Case Study: Van’t Veer breast cancer study study

Investigate whether tumor ability for metastasis is obtained later in development or inherent in the initial gene expression signature.

- Retrospective sampling of node-negative women: 44 non-recurrences within 5 years of surgery and 34 recurrences. Additionally, 19 test sample (12 recur. and 7 non-recur)
- Want to demonstrate that gene expression profile is independently predictive of the recurrence.

Reference
Data

- **Array**: 2-color Agilent oligo arrays, ~24,000 genes
- **Pre-processing**: image quantification, background subtraction and normalization
- **Filtering**: at least 2-fold differential expression relative to the reference and p-value for being expressed < 0.01 in at least 5 samples. 4,348 such genes are remaining.
- **Imputation**: k-nearest neighbors imputation procedure (Troyanskaya et al):
  - For each gene with at least 1 missing values, find 5 genes most highly correlated with it
  - Replace missing value with the average of those 5 gene values in the samples with a missing value for a gene.
**Tumor Classification Using Gene Expression Data**

Three main types of statistical problems associated with tumor classification:

- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning – clustering)
- Classification of malignancies into known classes (supervised learning – discrimination)
- Identification of “marker” genes that characterize the different tumor classes (feature or variable selection).

**Basic principles of discrimination**

- Each object associated with a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and a feature vector (vector of predictor variables) of $G$ measurements: $X = (X_1, \ldots, X_G)$

Aim: predict $Y$ from $X$.

- Predefined Class $\{1,2,\ldots,K\}$
- Objects

\[Y = \text{Class Label} = 2\]

\[X = \text{Feature vector (colour, shape)}\]

\[X = \{\text{red, square}\}\]

\[Y = \ ?\]
Discrimination and Allocation

Learning Set
Data with known classes

Classification Technique

Classification rule

Data with unknown classes

Class Assignment

Prediction

Objects
Array

Feature vectors
Gene expression

Predefine classes
Clinical outcome

Bad prognosis
recurrence < 5yrs

Good Prognosis
recurrence > 5yrs

Good Prognosis
Matesis > 5

Learning set

Reference

new array

Classification rule
Classification Rule

- Classification procedure,
- Feature selection,
- Parameters [pre-determine, estimable],
- Distance measure,
- Aggregation methods

- One can think of the classification rule as a black box, some methods provides more insight into the box.
- Performance assessment needs to be looked at for all classification rule.

Why select features

- Lead to better classification performance by removing variables that are noise with respect to the outcome
- May provide useful insights into etiology of a disease
- Can eventually lead to the diagnostic tests (e.g., “breast cancer chip”)
Note how selecting genes based on their association with recurrence biases the results towards clear grouping.

Approaches to feature selection

- Methods fall into three basic category
  - Filter methods
  - Wrapper methods
  - Embedded methods

- The simplest and most frequently used methods are the filter methods.
Filter methods

\[ \mathbf{R}^p \rightarrow \text{Feature selection} \rightarrow \mathbf{R}^s \rightarrow \text{Classifier design} \]

\[ s \ll p \]

• Features are scored independently and the top s are used by the classifier

• Score: correlation, mutual information, t-statistic, F-statistic, p-value, tree importance statistic etc

Easy to interpret. Can provide some insight into the disease markers.

Classification rule

Maximum likelihood discriminant rule

• A maximum likelihood estimator (MLE) chooses the parameter value that makes the chance of the observations the highest.

• For known class conditional densities \( p_k(X) \), the maximum likelihood (ML) discriminant rule predicts the class of an observation \( X \) by

\[
C(X) = \arg \max_k p_k(X)
\]
Gaussian ML discriminant rules

- For multivariate Gaussian (normal) class densities $X|Y=k \sim N(\mu_k, \Sigma_k)$, the ML classifier is
  \[
  C(X) = \arg\min_k \{ (X - \mu_k) \Sigma_k^{-1} (X - \mu_k)^\top + \log|\Sigma_k| \}
  \]
- In general, this is a quadratic rule (Quadratic discriminant analysis, or QDA)
- In practice, population mean vectors $\mu_k$ and covariance matrices $\Sigma_k$ are estimated by corresponding sample quantities

ML discriminant rules - special cases

[DLDA] Diagonal linear discriminant analysis
class densities have the same diagonal covariance matrix $\nabla = \text{diag}(s_1^2, \ldots, s_p^2)$

[DQDA] Diagonal quadratic discriminant analysis
class densities have different diagonal covariance matrix $\nabla_k = \text{diag}(s_{1k}^2, \ldots, s_{pk}^2)$

Note. Weighted gene voting of Golub et al. (1999) is a minor variant of DLDA for two classes (different variance calculation).
Nearest neighbor classification

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation).

- k-nearest neighbor rule (Fix and Hodges (1951)) classifies an observation $X$ as follows:
  - find the $k$ observations in the learning set closest to $X$
  - predict the class of $X$ by majority vote, i.e., choose the class that is most common among those $k$ observations.

- The number of neighbors $k$ can be chosen by cross-validation (more on this later).
Classification tree

- Partition the feature space into a set of rectangles, then fit a simple model in each one.
- **Binary tree structured classifiers** are constructed by repeated splits of subsets (nodes) of the measurement space $X$ into two descendant subsets (starting with $X$ itself).
- Each terminal subset is assigned a class label; the resulting partition of $X$ corresponds to the classifier.

Gene 1
$M_{i1} < -0.67$

Gene 2
$M_{i2} > 0.18$

0 = no recurrence
1 = recurrence
Performance assessment

- Any classification rule needs to be evaluated for its performance on the future samples. It is almost never the case in microarray studies that a large independent population-based collection of samples is available at the time of initial classifier-building phase.

- One needs to estimate future performance based on what is available: often the same set that is used to build the classifier.

- Assessing performance of the classifier based on
  - Cross-validation.
  - Test set
  - Independent testing on future dataset
Performance assessment (III)

- Common practice to do feature selection using the learning, then CV only for model building and classification.

- However, usually features are unknown and the intended inference includes feature selection. Then, CV estimates as above tend to be downward biased.

- Features (variables) should be selected only from the learning set used to build the model (and not the entire set).

Summary of Case Study Analysis

- Classifier
- Feature-ranking statistic
- Performance assessment method
- Optimization parameters

- K-NN
- F-statistic
- Leave-one-out cross-validation error rate
- Number of neighbors, Number of Features

(k = 1,3 and 
p = 10,30,50,100,150, 200,250,300 )

Apply classifier with the smallest leave-one-out cross validation error rate to the test set.
Results

Best classification is with k=3 and p=200
23/78 or 29% misclassified

Confusion matrix:

<table>
<thead>
<tr>
<th>True</th>
<th>no rec</th>
<th>rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>no rec</td>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td>rec</td>
<td>12</td>
<td>22</td>
</tr>
</tbody>
</table>

Fisher exact test p-value for association between true and predicted class labels.
Green horizontal line is drawn at p=.01

Application to the test set

1. Identify 200 variables with highest F-statistic using entire training set.
2. Build k-NN classifier with k=3 neighbors using entire training set
3. Apply that classifier to the test set

Confusion matrix:

<table>
<thead>
<tr>
<th>True</th>
<th>no rec</th>
<th>rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>no rec</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>rec</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

6/19 or 32% of samples misclassified
Microarrays: Case Studies and Advanced Analysis

Samples are clustered within class. Genes are clustered based on training set and the resulting order is used for the test set.

Learning set

Classification Rule

Feature selection. Correlation with class labels, very similar to t-test.

Using cross validation to select 70 genes

Case studies

Reference 1
Retrospective study

Reference 2
Cohort study

Reference 3
Prospective trials.
Aug 2003
Clinical trials http://www.agendia.com/

Agendia (formed by researchers from the Netherlands Cancer Institute)

Has started in Oct, 2003

1) 5000 subjects [Health Council of the Netherlands]
2) 5000 subjects New York based Avon Foundation.

Custom arrays are made by Agilent including 70 genes + 1000 controls


Results’ Gene expression profile is a more powerful predictor than standard systems based on clinical and histologic criteria.
van de Vuver’s breast data (NEJM, 2002)

- Consequent cohort of 295 additional breast cancer patients, mix of node-negative and node-positive samples.
- Want to use the predictor that was developed to identify patients at risk for metastasis.
- The predicted class was significantly associated with time to recurrence in the multivariate cox-proportional model.

Validation

- Apply predictor developed on the training set to the new samples to obtain binary label: recurrence or no recurrence
- Include that label as a covariate into cox-proportional model for survival along with known clinico-pathological risk factors
- Assess whether recurrence labels are independent predictors of time to recurrence (i.e. whether the covariate is significant in the multivariate model)

The answer is yes.
Software

- Various packages in R
- PAM
- MvE
  http://www.tigr.org/software/tm4/mev.html

Number of proprietary tools exists
# Acknowledgements

**SFGH**
- Agnes Paquet
- David Erle
- Andrea Barczac
- UCSF Sandler Genomics Core Facility.

**UCB**
- Terry Speed
- Jean Yang
- Sandrine Dudoit

**UCSF /CBMB**
- Ajay Jain
- Mark Segal
- UCSF Cancer Center Array Core
- Jain Lab