Methods & Myths in Prognostic & Diagnostic Prediction with Microarray & Proteomic Data

Richard Simon
National Cancer Institute
http://linus.nci.nih.gov/brb

Common Types of Objectives

- Class Comparison
  - Identify genes/proteins differentially expressed among predefined classes such as diagnostic or prognostic groups.
- Class Prediction
  - Develop multi-feature predictor of class for a sample
- Class Discovery
  - Discover clusters among specimens or among features

Statistical Methods Appropriate for Class Comparison Differ from Those Appropriate for Class Prediction

- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy.
- Demonstrating goodness of fit of a model to the data used to develop it is not a demonstration of predictive accuracy.
- Most statistical methods were developed for inference, not prediction.
- Most statistical methods for were not developed for p>>n settings

Components of Class Prediction

- Feature selection
  - Which genes or proteins will be included in the model
- Select model type
  - E.g. DLDA, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model

Feature Selection

- Key component of supervised analysis
- Usually features are selected that are univariately differentially expressed among the classes at a nominal significance level α (e.g. 0.01)
  - The α level is selected to control the number of features in the model, not to control the false discovery rate
  - The accuracy of the significance test used for feature selection is not of major importance as identifying differentially expressed genes is not the ultimate objective
  - For survival prediction, the features with significant univariate Cox PH regression coefficients

Feature Selection

- Small subset of features which together give most accurate predictions
  - Step-up regression
  - Greedy pairs
  - Combinatorial optimization algorithms
    - Genetic algorithms
- Some published complex methods for selecting combinations of features do not appear to have been properly evaluated

Linear Classifiers for Two Classes

$$l(\underline{x}) = \sum_{i \varepsilon F} w_i x_i$$

$\underline{x}$ = vector of features

F = features included in model

w_i = weight for i'th feature

decision boundary $l(\underline{x})$ > or < d

Linear Classifiers for Two Classes

- Fisher linear discriminant analysis (weights based on assumed multivariate normal distribution of expression vector in each class with common covariance matrix)
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
  - Naïve Bayes estimator
- Compound covariate predictor and Golub's weighted voting method are variants of DLDA

Linear Classifiers for Two Classes

- Compound covariate predictor

$$w_i \propto \frac{\overline{x}_i^{(1)} - \overline{x}_i^{(2)}}{\hat{\sigma}_i}$$

Instead of for DLDA

$$w_i \propto \frac{\overline{x}_i^{(1)} - \overline{x}_i^{(2)}}{\hat{\sigma}_i^2}$$

Linear Classifiers for Two Classes

- Support vector machines with inner product kernel are linear classifiers with weights determined to minimize errors
- Perceptrons with principal components as input are linear classifiers with no well defined criterion for defining weights

Advantages of Simple Linear Classifiers

- Do not over-fit data
  - Incorporate influence of multiple features without attempting to select the best small subset of variables
  - Do not attempt to model the multivariate interactions among the predictors and outcome

When p>>n

- For the linear model, many weight vectors w can always be found that give zero classification errors for the training data.
- Why consider more complex models?
- The number of parameters for this simple model is generally too large relative to the number of specimens to achieve accurate prediction for future samples if we select a single model by minimizing training errors

Myth

- That complex classification algorithms such as neural networks perform better than simpler methods for class prediction.

- Artificial intelligence sells to naïve journal reviewers and readers.
- Comparative studies indicate that simpler methods that avoid overfitting work better for p>>n problems.

- Fitting complex functions to training data results in unstable classifiers unless there is a huge training dataset
- For unstable classifiers, the test sample error rate is generally much less than the generalization error rate

Model Stability Can Be Improved By

- Restriction to models with fewer parameters
  - Complexity depends on number of parameters per candidate feature, not per selected feature
- Reducing number of candidate features
  - Principal components
  - Cluster averages
- Not minimizing training error
  - Equivalent to including penalty for complexity
- Aggregating models
- Use fitting criterion incorporating robustness to changes in data

- With unstable classifiers, we obtain both large bias and large variance in estimating the true classifier function
  - Large bias because there are many classifiers with zero training set errors that are far from the true classifier function
  - Large variance because the selected classifier varies substantially with small variations in the data

Evaluating a Classifier

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data.

Split-Sample Evaluation

- Training-set
  - Used to select features, select model type, determine parameters and cut-off thresholds
- Test-set
  - Withheld until a *single* model is *fully* specified using the training-set.
  - Fully specified model is applied to the expression profiles in the test-set to predict class labels.
  - Number of errors is counted
  - Ideally test set data is from different centers than the training data and assayed at a different time

Split-Sample Evaluation

- Used for Rosenwald et al. study of prognosis in DLBL lymphoma.
  - 200 cases training-set
  - 100 cases test-set

Leave-one-out Cross Validation

- Omit sample 1
  - Develop multivariate classifier *from scratch* on training set with sample 1 omitted
  - Predict class for sample 1 and record whether prediction is correct

Leave-one-out Cross Validation

- Repeat analysis for training sets with each single sample omitted one at a time
- e = number of misclassifications determined by cross-validation
- Subdivide e for estimation of sensitivity and specificity

Myth

- Cross-validation of a model can occur after selecting the features to be used in the model

- Cross validation is only valid if the test set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed from scratch for each leave-one-out training set. This means that feature selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error is an estimate of the prediction error for model fit using specified algorithm to full dataset
- If you use cross-validation estimates of prediction error for a set of algorithms and select the algorithm with the smallest cv error estimate, you do not have a valid estimate of the prediction error for the selected model

Partial Cross-Validation of Random Data

- Generate data for p features and n cases identically distributed in two classes
  - No model should predict more accurately than the flip of a fair coin
- Using all the data select k<<p features that appear most differentially expressed between the two classes
- Cross validate the estimation of model parameters using the same k features for all LOOCV training sets
- The cross-validated estimate of prediction error will be 0 over 99% of the time.

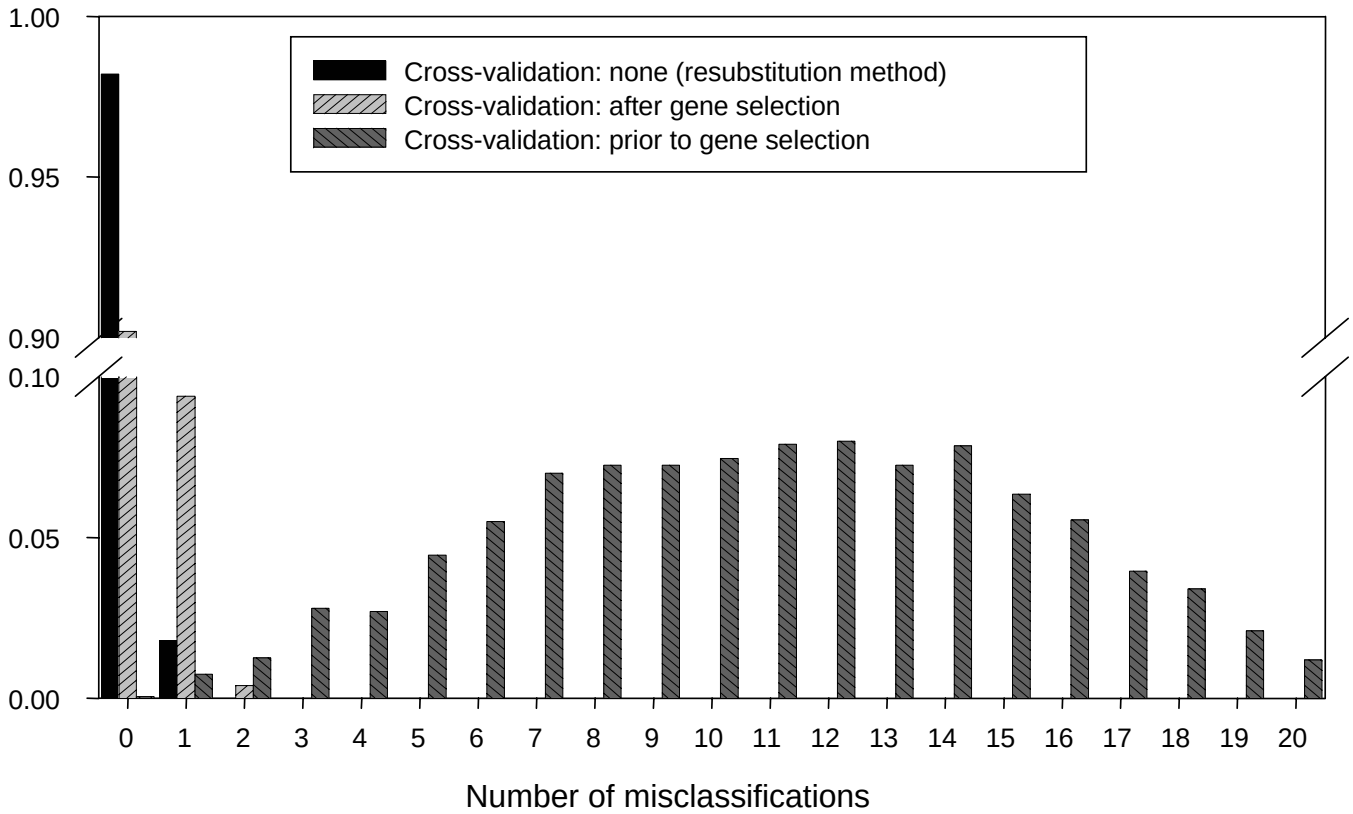
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

- 14 specimens (P_i is the expression profile for specimen i)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

Prediction Method

- Compound covariate prediction (*discussed later*)
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.


1.00
0.95
0.90
0.10
0.05
0.00
Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications
Proportion of simulated data sets

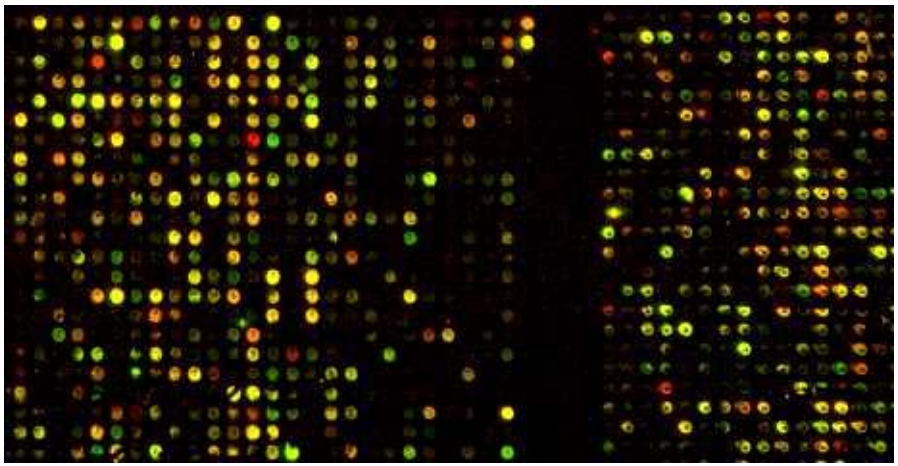
Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

- Randomly permute class labels and repeat the entire cross-validation
- Re-do for all (or 1000) random permutations of class labels
- Permutation p value is fraction of random permutations that gave as few misclassifications as e in the real data

Gene-Expression Profiles in Hereditary Breast Cancer

cDNA Microarrays

Parallel Gene Expression Analysis



- Breast tumors studied:
  - 7 *BRCA1+* tumors
  - 8 *BRCA2+* tumors
  - 7 sporadic tumors

- Log-ratios measurements of 3226 genes for each tumor after initial data filtering

Can we distinguish *BRCA1+* from *BRCA1–* cancers and *BRCA2+* from *BRCA2–* cancers based solely on their gene expression profiles?

BRCA1

α_g	# of significant genes	# of misclassified samples (m)	% of random permutations with *m* or fewer misclassifications
10^{-2}	182	3	0.4
10^{-3}	53	2	1.0
10^{-4}	9	1	0.2

BRCA2

α_g	# of significant genes	m = # of misclassified elements (misclassified samples)	% of random permutations with m or fewer misclassifications
10^{-2}	212	4 (s11900, s14486, s14572, s14324)	0.8
10^{-3}	49	3 (s11900, s14486, s14324)	2.2
10^{-4}	11	4 (s11900, s14486, s14616, s14324)	6.6

Classification of BRCA2 Germline Mutations

Classification Method	LOOCV Prediction Error
Compound Covariate Predictor	14%
Fisher LDA	36%
Diagonal LDA	14%
1-Nearest Neighbor	9%
3-Nearest Neighbor	23%
Support Vector Machine (linear kernel)	18%
Classification Tree	45%

Invalid Criticisms of Cross-Validation

- "You can always find a set of features that will provide perfect prediction for the training and test sets."
  - For complex models, there may be many sets of features that provide zero training errors.
  - A modeling strategy that either selects among those sets or aggregates among those models, will have a generalization error which will be validly estimated by cross-validation.

Sources of Bias in Estimation of Error Rates

- Confounding by sample handling or assay effects
  - Cases collected and assayed at different times than controls
- Failure to incorporate important sources of future variability
  - Assay drift
- Change in distribution of unmodeled variables
  - In split sample validation, split samples by institution

BRB-ArrayTools

- Integrated software package using Excel-based user interface but state-of-the art analysis methods programmed in R, Java & Fortran
- Publicly available for non-commercial uses from BRB website:
- Amy Peng & R Simon

http://linus.nci.nih.gov/brb

Acknowledgements

- Michael Radmacher
- Sudhir Varma