results of running the methodology on the various datasets. General conclusions and interpretation of the results will be presented in the next section.

Tumor vs. normal tissue

This dataset consists of expression levels for 180 samples of a variety of different primary tumors (breast, prostate, lung etc…) and 100 normal samples from the corresponding tissue of origin (again breast, prostate, lung etc…) [Ramaswamy et al. 2001b]. The dimensionality of the dataset is 16063 (throughout this section by dimensionality we mean the number of gene expression values recorded for a sample). No preprocessing was performed. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel and no feature selection. Error rates were estimated for training set sizes of $\boldsymbol{n} = (30,40,50,60,80,90,130,170,210)$. A leave-one-out model built with all the available samples (280) was used to validate the method and to compare the scaling model to the error rate achieved when using almost the entire dataset, this corresponds to a training set size of $\boldsymbol{n} = 279$.

Figure 3 illustrates the results of the significance permutation test for this dataset, i.e., the statistical significance of classifiers with training sets of 15 and 30 samples. As can be seen in Figure 3b, with 30 samples most of the random classifiers attain larger error rates than the actual classifier. For the case using 15 samples, about one in 20 of the random classifiers attain the same or better error rates and therefore a p-value of 5% is achieved. To fit the learning curves we will use only data points obtained from training sets of size greater than or equal to 15.

To study the improvement of the learning curve estimates as a function of the number of training set sizes used to fit the learning curves we constructed four learning curves using: 1) the error rates for all training set sizes (up to 210), 2) the error rates for the first 8 training set sizes, 3) the error rates for the first 6 training set sizes, 4) the error rates for the first 4 training set sizes. The plots of these learning curves along with the leave-one-out error for 280 samples is given in Figure 4. As expected the model improves as more and larger training set sizes are used in the fit. The actual leave-one-out error rate achieved with 280 samples is only about 2% less than the error rate estimated for 279 training samples by extrapolating the learning curve model. Figure 5 shows the curve for the power law that results from applying the methodology to a) all training samples sizes stated above (up to 210) and b) using the first 6 training set sizes (up to 90), along with the leave-one-out error for the entire dataset (280 samples). The expression for the error rate as a function of n estimated using training sets sizes (up to 210) is

$$\boldsymbol{error}(\boldsymbol{n}) = 1.42\boldsymbol{n}^{-0.52} + .0098.$$

The error rates for the 25th and 75th quantiles are

$$\boldsymbol{error}_{25}(\boldsymbol{n}) = 1.89\boldsymbol{n}^{-0.63} + .0032$$

$$\boldsymbol{error}_{75}(\boldsymbol{n}) = 1.17\boldsymbol{n}^{-0.43} + .000.$$

Based on this model one can see clearly how fast the error rate decreases with increasing dataset size. The asymptotic Bayes error rate $\boldsymbol{b}$ is very small indicating that indeed very low errors can be achieved if a large dataset were used to train the classifier. The decay rate α is about .5 indicating that, in scaling terms, this is a rather difficult problem for the model to learn. The size of the 25th and 75th quantiles envelope is about +/- 2% and it indicates that the model is relatively accurate. If we were going to collect 400 training samples this model can be used to extrapolate the error rate as follows:

$$\boldsymbol{error}(400) = 1.42(400)^{-0.52} + .0098 = 7.3\%$$

$$\boldsymbol{error}_{25}(400) = 1.89(400)^{-0.63} + .0032 = 4.7\%$$

$$\boldsymbol{error}_{75}(400) = 1.17(400)^{-0.43} + .000 = 8.9\%.$$

The achievable error rate using 400 samples according to the model is $7.3 \pm$ 2.6% and perhaps as low as 4.7% (25th quantile envelope).

Leukemia morphology

The dataset consists of expression levels for 48 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (ALL) [Golub et al. 1999]. The dimensionality of the dataset is 7129. No preprocessing was performed. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel and no feature selection. Error rates were estimated for training set sizes of $\boldsymbol{n} = (10,15,20,25,30,35)$. In Figure 6 a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 73 samples. A p-value of 5% is achieved at about 5 samples. The learning curve estimate of the error rate as a function of $\boldsymbol{n}$ is

$$\boldsymbol{error}(\boldsymbol{n}) = .7706\boldsymbol{n}^{-.65} + .009\,.$$

In this case, the learning and decay rates are such that the model clearly learns more quickly than in the previous example as a function of training set size. It achieves practically a zero error rate at 73 samples (consistent with the 25th quantile envelope). The envelope is wider in this case because we fit the model using a narrower range of dataset sizes over which the empirical error rates display more variation than the previous dataset.

Colon cancer

The dataset consists of expression levels for 22 samples of normal colon tissue and 40 samples of malignant tissue [Noterman et al. 2001]. The dimensionality of the dataset is 2000. The data was preprocessed by taking the natural logarithm of all input values, and then applying a hyperbolic-tangent function. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel and no feature selection. Error rates were estimated for training set sizes of $\boldsymbol{n} = (10,15,20,25,30,35,40,45,50)$. In Figure 7, a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 62 samples.

A p-value of 5% is achieved at about 10 samples. The learning curve estimate of the error rate as a function of $\boldsymbol{n}$ is

$$\boldsymbol{error}(\boldsymbol{n}) = .4798\boldsymbol{n}^{-0.2797}.$$

Ovarian cancer

The dataset consists expression levels for 24 samples of normal ovarian tissue and 30 samples of malignant tissue [Schummer et al. 1999]. The dimensionality of the dataset was 1536. The data was preprocessed by adding 1 and taking the natural logarithm of all input values. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel and no feature selection. Error rates were estimated for training set sizes of $\boldsymbol{n} = (10,15,20,25,30,35,40)$. In Figure 8, a plot of the learning curve and its 25th and 75th

quantiles is given along with the leave-one-out error of the 54 samples. A p-value of 5% is achieved at about 10 samples. The learning curve estimate of the error rate as a function of n is

$$error(n) = .7362n^{-0.6864}.$$

Lymphoma morphology

The dataset consists of expression levels for 24 samples of diffuse large B-cell lymphoma and 12 samples of follicular lymphoma and chronic lymphocyptic [Alizadeh et al. 2000]. The dimensionality of the dataset was 18,432. The data was preprocessed by taking the base 2 logarithm of all input values. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel and no feature selection. Error rates were estimated for training sizes of $n = (5,10,15,20,25,30,35,40)$. In Figure 9, a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 36 samples. A p-value of 5% is achieved at about 5 samples. The learning curve estimate of the error rate as a function of n is

$$error(n) = .57n^{-0.7073} + .0006.$$

Brain cancer treatment outcome

The dataset was obtained from 39 samples of patients that had successful treatment outcome (alive two years after treatment) and 21 samples of patients with poor

treatment outcome. All patients had childhood Medulloblastomas [Pomeroy et al. 2002]. The dimensionality of the dataset is 7129. No preprocessing was performed. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel selecting 150 features using the radius-margin criteria [Chapelle et al. 2001]. Error rates were estimated for training set sizes of $\boldsymbol{n} = (20,25,30,35,40)$.

Statistical significance on this dataset (a p-value of 5%) is achieved at about 45 samples, which is larger than any of the training set sizes for which error rates were estimated so strictly speaking we cannot apply the methodology.

However, we can examine how accurately a learning curve fit to the error rates for the above training set sizes can extrapolate. In Figure 10, a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 60 samples. As expected, this model is not very accurate and over estimates the error rate at 59 samples by more than 7%. The learning curve estimate of the error rate as a function of $\boldsymbol{n}$ is

$$\boldsymbol{error}(\boldsymbol{n}) = 1.115\boldsymbol{n}^{-.3295} + .006 .$$

Lymphoma treatment outcome

The dataset was obtained from 32 samples of patients that had successful treatment outcome (alive two years after treatment) and 26 samples of patients with poor treatment outcome. All patients had diffuse large cell lymphoma (DLCL) [Shipp et al.

2001]. The dimensionality of the dataset is 7129. No preprocessing was performed. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel selecting 150 features using the radius-margin criteria [Chapelle et al. 2001]. Error rates were estimated for training set sizes of $\boldsymbol{n} = (20,25,30,35,40)$. Statistical significance on this dataset (a p-value of 5%) is achieved at about 50 samples. In Figure 11 a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 58 samples. As expected, this model is not very accurate and over estimates the error rate at 57 samples by more than 9%. The learning curve estimate of the error rate as a function of $\boldsymbol{n}$ is

$$\boldsymbol{error}(\boldsymbol{n}) = .9431\boldsymbol{n}^{-.2957} + .01.$$

Breast cancer treatment outcome

The dataset consists expression levels of 34 samples from patients with breast tumors that metastasized within five years of disease onset and 44 samples from patients that were disease free for at least five years [Van't Veer et al. 2002]. The dimensionality of the dataset was 24,624. No preprocessing was performed. The classifier used was a Support Vector Machine [Vapnik 1998] with a linear kernel without feature selection. Error rates were estimated for training set sizes of $\boldsymbol{n} = (10,20,30,40,50,60,70)$. Statistical significance on this dataset (a p-value of 5%) is achieved at about 65 samples. In Figure 12 a plot of the learning curve and its 25th and 75th quantiles is given along with the leave-one-out error of the 78. As expected, this model is not very accurate and over

estimates the error rate at 77 samples by more than 6%. The learning curve estimate of the error rate as a function of n is

$$error(n) = .4852n^{-.0733} + .01.$$

Section 5. Conclusions

We have described a methodology for assessing the significance of a classifier's performance via a permutation test and constructing a learning curve to extrapolate error statistics for larger data sets that include yet unseen samples. We applied the method to eight cancer classification problems of varying levels of difficulty. Based on the results of the previous section, one can see that the inverse power-law scaling model proposed fits the empirical error rates reasonably well. The classifier we used was an SVM but the methodology is applicable to other algorithms (e.g. weighted voting, k-nearest neighbors, logistic regression etc.). For the morphology classification problems the extrapolation is quite accurate. For the treatment outcome classification problems the combination of the increased complexity of the problems and the limited dataset sizes yield a less accurate, but still indicative extrapolation. As expected, the model improves as larger training samples sizes are used in the learning curve fit (see Figs. 4 and 5). The learning curves bear out the empirical observation that morphological distinctions are more dramatic and thus, in general, "simpler" problems than the more subtle distinctions that must be determined for treatment outcome prediction. Significance on morphology problems is achieved with 10-20 training samples and “reasonably accurate” extrapolation requires 30-40 training samples. In contrast, for treatment outcome, significance is achieved with 45-60 training samples and “reasonably accurate” extrapolation requires on the order of 75-100 training samples. For morphological distinctions the learning curve prediction is reasonably close to the actual leave-one-out error measured at a larger size. The 25^{th} and 75^{th}

quantile models provide useful error bar envelopes that enclose the observed error rates for those problems. For treatment outcome prediction, due to the large training set size required to achieve significance and small available dataset sizes, we do not have enough significant classifiers with which to construct an accurate learning curve. Consequently, we get less accurate estimates of the leave-one-error on the entire dataset for the outcome treatment examples with differences of 7% for brain tumor outcome, 9% for Lymphoma treatment outcome, and 8% for breast tumor metastasis.

The estimation of the asymptotic Bayes error ***b***, the learning rate ***a***, and decay rate α, can also be used directly to characterize the difficulty of a problem and the complexity of a model. They can provide a basis for comparing and contrasting models and problems. To illustrate, we show in Figure 13 the values of these parameters for the examples discussed in the paper. The morphology and treatment outcome datasets cluster with respect to α, and ***b***. We have not elaborated on this aspect of the analysis but it is certainly an interesting direction to pursue in the future.

In summary, our methodology produces reasonable, non-trivial dataset size estimates when applied to a fairly general set of molecular cancer classification problems. In this context it can serve as a valuable tool when designing future experiments, either for evaluating whether it is worthwhile to collect additional samples, or for obtaining a deeper insight into the complexity of a given classification problem based on preliminary data. Table 1 shows a summary of the results for the examples described in this paper. The results of applying this method to those examples suggest that minimum dataset

size requirements for morphological classifications are typically in the 10-20 samples range and upwards of 50 samples for treatment outcome classification. These results can be used to provide general rule of thumb guidelines but the exact numbers for a given problem are dataset and classifier dependent. This method can also be applied to other domains where a prospective estimation of the number of samples is relevant as is the case in many problems using molecular features to classify biological samples, e.g., classification based on proteomic mass spec. data, chemosensitivity prediction, survival analysis, and putative class discovery using clustering.

Acknowledgements

We are indebted to members of the Cancer Genomics Group, Whitehead / MIT Center for Genome Research and the Golub Laboratory, Dana-Farber Cancer Institute for many valuable discussions. This work is supported in part by grants from Affymetrix Inc., Millennium Pharmaceuticals Inc., Bristol-Myers Squibb Company, and the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-01ER63185

References

Adcock, C. 1997. Sample size determination: a review. The Statistician 46, 261-283.

Alizadeh A.A., et al. 2000. Distinct Types of Diffuse Large B-Cell Lymphoma identified by gene expression profiling. Nature 403, 503-511.

Anderson, J. 1983. Architecture of Cognition Harvard University Press, Cambridge, MA.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M. 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA 98, 13790–13795.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S. 2001. Choosing Multiple Parameters for Support Vector Machines. Machine Learning.

Chipping Forecast 1999. Nature Genetics volume 21, Supplement.

Cortes, C. Jackel, L., Solla, S., Vapnik, V., and Denker, S. 1994. Asymptotic values and rates of convergence. Advances in Neural Information Processing Systems VI, Morgan Kaufmann Publishers, San Francisco, CA.

Devoye, L., Gyorfi, L. and Lugosi, G. 1997. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.

Dietrich, R., Opper, M., and Sompolinsky, H. 1999. Statistical Mechanics of Support Vector Networks. Physical Review Letters 82, 2975–2978.

Dietrich, R., Opper, M., and Sompolinsky, H. 2000. Support Vectors and Statistical Mechanics, 359-368. In Smola, A.J., and Bartlett, P.J., eds., Advances in Large Margin Classifiers, MIT Press, Cambridge, MA.

Engel, A., and Van den Broeck, C. 2001. Statistical Mechanics of Machine Learning. Cambridge University Press.

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, 531–537.

Golland, P., Mukherjee, S., Tamayo, .P., and Fischl, B. 2002. Permutation Tests for Classification. Submitted to Advances in Neural Information Proceedings.

Good, P. 1994. Permutation Tests: A Practical guide to Resampling Methods for Testing Hypothesis. Springer-Verlag, New York.

Guyon, I., Makhoul, J., Schwartz, R., and Vapnik, V. 1998. What size test set gives good error estimates? IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 52-64.

Hertz, J., Krogh, A. and Palmer, R. 1991. Introduction to the Theory of Neural Computation. Addison-Wesley, Boston, MA.

Kerr, A., and Churchill, G. 2001a. Experimental design for gene expression microarrays. Biostatistics 2, 183-201.

Kerr, A., and Churchill, G. 2001b. Statistical design and the analysis of gene expression microarrays. Genetic Research 77, 123-128.

Li, C., and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2(8): research 0032.1-0032.11.

Niyogi, P., and Girosi, F. 1996. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. Neural Computation 8, 819-842.

Noterman, D., Alon, U., Sierk, A., Levine, A. 2001. Transcriptional Gene Expression Profiles for Colorectal Adenoma, Adenocarcinoma and Normal Tissue Examined by Oligonucleotide Array. Cancer Research 61, 3124-3130.

Opper, M., Kinzel, W., Kleinz, J., and Nehl, R. 1990. On the Ability of the Optimal Perceptron to Generalise. Journal of Physics 23, 581-586.

Opper M. and Haussler D. 1991. Generalization performance of Bayes optimal classification algorithms for learning a perceptron. Physical Review Letters, 66, 2677-2680.

Opper, M., and Kinzel, W. 1995. Statistical Mechanics of Generalisation in Models of Neural Networks, ed. Domany, E. and van Hemmen, J.L. and Schulten, K., Springer Verlag, Heidelberg.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky,

G., Louis, D., Mesirov, J.P., Lander, E., Golub, T. 2002. Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. Nature 415, 436-442.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R. 2001a. Multiclass cancer diagnosis by using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98, 15149-15154.

Ramaswamy, S., Osteen, R.T., & Shulman, L.N. 2001b. Metastatic Cancer from an Unknown Primary Site, 711-719. in Lenhard, R.E., Osteen, R.T., & Gansler, T., eds., Clinical Oncology, American Cancer Society, Atlanta, GA.

Rosenblatt F. 1962. Principles of Neurodynamics. Spartan Books, New York.

Schummer, M., Ng, W., Bumgarner, R., Nelson, P., Schummer, B., Bednarski, D., Hassell, R., Baldwin, R., Karlan, B., Hood, L. 1999. Comparitive hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 238, 375-385.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M.,Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T.,

Mesirov, J.P., Neuberg, D., Lander, E., Aster, J.,Golub, T. 2001. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medicine 8, 68-74.

Shrager, J., Hogg, T. and Huberman, B.A. 1988. A graph-dynamic model of the power law of practice and the problem-solving fan effect. Science 242, 414-416.

Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S. 2000. Class prediction and discovery using gene expression data, 263–272. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology. RECOMB. ACM Press, New York.

George, T., Oh, M., Rohlin, L., Liao, J.C. and Wong, W.H. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. Nucleic Acids Research 29, 2549-2557.

Watkin T., Rau A., and Biehl M. 1993. The Statistical Mechanics of Learning a Rule. Reviews of Modern Physics 65, 499-556.

Vapnik, V.N. 1998. Statistical Learning Theory. John Wiley & Sons, New York.

Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J.P., Golub, T. 2001. Molecular classification of multiple tumor types. Bioinformatics 17 (Suppl. 1):S316–S322.

Appendix 1: Bias properties of the mean, variance, and quantiles of leave-p-out estimators

1a) The mean of the leave-p-out estimator is unbiased.

Statement 1: The procedure of excluding p samples from a dataset of size ℓ, constructing a classifier, and then testing on the p samples left out is designated as follows:

$$L_p(z_1,...z_\ell) = \frac{1}{p}\sum_p Q\left(z_p, f\left(z_{\ell-p}\right)\right)$$

where $z = (x, y)$, $f\left(z_{\ell-p}\right)$ is the classifier constructed with p samples left out and $Q\left(z_p, f\left(z_{\ell-p}\right)\right)$ is the error of this classifier on the p samples left out. This procedure is unbiased:

$$E\ L_p(z_1,...z_\ell) = E\ Q\left(z, f\left(z_{\ell-p}\right)\right)$$

which means that the expected error when the classifier is trained with $\ell - p$ samples is the same as the expected error of procedure L_p.

Proof:

The proof is a straightforward extension of the leave-one-out case which was derived by Luntz and Brailovsky [Luntz et al. 69] by the following series of transformations:

$$
\begin{aligned}
\boldsymbol{E}\ \ \boldsymbol{L}_1(z_1,...,z_\ell) &= \frac{1}{\ell}\int \sum_{i=1}^{\ell} \boldsymbol{Q}\left(z_i, \boldsymbol{f}\left(z_{\ell-1}\right)\right) \boldsymbol{dP}(z_1)...\boldsymbol{dP}(z_\ell) \\
&= \frac{1}{\ell}\int \sum_{i=1}^{\ell} \left(\boldsymbol{Q}\left(z_i, \boldsymbol{f}\left(z_{\ell-1}\right)\right) \boldsymbol{dP}(z_i)\right) \boldsymbol{dP}(z_1)...\boldsymbol{dP}(z_{i-1})\boldsymbol{dP}(z_{i+1})...\boldsymbol{dP}(z_\ell) \\
&= \frac{1}{\ell}\sum_{i=1}^{\ell} \boldsymbol{E}\ \ \boldsymbol{Q}\left(z_i, \boldsymbol{f}\left(z_{\ell-1}\right)\right) \\
&= \boldsymbol{E}\ \ \boldsymbol{Q}\left(z, \boldsymbol{f}\left(z_{\ell-1}\right)\right)
\end{aligned}
$$

The implication of this statement is that the subsampling procedure proposed is unbiased and in expectation gives us more accurate estimates of the true error of a classifier trained with $\ell - \boldsymbol{p}$ samples as the number of subsamples increase.

1b. The variance of the leave-p-out estimator is biased and optimistic.

Statement 2: The variance of the leave-p-out estimator is less than or equal to the variance of a classifier trained with $\ell - \boldsymbol{p}$ samples and tested on an independent sample, so the variance of the leave-p-out procedure is not necessarily unbiased. This procedure is not necessarily unbiased and in general will be optimistic:

$$
\boldsymbol{V}\ \boldsymbol{L}_p\left(z_1,...,z_\ell\right) \le \boldsymbol{V}\ \boldsymbol{Q}\left(z, \boldsymbol{f}\left(z_{\ell-p}\right)\right)
$$

which means that the expected variance when the classifier is trained with $\ell - \boldsymbol{p}$ samples is greater than or equal to the variance of the procedure $\boldsymbol{L}_p$.

Proof:

Again we will prove the leave-one-out case and the leave-p-out case is a straightforward extension.

The variance of training sets of size $\ell - 1$ is

$$\boldsymbol{V}\ \boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1})) = \boldsymbol{E}\left[\boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1}))^2 - [\boldsymbol{E}\ \boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1}))]^2\right].$$

The variance of the estimator

$$\boldsymbol{V}\ \frac{1}{\ell}\boldsymbol{L}_1(z_1, ..., z_\ell) = \boldsymbol{V}\ \left[\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right]$$

where $\boldsymbol{t}_i$ is whether an error is made or not when the $\boldsymbol{i}^{\text{th}}$ point is left out,

$$\boldsymbol{Q}(z_i, \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_\ell)).$$

We can rewrite this as

$$\boldsymbol{V}\ \left[\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right] = \boldsymbol{E}\left(\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right)^2 - \left(\boldsymbol{E}\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right)^2$$

from statement 1 we know

$$\boldsymbol{E}\ \frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i = \boldsymbol{E}\ \boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1})).$$

So

$$\boldsymbol{V}\ \left[\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right] = \boldsymbol{E}\left(\frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell}\boldsymbol{t}_i\right)^2 - [\boldsymbol{E}\ \boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1}))]^2.$$

One can show that

$$E\left(\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{t}_i\right)^2 \le E\left[\boldsymbol{Q}(z, f(z_{\ell-1}))^2\right].$$

One can write

$$E\left(\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{t}_i\right)^2 = E\left(\frac{1}{\ell^2}\sum_{i=1}^{\ell}\boldsymbol{t}_i^2 + \frac{1}{\ell^2}\sum_{i\neq j}^{\ell}\boldsymbol{t}_i\boldsymbol{t}_j\right)$$

if the random variables $\boldsymbol{t}_i, \boldsymbol{t}_j$ are identical and independent then the above equation can be rewritten

$$E\left(\frac{1}{\ell^2}\sum_{i=1}^{\ell}\boldsymbol{t}_i^2 + \frac{1}{\ell^2}\sum_{i\neq j}^{\ell}\boldsymbol{t}_i\boldsymbol{t}_j\right) = E\left(\frac{\boldsymbol{t}^2}{\ell} + \frac{\ell^2-\ell}{\ell^2}\boldsymbol{t}^2\right) = E\ \left[\boldsymbol{Q}(z, f(z_{\ell-1}))\right]^2,$$

however $\boldsymbol{t}_1,...,\boldsymbol{t}_\ell$ are not independent and are correlated so

$$E\left(\sum_{i\neq j}^{\ell}\boldsymbol{t}_i\boldsymbol{t}_j\right) \le E\left(\frac{\ell^2-\ell}{\ell^2}\boldsymbol{t}^2\right).$$

The implication of this statement is that the variance of sub-sampling procedure proposed is biased and does not give an accurate estimate of the variance of a classifier trained with $\ell - \boldsymbol{p}$ samples and in general the variance of the sub-sampling procedure will be smaller.

1c. Quantiles of the leave-p-out estimator are unbiased.

Statement 3: Quantiles of the leave-p-out estimator estimated give an accurate estimate of quantiles of a classifier trained with $\ell - \boldsymbol{p}$ samples and tested on an independent sample.

Proof:

Again we will prove the leave-one-out case and the leave-p-out case is a straightforward extension.

The cumulative distribution function of the random variable $\boldsymbol{t} = \boldsymbol{Q}(z, \boldsymbol{f}(z_{\ell-1}))$ is $\boldsymbol{P}_{\boldsymbol{D}_{\ell-1}}\{\boldsymbol{t} < \xi\}$. The cumulative distribution function of the random variable $\boldsymbol{r} = \frac{1}{\ell}\boldsymbol{L}_{\ell-1}(z_1, ...z_\ell)$ is $\boldsymbol{P}_{\boldsymbol{D}_\ell}\{\boldsymbol{r} < \xi\}$. If we show that these distribution functions are equal then the quantiles of the leave-p-out estimator is unbiased. The distribution function of the random variable $\boldsymbol{t}$ is

$$\boldsymbol{P}(\boldsymbol{t} < \xi) = \int_{-\infty}^{\xi} \boldsymbol{t}\ \boldsymbol{p}(\boldsymbol{t})d\boldsymbol{t} = \int_{-\infty}^{\infty} \boldsymbol{t}\ \boldsymbol{\theta}(\xi - \boldsymbol{t})\ \boldsymbol{p}(\boldsymbol{t})d\boldsymbol{t}\,.$$

The distribution function for the random variable $\boldsymbol{r} = \frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell} \boldsymbol{r}_i$ can be written as follows by a similar sequence of transformations as used in the proof of statement 1

$$\boldsymbol{P}(\boldsymbol{r} < \xi) = \frac{1}{\ell}\sum_{\boldsymbol{i}=1}^{\ell} \int_{-\infty}^{\infty} \boldsymbol{r}_i\ \boldsymbol{\theta}(\xi - \boldsymbol{r}_i)\ \boldsymbol{p}(\boldsymbol{r}_i)d\boldsymbol{r}_i$$

which is the same as $\boldsymbol{P}(\boldsymbol{t} < \xi)$.

We have now shown that the cumulative distribution function of the error measured of the leave-p-out procedure is equivalent to the cumulative distribution of a classifier trained on $\ell - \boldsymbol{p}$ samples.

However, we do not have this distribution when we run the leave-p-out procedure

We have a sequence of p error rates and we take the 25th and 75th quantiles of the empirical distribution function constructed from the sequence. We can use the Kolmogorov-Smirnov or Smirnov distributions to show that the empirical quantiles values are close to those for the true underlying distribution. For ℓ large enough ($\ell > 20$) the Kolmogorov-Smirnov distribution gives us

$$P\left\{\sqrt{\ell}\sup_{x}\left|F(x)-F_{\ell}(x)\right|<\varepsilon\right\}\approx 1-2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^{2}\varepsilon^{2}}$$

where $F(x)$ is the distribution function of the error rates, $F_{\ell}(x)$ is the empirical distribution function constructed from a sequence of ℓ error rates. We can use this result to state that with probability $1-\delta$ the difference between the estimate of a quantile and the true quantile value will be bounded. For the case where $\ell \le 20$ instead of using the Kolmogorov-Smirnov distribution one can use tabulated values of the Smirnov distribution.

♦

The implication of this statement is that the sub-sampling procedure proposed gives us more accurate estimates of the quantiles of the true error of a classifier trained with $\ell - p$ samples as the number of sub-samples increase.

Appendix 2: Motivation of the inverse power law for the error rate as a function of training set size

2a. A motivation from the statistical mechanics approach to learning

In this appendix we will describe results derived within the statistical mechanics (SM) approach to generalization [Engel 2001, Watkin 1993] which strongly motivates the use of equation (1). In this approach the *average* generalization error can be calculated as a function of n. In order to derive analytic expressions, the data is assumed to consist of randomly constructed and uncorrelated input patterns. This assumption is unrealistic for practical datasets, of course, but we can assume that the functional relation derived between $e(n)$ and n largely holds for real-life data. In the SM approach a teacher (the rule to be discovered) and student (the learner) are used, with the extent of correlation between teacher and student quantifying generalization ability. To be more specific let us consider a simple perceptron [Rosenblatt 1962] rule with a decision function

$$y = sign(w \cdot z)$$

where z is an input vector, w is the weight vector for the perceptron (which *weight* the relevance of particular inputs or attributes), and $y = \pm 1$ is the output. Suppose the weight vector for the teacher perceptron is t and the weight vector for the student perceptron is w then the number of generalization errors made by the student perceptron on a set of p new samples will be

$$\sum_{i=1}^{p} \theta\left[-(\boldsymbol{t}\cdot z_i)(\boldsymbol{w}\cdot z_i)\right]$$

where $\theta(\boldsymbol{x}) = 0$ if $\boldsymbol{x} < 0$ and 1 otherwise. This general approach leads to a set of equations for determining the generalization error via a replica approach [Hertz 1991] involving integrations in the weight space $\boldsymbol{w}$. The generalization error is given as a function of the ratio $\beta = \boldsymbol{n}/\boldsymbol{m}$ (where $\boldsymbol{m}$ is the number of attributes). However, with $\boldsymbol{m}$ fixed we can assume the same functional dependence on $\boldsymbol{n}$ as for β. From this analysis we find that the generalization error depends on the algorithm used and generally assumes a power law.

It can be argued [Dietrich et al. 1999] that a Support Vector Machine with linear kernel, used in our numerical simulations, has the same generalization error dependence as the optimal perceptron [Opper et al. 1990]. We have solved the system of equations in [Opper et al. 1990] in the low β limit and find a very close fit to equation (1). With few samples and a large number of measured attributes the low β limit is most appropriate when considering microarray data. However, some further insights can also be gained by considering the high β limit where the dependence of generalization error on β (or equivalently $\boldsymbol{n}$) can be extracted explicitly. Thus for the optimal perceptron the generalization error scales as $.50\boldsymbol{n}^{-1}$ [Opper et al. 95]. Similarly for other rules this scaling can be extracted explicitly [Engel 2001]. For example, for the Bayes optimal classifier (derived from the Bayes Point or center of mass of version space - the space

of all hypotheses consistent with the data) the generalization error scales as $.44\boldsymbol{n}^{-1}$ [Opper et al. 91]. For the Adaline learning rule the error scales as $.24\boldsymbol{n}^{-1/2}$, and for the Hebb rule as $.40\boldsymbol{n}^{-1/2}$ (see [Watkin 1993] for a review). The dependence on $\boldsymbol{n}$ is thus approximately $\boldsymbol{n}^{-\alpha}$ with α near 1 for the more efficient rules such as the optimal perceptron and Bayes optimal classifier. The SM approach to generalization has also been used to quantify the effects of input noise, output noise and noise affecting the parameters in the model (e.g. the weights $\boldsymbol{w}$). Thus, for example, white noise added to examples in the training set appears as an additive constant term to the generalization error (justifying the $\boldsymbol{b}$ term in equation (1)). In summary, then, this approach strongly motivates use of $\boldsymbol{e}(\boldsymbol{n}) = \boldsymbol{a}\boldsymbol{n}^{-\alpha} + \boldsymbol{b}$ for modeling the generalization error.

2b. A motivation from an approximation theory point of view

Another justification for a power law for regression or classification comes from approximation theory [Niyogi et al. 96]. In the approximation theory framework the classification functions come from some restricted function class $\boldsymbol{f} \in \boldsymbol{H}$ and the optimal classification function $\boldsymbol{f}_T$ is a more complicated function that is not included in the function class $\boldsymbol{H}$. For a wide variety of algorithms the distance between the optimal function in the function class $\boldsymbol{f}_o \in \boldsymbol{H}$ and $\boldsymbol{f}_T$ is characterized as

$$\boldsymbol{d}(\boldsymbol{f}_o, \boldsymbol{f}_T) = \boldsymbol{O}(\boldsymbol{n}^{-\alpha}),$$

where $\alpha > 0$. Worst case analyses measure the distance between the two functions as the value of the point of greatest deviation between these functions.

For function classes used in most algorithms the worst case analysis yields $\alpha = .5$. In general an empirical result should have quicker convergence since the worst case assumptions need not be made. When the loss function

$V(\cdot,\cdot)$ is smooth then the difference in error measured using the loss function between the functions f_o and f_T is $V(f_o(x), y) - V(f_T(x), y) = O(n^{-\alpha})$ for all $x \in X$, and $y \in Y$. By a smooth loss function we mean loss functions that are ℓ_P with $p \geq 1$ or Lipschitz over a bounded domain. Note that the classification loss,

$V(f(x), y) = \theta(-yf(x))$ is not Lipschitz and when the classifier outputs $f(x) = \pm 1$ the loss function is ℓ_0. However, for most algorithms the loss function optimized to set the parameters of the classifier is Lipschitz (for computational reasons the ℓ_0 loss is not used). For example in Support Vector Machines for classification the loss function is Lipschitz. For this reason this analysis is still appropriate.

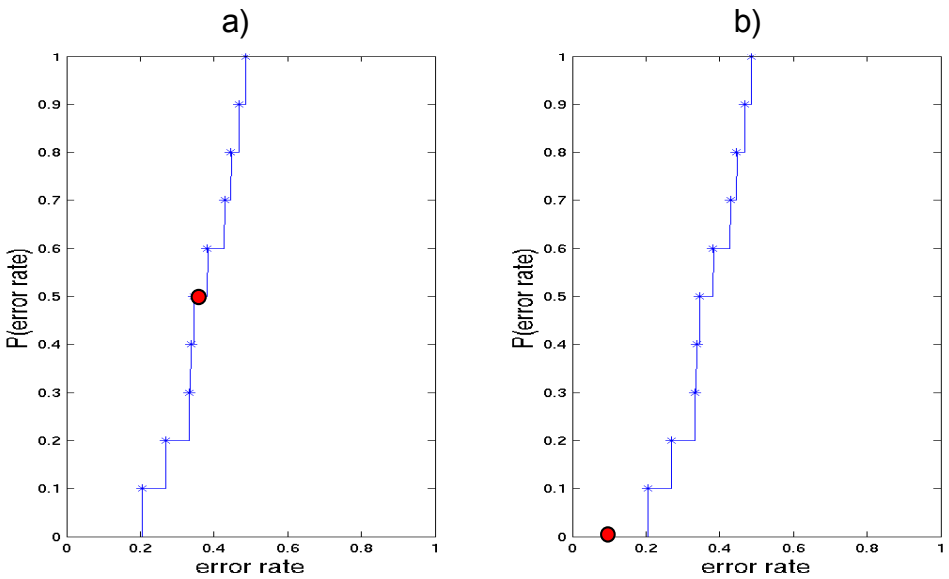


Figure 1. The statistical significance for the fictitious dataset example with a) 15 samples and b) 45 samples. The blue line is the empirical distribution function for the random classifiers and the red point is the average error rate for the classifier with randomization of labels.

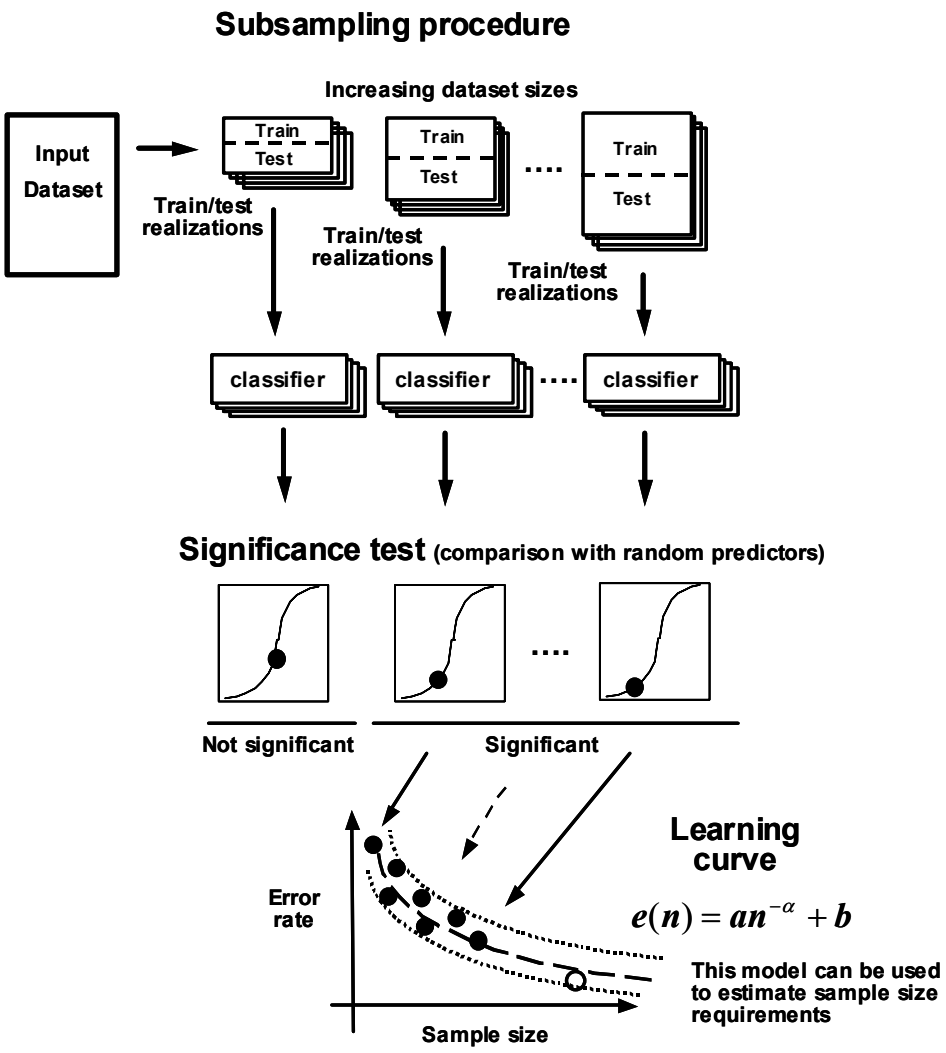


Figure 2. Dataset size estimation statistical methodology.

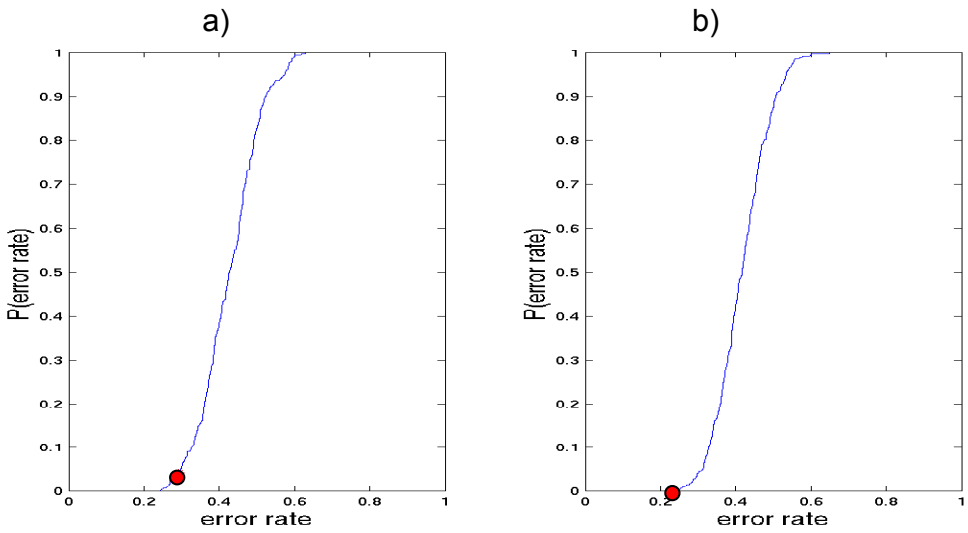


Figure 3. The statistical significance in the tumor vs. non-tumor classification for a) 15 samples and b) 30 samples. The blue line is the empirical distribution function for the random classifiers and the red circle is the average error rate for the actual classifier.

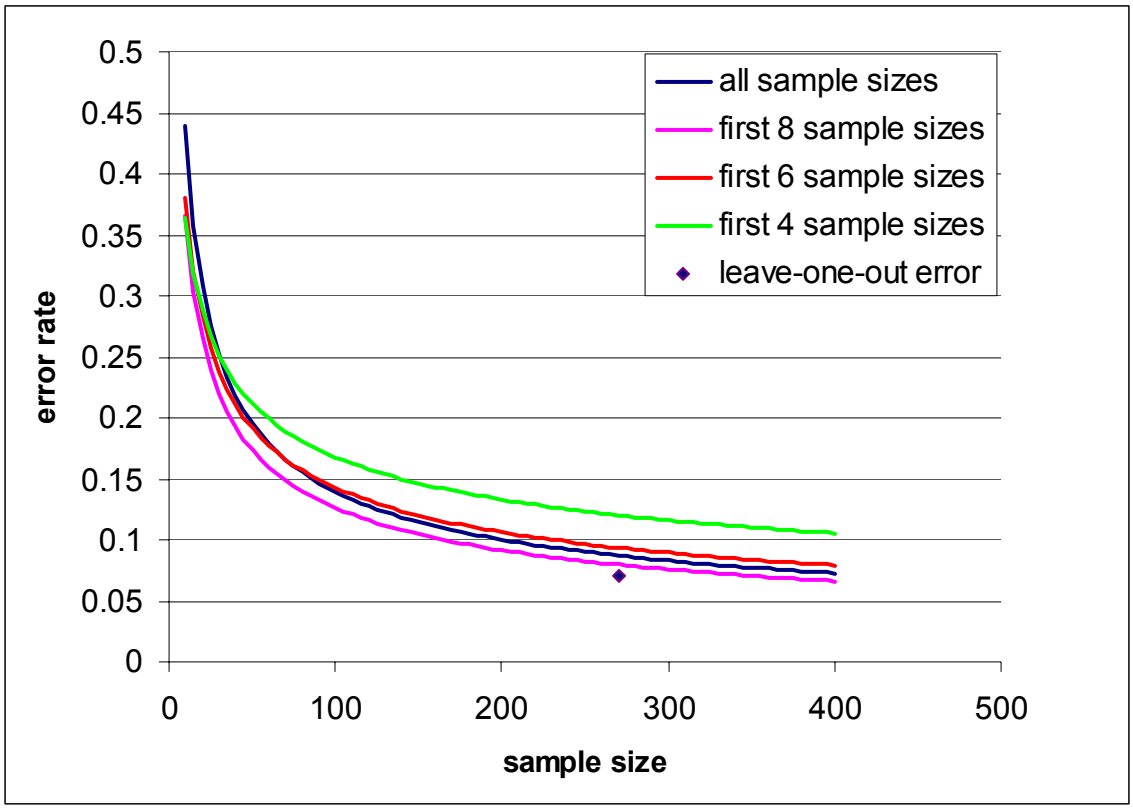
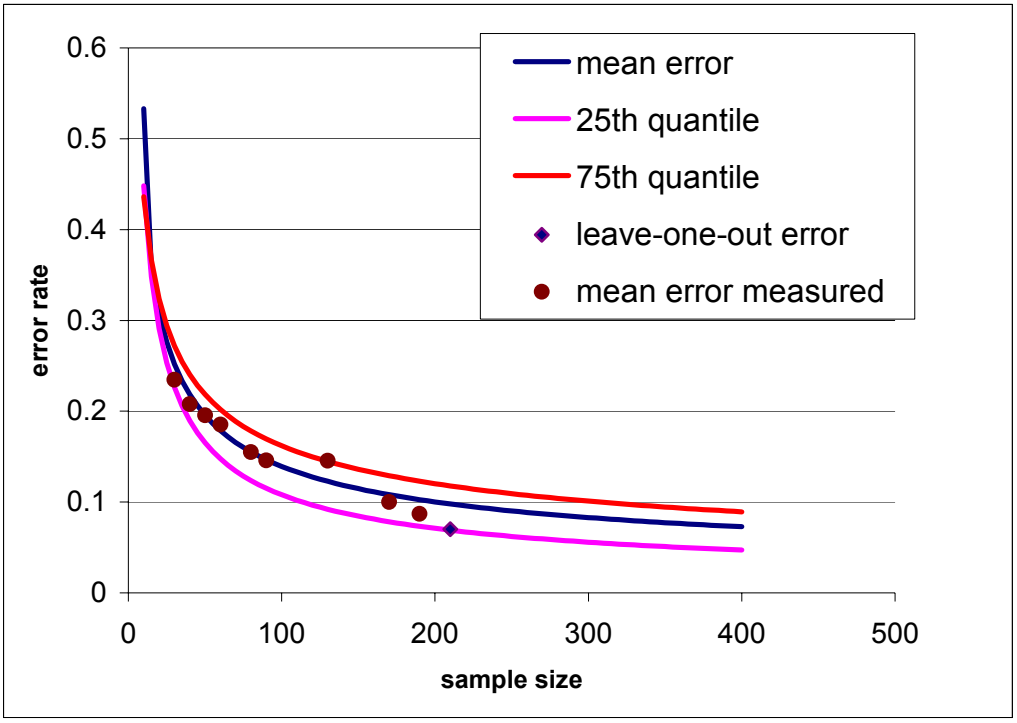


Figure 4. Error rate as a function of sample size. The blue diamond is the leave-one-out error using 279 samples. The green curve is the learning curve using the first 4 sample sizes. The magenta curve is the learning curve using the all sample sizes. The blue and the red curves (which basically overlap) are the learning curves using the first 6 (red) and first 8 (blue) sample sizes.



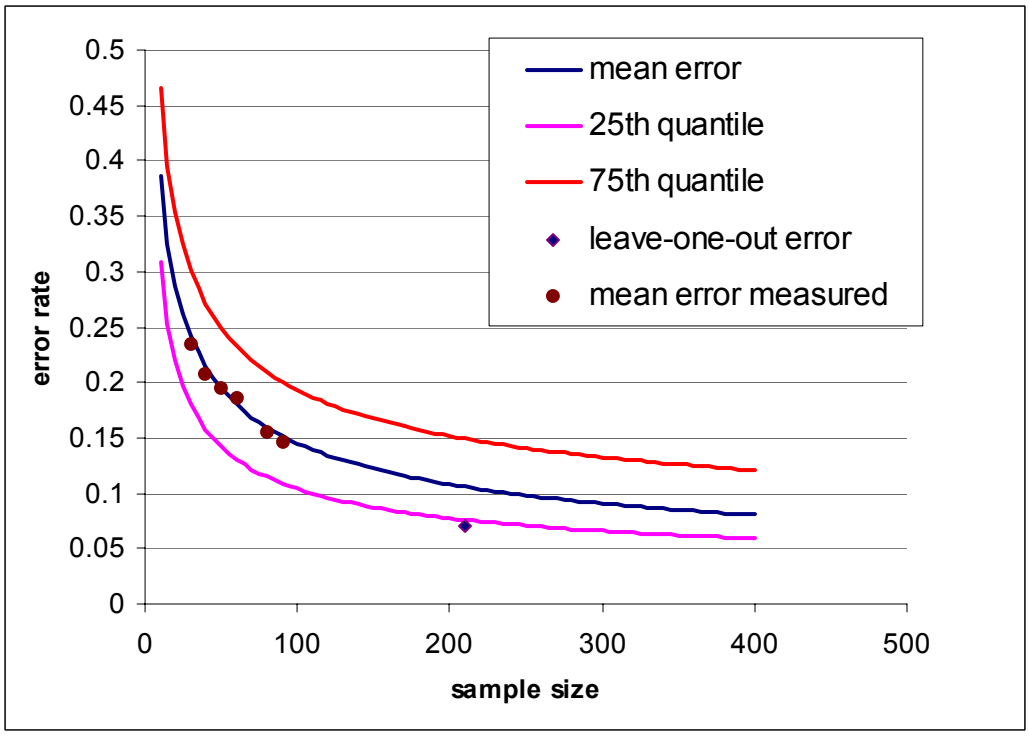


Figure 5. Learning curves in the tumor vs. non-tumor classification constructed using a) all samples sizes stated above and b) using the first 6 sample sizes stated above. The blue line is the learning curve for the mean error. The magenta line is

for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

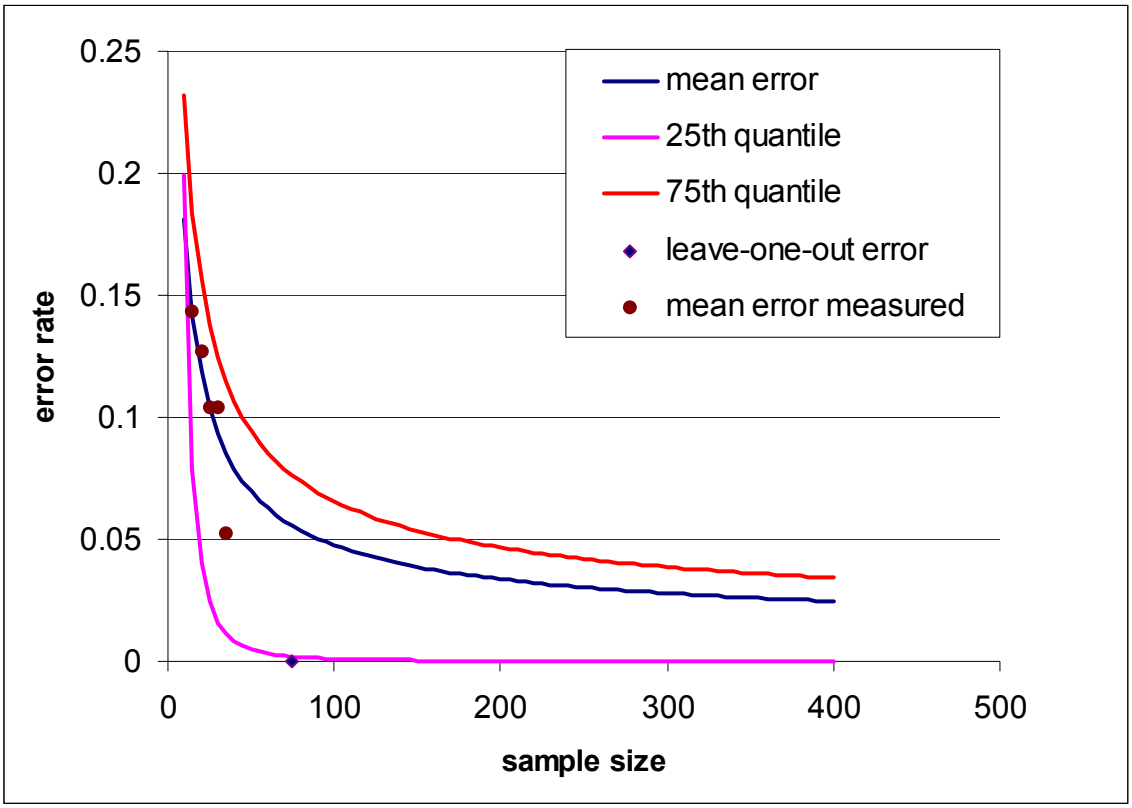


Figure 6. Learning curves in the AML vs. ALL classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

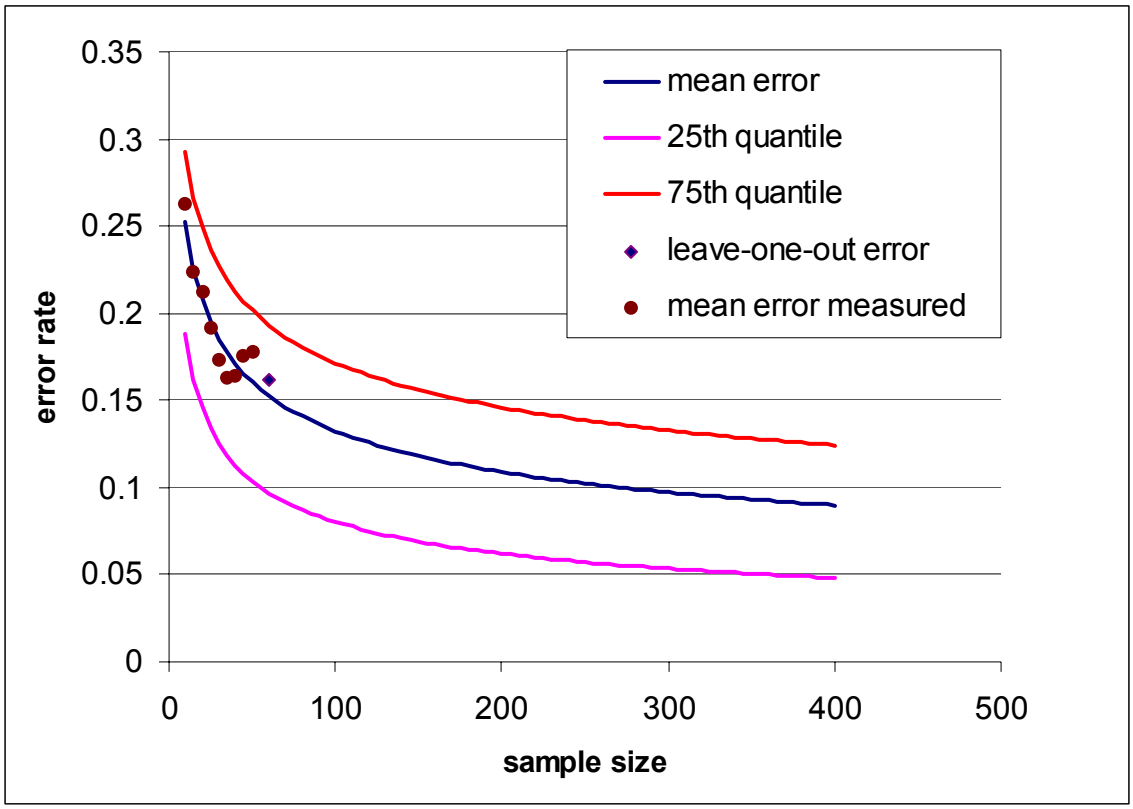


Figure 7. Learning curves in the malignant vs. normal colon tissue classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

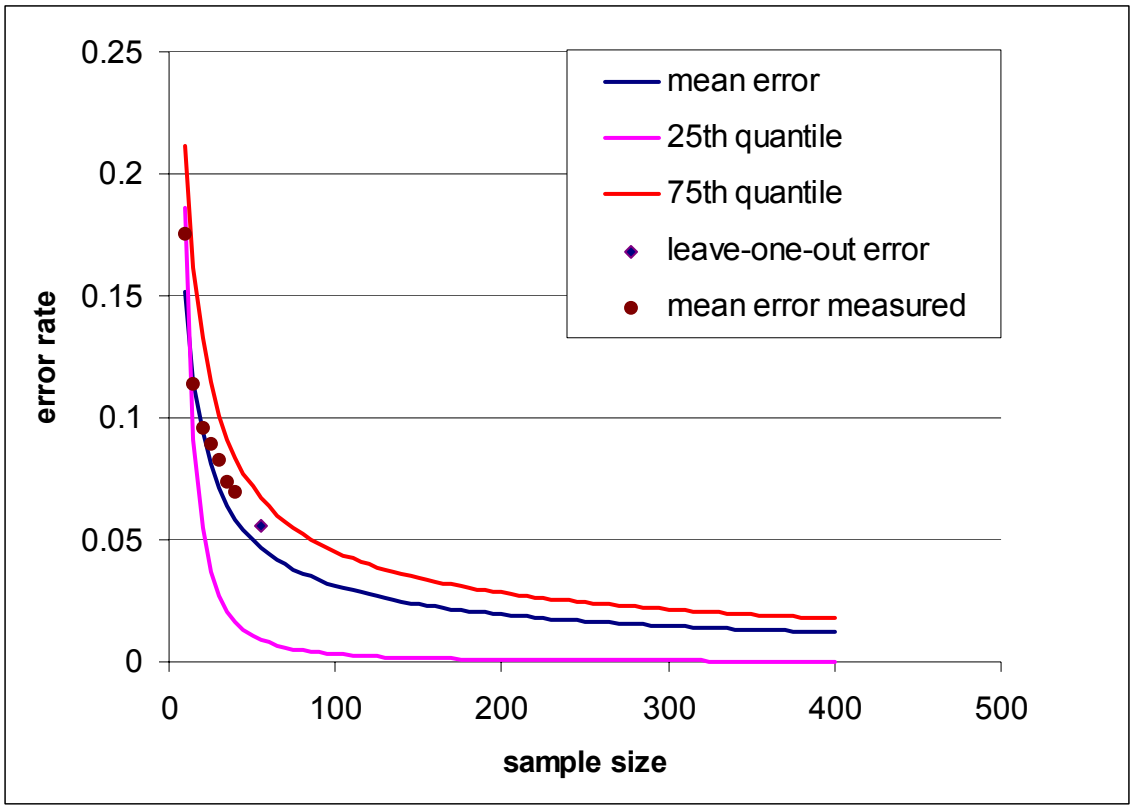


Figure 8. Learning curves in the cancerous vs. normal ovarian tissue classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

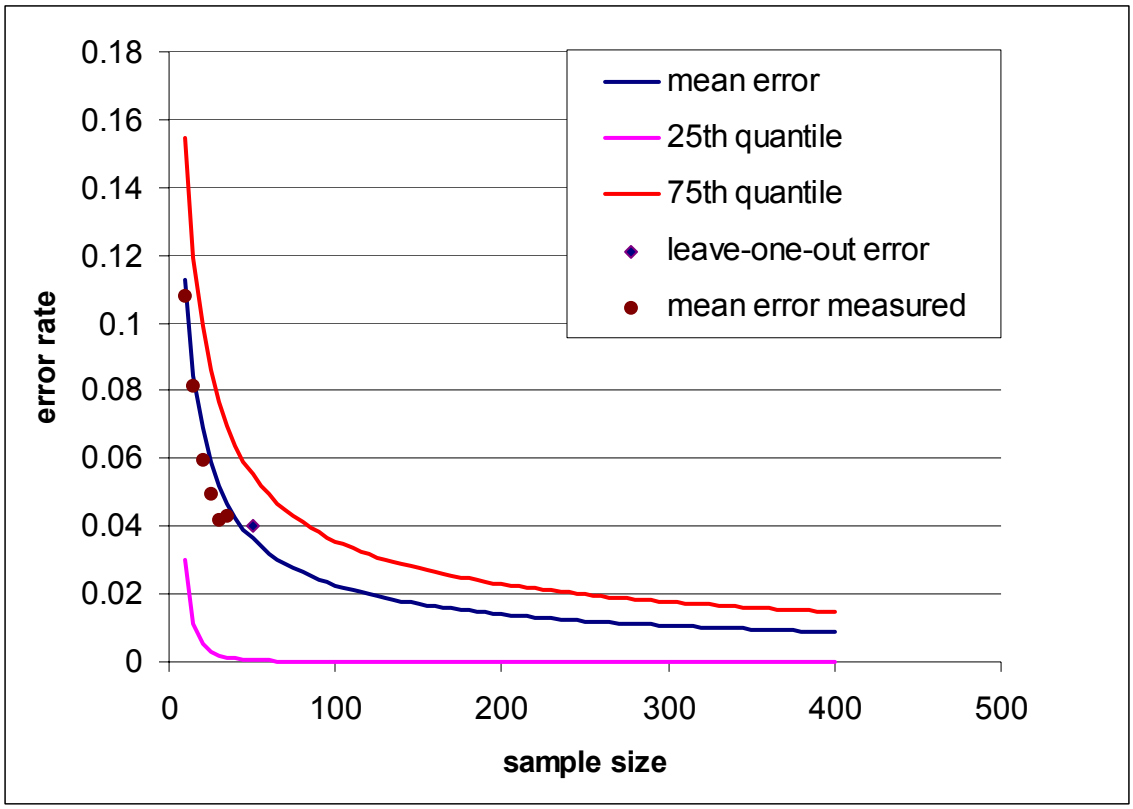


Figure 9. Learning curves in the diffuse large B-cell vs. follicular morphology classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The magenta star is the leave-one-out error and the blue points are the measured average error rates.

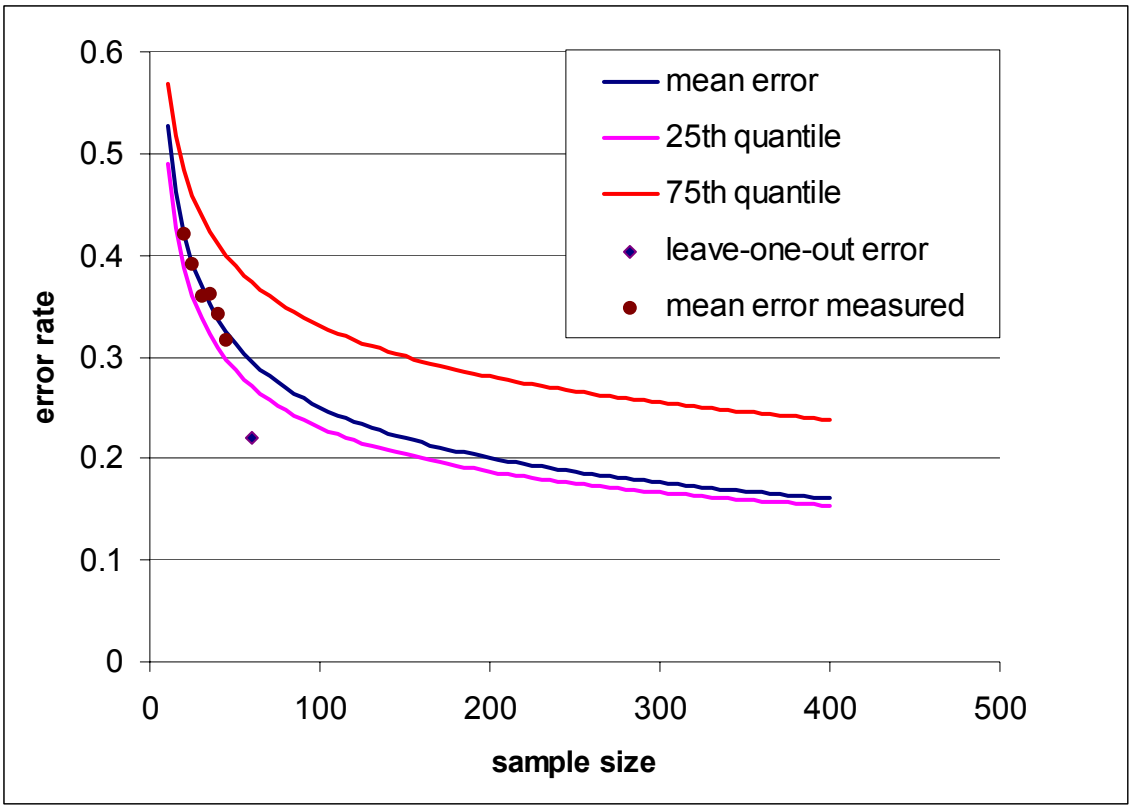


Figure 10. Learning curves in the brain tumor treatment outcome classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

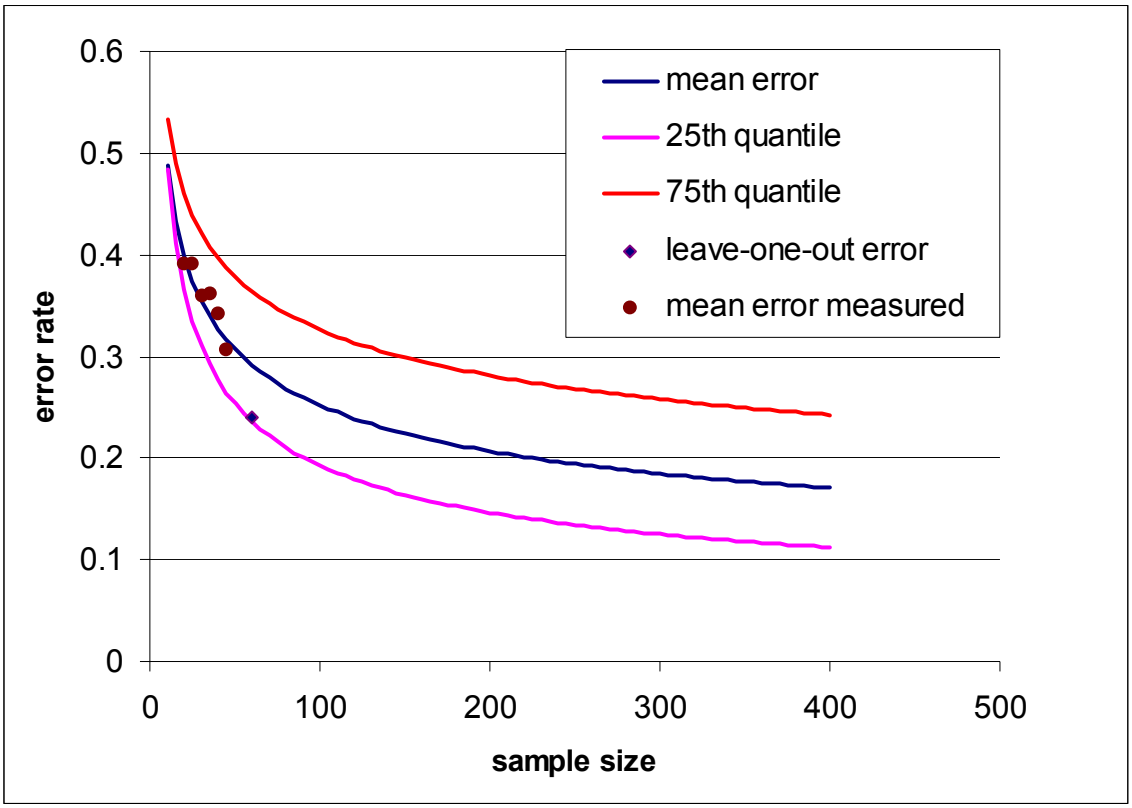


Figure 11. Learning curves in the lymphoma treatment outcome classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

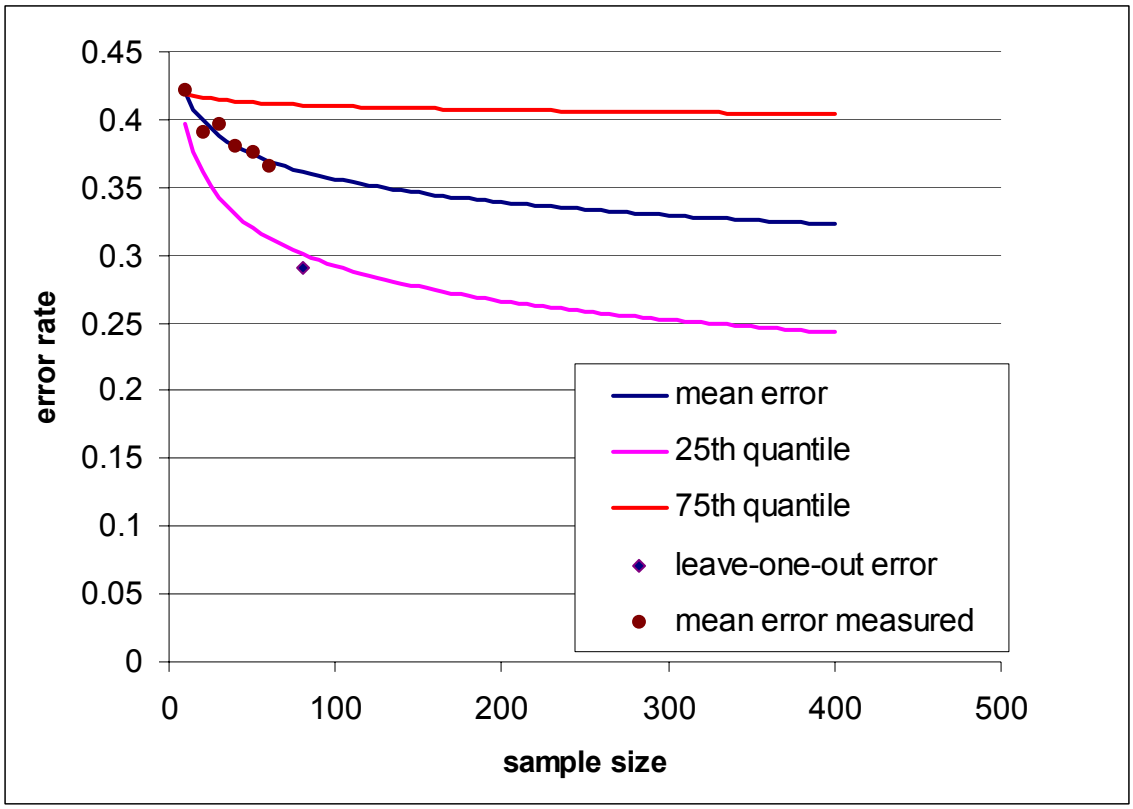


Figure 12. Learning curves in the breast cancer treatment outcome classification constructed using samples sizes stated above. The blue line is the learning curve for the mean error. The magenta line is for the 25th quantile. The red line is for the 75th quantile. The blue diamond is the leave-one-out error and the red points are the measured average error rates.

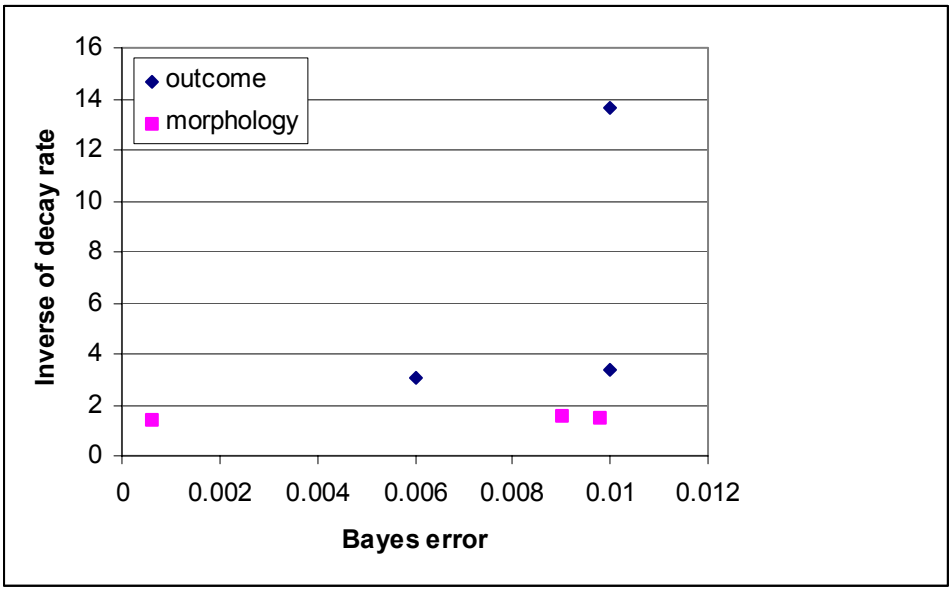


Figure 13. A plot of $1/\alpha$ vs. β for the eight datasets. The blue diamonds correspond to treatment outcome problems and the magenta squares correspond to morphology prediction problems.

Dataset/Problem	Predicted error rate at 400 samples			Learning Curve Parameters			
	25th quantile	Mean	75th quantile		a learning rate	α decay rate	b asymp. Bayes error
Multiple tissues Cancer vs. normal	4.7%	7.3%	8.9%	25th quantile mean 75th quantile	1.89 1.42 1.17	–0.63 –0.52 –0.43	0.003 0.010 0.000
Leukemia AML vs. ALL	0.6%	2.5%	3.4%	25th quantile mean 75th quantile	39.8 .771 .884	–2.3 –0.65 –0.6	0.006 0.009 0.010
Colon cancer Cancer vs. normal	4.8%	9.0%	12.4%	25th quantile mean 75th quantile	.439 .480 .500	–0.37 –0.28 –0.23	0.00 0.00 0.00
Ovarian cancer Cancer vs. normal	.03%	1.2%	1.8%	25th quantile mean 75th quantile	10.7 .736 .995	–1.76 –0.69 –0.67	0.00 0.00 0.00
Lymphoma DLBC vs. Follicular	0%	0.9%	1.5%	25th quantile mean 75th quantile	8.99 .57 .671	–2.48 –0.70 –0.64	0.00 0.001 0.00
Brain cancer Treatment outcome	14.5%	16.1%	23.8%	25th quantile mean 75th quantile	1.06 1.12 .980	–0.37 –0.33 –0.24	0.004 0.006 0.00
Lymphoma Treatment outcome	11.2%	17%	24.3%	25th quantile mean 75th quantile	1.23 .943 .872	–0.41 –0.30 –0.21	0.008 0.01 0.00
Breast cancer Treatment outcome	24.3%	32.3%	40.4%	25th quantile mean 75th quantile	.532 .485 .429	–0.14 –0.07 –0.01	0.01 0.01 0.00

Table 1. Summary of learning curve parameters and extrapolated error rates.

Dataset/Problem	Dataset size	Size to achieve classifier's statistical significance (p-val < 0.05)	Learning curve fitted with significant data	Maximum training set size used in fitting learning curve	Classification error			Conclusions
					Actual	Learning Curve (extrapolated)	% Error	
Multiple tissues Cancer vs. normal	280	15	yes	210	7.0%	8.6%	2.6%	Dataset size is large enough for extrapolation and to achieve a low error rate ~0.10. Even lower error rates <0.07 are achievable with >400 samples.
Leukemia AML vs. ALL	73	10	yes	35	0.0%	5.6%	5.6%	Dataset size is large enough for extrapolation and to achieve a low error rate ~0.05. Additional samples may lower error to ~0.
Colon cancer Cancer vs. normal	62	10	yes	50	16.3%	15.2%	1.1%	Dataset size is large enough for extrapolation and to achieve an error rate of ~0.16. Lower error rates <0.09) are achievable with >400 samples.
Ovarian cancer Cancer vs. normal	54	10	yes	40	5.6%	4.8%	0.8%	Dataset size is large enough for extrapolation and to achieve a low error rate ~0.05). Additional samples may lower error rate to ~0.
Lymphoma DLBC vs. Follicular	53	5	yes	40	4%	4.7%	0.7%	Dataset size is large enough for extrapolation and to achieve a low error rate ~0.035). Additional samples may lower error rate to ~0.
Brain cancer Treatment outcome	60	45	no	40	22%	29.6%	7.6%	Dataset is not large enough for extrapolation. Fitting of learning curve is inaccurate but suggests error rate could be <0.14 with > 400 samples.
Lymphoma Treatment outcome	58	50	no	40	23%	29.5%	6.5%	Dataset is not large enough for extrapolation. Fitting of learning curve is inaccurate but suggests error rate could be <0.17 with > 400 samples.
Breast cancer Treatment outcome	78	65	no	70	30%	36.3%	6.3%	Dataset is not large enough for extrapolation. Fitting of learning curve is inaccurate but suggests error rate could be <0.32 with > 400 samples.

Table 2. Summary of results for datasets included in the study.