

Optimally splitting cases for training and testing high dimensional classifiers

Kevin K. Dobbin, University of Georgia

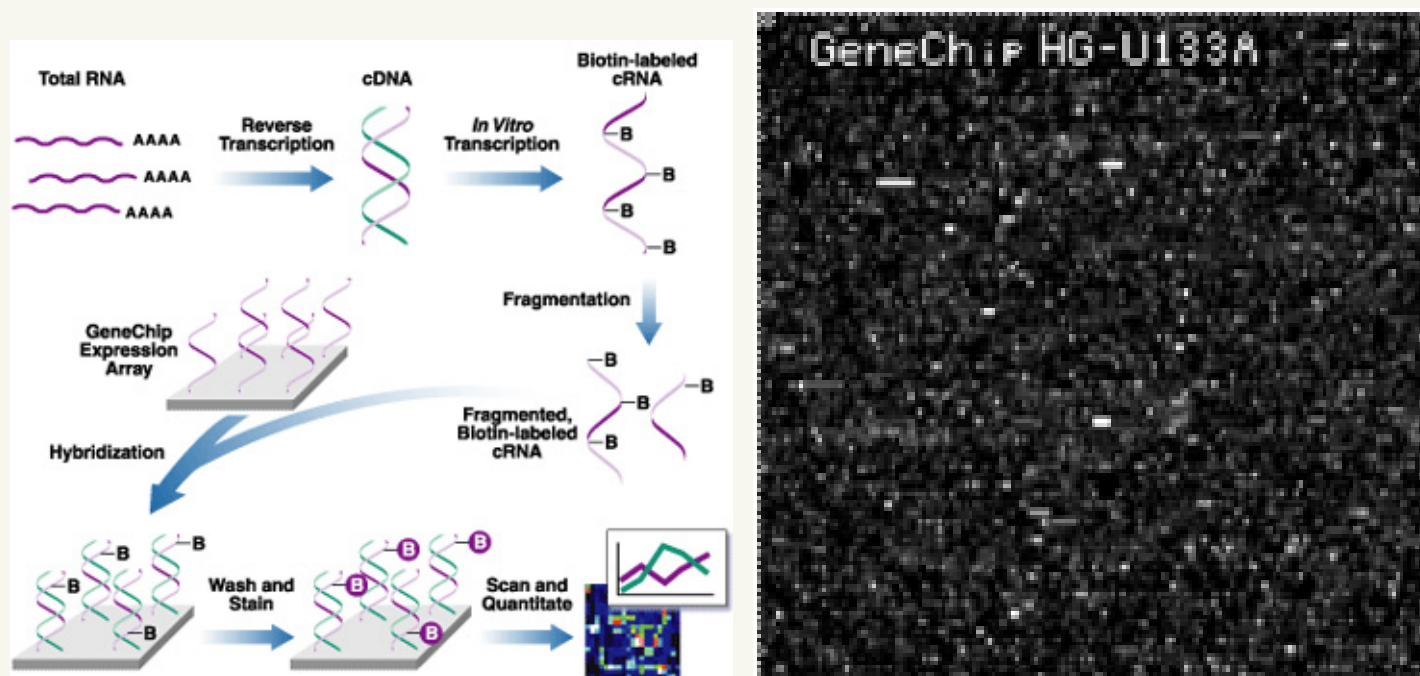
Richard Simon, National Cancer Institute

This research was partially supported by GCC

Classifiers in cancer

- Cancer classifiers can be clinically useful.
 - Classifiers can identify which patients will benefit from a targeted agent (e.g., EGFR inhibitor).
 - Classifiers can identify patients who will/will not benefit from adjuvant chemotherapy.
- New assays, like microarrays, measure levels of tens of thousands of RNA at once.
- This rich information may be useful in developing better classifiers.
 - The ultimate classifier need not require the microarray platform, which can be difficult to use in clinical settings.

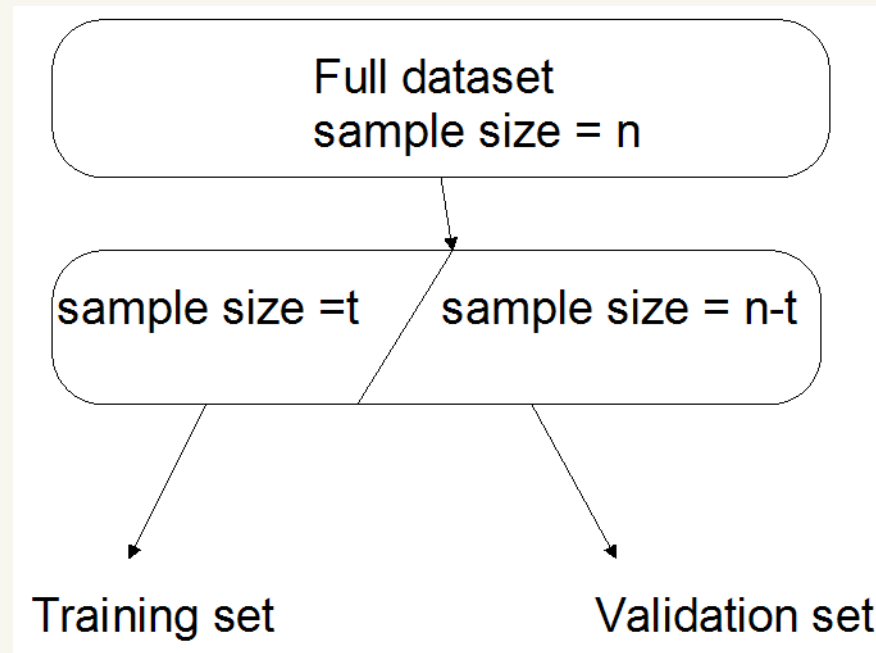
Microarrays



- Microarrays measure the mRNA expression levels of tens of thousands of genes in a specimen of cells.
- The results are images (right) where the brightness of a square represents gene expression level.

Motivation

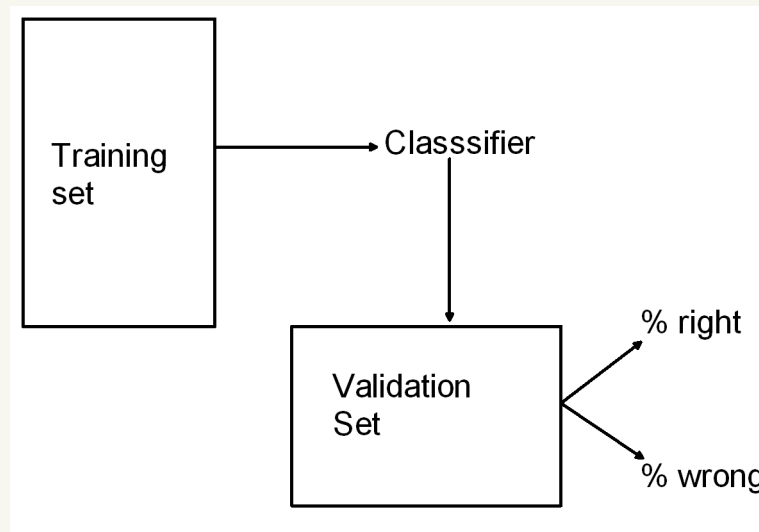
- In microarray studies, patient samples are typically split into a *training set* and a *validation set*.
- Why split the dataset? To ESTIMATE ACCURACY.
- SCIENTIFIC QUESTION: What proportion of the samples should be used for the training set?



Split sample approach

- A classifier is developed completely on the training set ONLY.
- The classifier is then applied to every sample in the validation set.
- The accuracy on the validation set is the estimate of the classifier's accuracy (on future samples).

Sometimes, after this process, a final classifier is produced using the whole dataset.



What's been done before?

- Nothing directly on topic we could find.
- Related research on sample size and classification in high dimensions include
 - Mukherjee et al. (2003) J Comput Biol: 10:119-42
 - Dobbin and Simon (2007) Biostatistics: 8: 101-7.
 - Dobbin et al. (2008) Clin Cancer Res: 14: 108-14.
- What do people do?
 - Use rules-of-thumb.
 - Typical, 50% or 67% are assigned to the training set.

Our approach

HOW DO WE DEFINE THE BEST SPLIT?

- The sample is split in order to estimate the accuracy of the predictor.
- The best split is the one that minimizes the mean squared error (MSE) of the accuracy estimate.

Decomposition of the MSE

- We prove that the mean squared error (MSE) can be decomposed into 3 intuitive parts.

Squared bias term: From using $t < n$ to estimate accuracy.

Accuracy variance term: From developing the predictor.

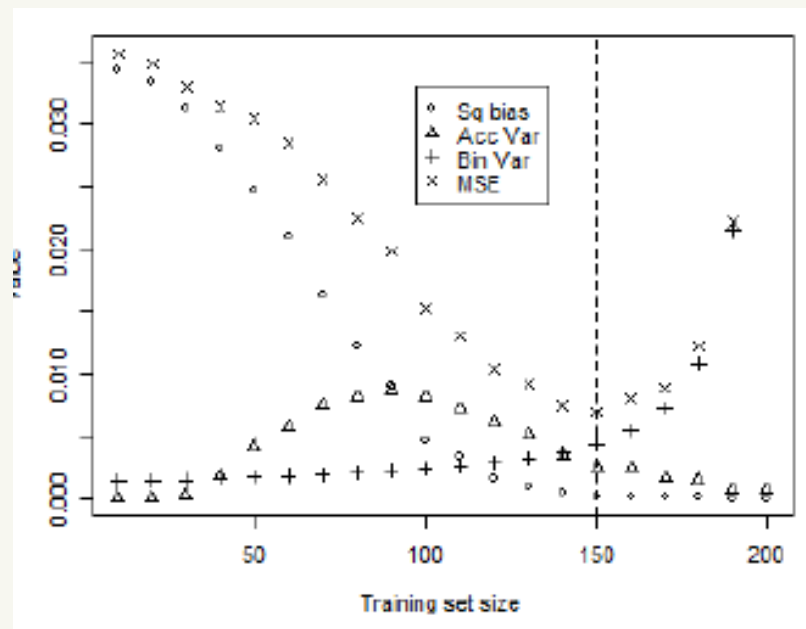
Binomial variance term: From applying the predictor to the validation set.

Our estimation methods

- A novel MSE estimation method is developed.
 - Step 1: The nonparametric bootstrap is used to estimate estimate variance terms
 - Step 2: A resampling method and nonparametric learning curve regression is used to estimate the squared bias term.
- The novel method is used to estimate MSE over a range of training set sizes.

Results

- A wide range of simulations were carried out.
 - Squared bias term decreases as training set size increases.
 - Accuracy variance term remains relatively small.
 - Binomial variance term increases as the training set size increases



The tradeoff: Example with $n=200$ samples.

Example application

We applied the method to 4 real microarray datasets and the results supported the simulation studies.

One Example dataset

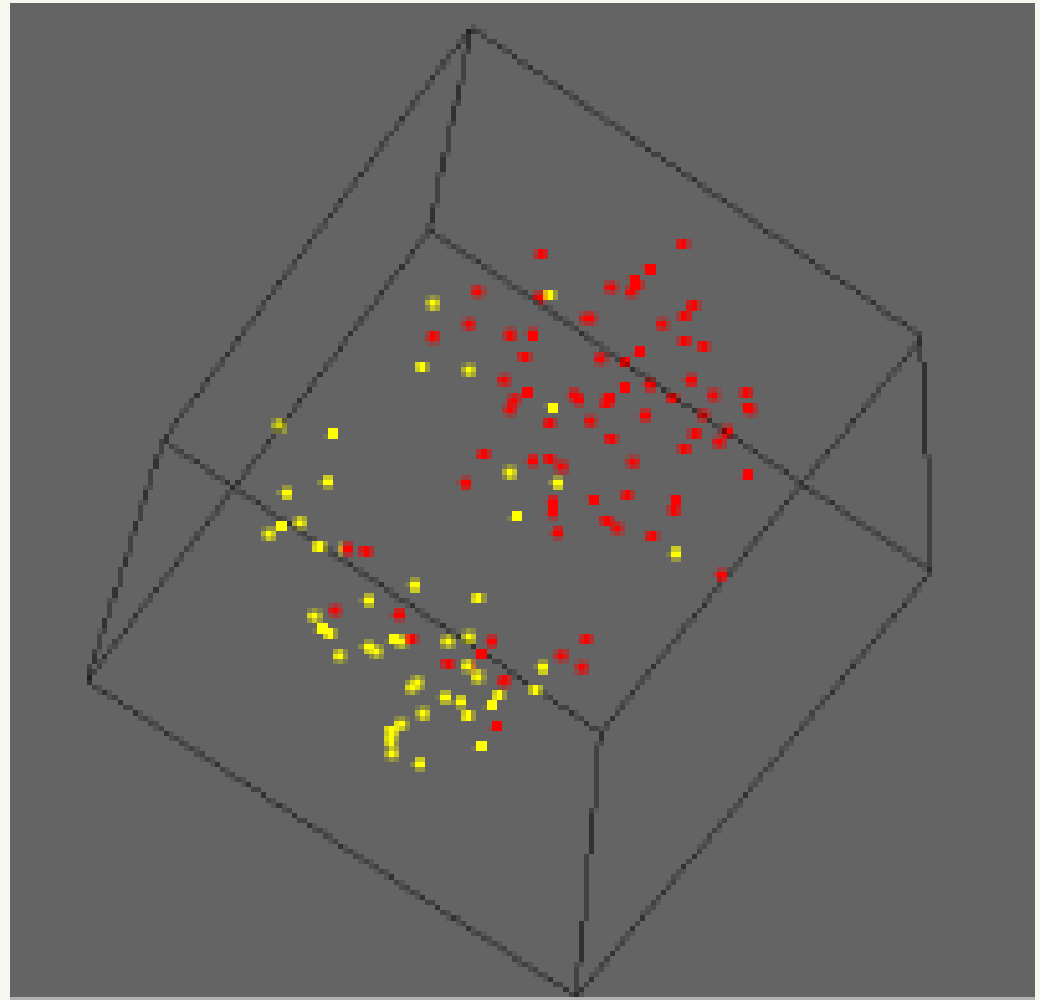
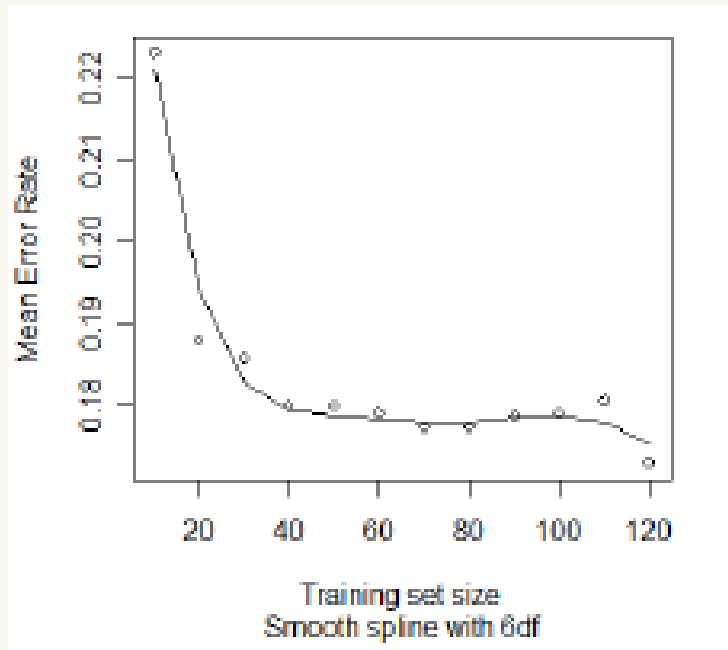
Dataset	Rosenwald et al., 2002
Classes	GCBCCL vs. non-GCBCCL
Total samples	240
Prevalence of GCBCCL	52%
<u>Estimates from our method</u>	
Best accuracy possible ¹	96%
Optimal allocation to training	63%

¹on this dataset

Summary

- Developed a novel, robust, nonparametric method for identifying the optimal method of splitting a sample into a training set and a validation set.
 - This method should be used whenever feasible.
- Extensive simulation analysis showed that the 67% allocation rule-of-thumb is more robust than 50%.
 - If application of our nonparametric approach is not feasible, we suggest devoting 2/3rds to the training set for microarray studies.
- Our nonparametric method also useful for identifying anomalies in datasets:
 - We identified one such anomaly in the real datasets we examined.

The anomalous dataset



Left: Learning curve estimate.
Right: Multidimensional scaling plot.