Enriched methods for Megavariate Data

Javier Cabrera,
Rutgers University

Dhammika Amaratunga,
Janssen Phramaceutical

Current/Former Students: Lan Yi,
Zhenya Cherkas, YS Lee,

Hasselt University, December 14, 2012

Outline
1. Introduction: Megavariate Data from Genomics experiments
2. Enriched methods for supervised and unsupervised classification
3. Enriched PCA, and Biplots
4. Enriched Random Forest
5. Enriching other classification methods

(1) Megavariate Data

Many variables with few observations: $X$ is a $G \times n$ matrix, $G >> n$
Examples: Microarray data, Chip-seq, SNP and other Sequence data, Proteomics data, Mass spec, Imaging, Google data, Census Data,...
6 samples x 50000 Variables + Responses
50000 Genes x 6 samples + Responses

You can also think of the data as:
- 50000 scatter-plots or boxplots vs same Y
- 50000 regression (linear or logistic) vs the same Y

Some Properties
(i) The variables may all reflect one type of measurement
(ii) Mild to strong correlation structure among subgroups of variables.
(iii) Preprocessing may induce correlations among samples.
(iv) Data are not likely to be normally distributed. Sample outliers are difficult to detect.
(v) There is often subsidiary information available.
(vi) High throughput experiments that produce Megavariate data are exploratory in nature.

The central issue on modeling these data is how to avoid overfitting!

Examples
Experiments that produce Megavariate Data in Genomics:
(i) Microarray data
(ii) HT-qPCR
(iii) Chip-seq, RNA-seq
(iv) SNP’s, Genome Wide Association Studies (GWAS)
(v) Protein arrays
(vi) LC-MS: Liquid chromatography Mass spectrometry
    Protein or metabolites or other small molecules
(vii) Copy number variation
(viii) Molecular imaging: CAT, PET, etc

Other sources: Clinical Data? Internet Data?

Signal Structure in Microarray Data

Microarray experiments:

As dimension increases => Spurious signal may ultimately dominate all signals.

Spurious Signal => Likely to be made up of small incremental random signals

Solutions:
- Enriched methods (Enriched PCA …)
- Thresholding methods (GLMNET …)
Case study: Sialin gene(1429116_a) Knockout Experiment

Experiment: Compare the gene expression profiles of 6 KO mice vs 6 WT mice using a microarray with 45101 genes.

Questions: Are there differences between gene expressions for WT and KO at 6h? 18 days? Are the differences happening in the same cellular processes?

Gene expression set (matrix)

<table>
<thead>
<tr>
<th>Sample</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
<th>G6</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data: Expression measures for 6 genes in N samples:

Preprocess: normalize and log transform

Enriched methods for supervised and unsupervised classification

Objectives: Response Prediction (class or continuous), Variable Selection, Clustering...

Enriched Methods: Assigning weights to variables, not to observations
- Amaranthus, Cabrer, Kortun (2007). Microarray learning with ABC: Biostatistics
- Amaranthus, Cabrera, Li (2008). Enriched random forests, Bioinformatics
- Amaranthus, Cabrera, Cherkas, Li (2011). Enriched Ensembles, AMS series
- Philippe Höldermann thesi and Yuhenys Cherkas thesi (2010).

Issues:
- Modeling and Variable Selection are the obvious ideas but in practice are difficult because of the high level of spurious information.
- Our idea is to apply weights to variables that correct for spurious information contained by each variable.
- 1. Construct FDR corrected weights
- 2. Model the data using these weights: Directly or using Ensembles.

Enriched Methods

Assign a weight to each gene based on FDR

Weight functions: \( w_i(t) = -\log(t) \quad w_i(t) = 1/t \)

Suppose that is the p-value of some gene wise model or some other measure of how predictive a gene is.

In unsupervised analysis, we have used gene variance as a measure of how predictive a gene is.

To avoid over-fitting we need to remove any signal that could be explained by chance. To do this we correct p-values for multiplicity p-value FDR corrections (sometimes called q-values).

Benjamini & Hochberg, Yekutieli & Benjamini, Storey & Tibshirani

Enriched Methods: Enriched Principal Components Analysis

Steps for obtaining FDR based Weights:
- Perform Statistical analysis of individual variables: Obtain p-values by testing statistically significant difference across the groups (t, F, ...)
- Assume a null distribution for p-values, for example uniform p-value distribution \( p_i \sim \text{Uniform}(0,1) \), then \( p_{(0)} \sim \text{Beta}(i, 1-i) \) (order statistics)

FDR corrected p-values: \( q_{(0)} = p_{(0)}/F \) and make \( q_{(0)} \) monotone on (i)

FDR corrected weights: \( W_i = 1/q_i \) or \( W_i = -\log(q_i) \)

Suppose that a is a tuning constant and that \( p_{(a)} \) is the a-percentile of distribution of the i-th order statistic.

We fix \( a = 0.05 \) (Fisher) and calculate the weights.

Simple version: \( q_i = p_i/(1/G) \) and \( W_i = 1/q_i \) or \( W_i = -\log(q_i) \)
Summary of important genes from Enriched PCA:
day 0 gene expression data from Slc17A5 experiment

**Important Genes with Annotation**

<table>
<thead>
<tr>
<th>Vertical direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1429116_a</td>
<td>SLC gene that was knocked out in treatment group.</td>
</tr>
<tr>
<td>1435559_at</td>
<td>Myo6, Growth.</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone.</td>
</tr>
<tr>
<td>1454905_at</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizontal direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1417210_at</td>
<td>Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked EF-2,g,p, Sps, Thy</td>
</tr>
<tr>
<td>1426438_at</td>
<td>Box polypeptide 3, Y-linked 8030469F12Rik, D1Pas1-rs1, Dby</td>
</tr>
<tr>
<td>1427262_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
<tr>
<td>1436936_s_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone.</td>
</tr>
</tbody>
</table>

Enriched PCA Analysis:
Gene expression data from Slc17A5 experiment
Day 0 predicts day 18

<table>
<thead>
<tr>
<th>Vertical direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1429116_a</td>
<td>SLC gene that was knocked out in treatment group.</td>
</tr>
<tr>
<td>1435559_at</td>
<td>Myo6, Growth.</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone.</td>
</tr>
<tr>
<td>1454905_at</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizontal direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1417210_at</td>
<td>Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked EF-2,g,p, Sps, Thy</td>
</tr>
<tr>
<td>1426438_at</td>
<td>Box polypeptide 3, Y-linked 8030469F12Rik, D1Pas1-rs1, Dby</td>
</tr>
<tr>
<td>1427262_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
<tr>
<td>1436936_s_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
</tbody>
</table>

Summary of important genes from Enriched PCA:
day 18 gene expression data from Slc17A5 experiment

**Important Genes with Annotation**

<table>
<thead>
<tr>
<th>Vertical direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1429116_a</td>
<td>SLC gene that was knocked out in treatment group.</td>
</tr>
<tr>
<td>1435559_at</td>
<td>Myo6, Growth.</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone.</td>
</tr>
<tr>
<td>1454905_at</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizontal direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1417210_at</td>
<td>Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked EF-2,g,p, Sps, Thy</td>
</tr>
<tr>
<td>1426438_at</td>
<td>Box polypeptide 3, Y-linked 8030469F12Rik, D1Pas1-rs1, Dby</td>
</tr>
<tr>
<td>1427262_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
<tr>
<td>1436936_s_at</td>
<td>Inactive X specific transcripts A430022811, A314753.</td>
</tr>
</tbody>
</table>

Enriched PCA Analysis:
Gene expression data from Slc17A5 experiment
Day 18 predicts day 0
General FDR Weights
Methods that can incorporate weights:
- PCA, LDA and methods related to the Covariance Matrix, PLS
- LASSO, Elastic Net, SVM, KNN, Random Forest, Trees

Methods that cannot incorporate weights and should use the
Ensemble approach:
- Linear regression and Linear models Generalized LM
- LASSO, Elastic Net, SVM, KNN

Classification: Enriched Classifiers: Random Forest (ERF)
1. Draw a bootstrap sample from the data. Call those not in the
bootstrap sample the "out-of-bag" data.
2. Grow a "random" tree, where at each node, the best split
is chosen among \( m \) randomly selected variables. The tree is
grown to maximum size and not pruned back.
3. Use the tree to predict out-of-bag data.
4. use the predictions on out-of-bag data to form majority votes.
5. Repeat 1-4 \( N \) times and collect an ensemble of \( N \) trees.
Prediction of test data is done by majority votes from
predictions from the ensemble of trees.

Incorporate Weights
2. Grow a "random" tree, where at each node, the best split is
chosen among \( m \) randomly selected variables according to
the weights \( \{W_i\} \). The tree is grown to maximum size and
not pruned back.

The Lasso
\[
\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)
\]
subject to: \( \sum_{j=1}^{p} |\beta_j| \leq s \)

Quadratic programming algorithm needed to solve for the
parameter estimates. Choose \( s \) via cross-validation.

\[
\hat{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)
\]

FDR corrected p-value weights can be incorporated to some methods
allow for feature weighting. For example:

LASSO:
\[
\hat{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} W_j |\beta_j| \right)
\]

\[
= \arg \min_{\beta} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} W_{ij} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)
\]

Support Vector Machine (SVM):
The shaded area represents the separation region.
The arrows indicate the location of the support vectors.

Enriched SVM:
Make the Std. Dev. of \( X \leq W \)
Use the standard SVM algorithm
without scaling.
Comparing Enriched with LASSO and SVM methods for classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction Error</th>
<th>Prediction Error</th>
<th>Prediction Error</th>
<th>Prediction Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched PCA</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Enrich Forest – log</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Random Forest (500)</td>
<td>41%-75%(5-9)</td>
<td>16%-50%(2-6)</td>
<td>0%-25%(0-3)</td>
<td></td>
</tr>
<tr>
<td>Random Forest(1000000)</td>
<td>41%(5)</td>
<td>16%(2)</td>
<td>8%</td>
<td></td>
</tr>
<tr>
<td>GLMnet</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>SVM(e1071)</td>
<td>50%(6)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>SVM(e1071)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>SVM(e1071) (log)</td>
<td>50%(6)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Proximal SVM</td>
<td>50%(6)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Penalized SVM (LASSO)</td>
<td>0%</td>
<td>50%(6)</td>
<td>50%(6)</td>
<td></td>
</tr>
<tr>
<td>Penalized SVM (SCAD)</td>
<td>0%</td>
<td>50%(6)</td>
<td>50%(6)</td>
<td></td>
</tr>
</tbody>
</table>

Comparing Enriched PCA with Elastic Net and SVM methods for classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Day 0 predicts Day 10</th>
<th>Day 18 predicts day 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched PCA</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>GLMnet</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>SVM(e1071)</td>
<td>50%(6)</td>
<td>50%(6)</td>
</tr>
<tr>
<td>SVM(e1071)</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Proximal SVM</td>
<td>50%(6)</td>
<td>50%(6)</td>
</tr>
<tr>
<td>Penalized SVM (LASSO)</td>
<td>0%</td>
<td>50%(6)</td>
</tr>
<tr>
<td>Penalized SVM (SCAD)</td>
<td>0%</td>
<td>50%(6)</td>
</tr>
</tbody>
</table>

Comparing Enriched PCA with Elastic Net and SVM methods for classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Day 0 -&gt; Day 10</th>
<th>Day 0 -&gt; Day 18</th>
<th>Day 10 -&gt; Day 0</th>
<th>Day 10 -&gt; Day 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched PCA</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>GLMnet</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Elastic Net (a=0.9-1)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Proximal SVM</td>
<td>50%(6)</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Penalized SVM (SCAD)</td>
<td>0%</td>
<td>50%(6)</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Enriched Elastic Net (glmnet) compared with Standard:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Brain</th>
<th>Prostate</th>
<th>Srbct</th>
<th>Lymphoma</th>
<th>Colon</th>
<th>Leukemia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>42 x 5,597</td>
<td>102 x 6,033</td>
<td>63 x 2,308</td>
<td>49 x 7,129</td>
<td>62 x 6,500</td>
<td>72 x 3,571</td>
</tr>
</tbody>
</table>

Poor performance of Random Forest

Data: y={0,1} P(0)=P(1)=0.5
Signal x: N(2*y,1)
Noise x: N(0,1)
n=50
RF was run with 1000 trees and 1000000 (no change)

<table>
<thead>
<tr>
<th>Signal 10 V</th>
<th>Noise 0 V</th>
<th>Signal 10 V</th>
<th>Noise 1000 V</th>
<th>Signal 10 V</th>
<th>Noise 5000 V</th>
<th>Signal 10 V</th>
<th>Noise 10000 V</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0%</td>
<td>2%</td>
<td>18%</td>
<td>28%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLMnet</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERF</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Important variables from Enriched Forest

Enriched Forest also gives a list of the most important variables. Here is the list for the day 0 data.

SLC gene
### Relationship between Enrichment and Partial Least Squares

**PLS:**
- First developed by Wold (1960’s) in the field of chemometrics
- Idea of PLS regression is to find uncorrelated linear transformations of the original predictor variables which have high covariance with the response
- Gene expression data are very similar to spectroscopic data:
  - Large number of genes
  - Large amount of systematic variation
- PLS is very well suited for the analysis of high-dimensional problems arising from the genomic experiments
- Large number of algorithmic variants of PLS exist

### Basic PLS Algorithm (Univariate Continuous response)

Y - centered and scaled, each X has mean(X) = 0, Var(X) = 1 for all i

1. Initialize Y = \bar{Y}, \text{X}^0 = \text{X}.
2. Calculate the individual regression coefficients of \( Y \) on each \( X_k \)
   \[ w_k^0 = \text{X}_k^T \cdot Y, \]  
3. Form the PLS component as the weighted sum of \( X_k \)
   \[ t_0 = \sum w_k^0 \cdot X_k^0, \]  
4. Update by the residuals of the previous linear fit
   \[ t_0 = Y - A_0 \cdot t_0, \]  
5. Update by the residuals of the previous linear fit
   \[ Y_{i+1} = Y - A_i \cdot t_i, \]  
6. Iterate these 5 steps until convergence

### PLS Algorithm Extension

The PLS modification for binary response could be written as a set of the following steps:
1. Calculate the individual logistic regression coefficients of \( Y \) on each \( X_k \)
   \[ w_k^0 = \text{X}_k^T \cdot \text{Y}, \]  
2. Form the PLS component as the weighted sum of \( X_k \)
   \[ t_0 = \sum w_k^0 \cdot X_k^0, \]  
3. Update the \( X_k^0 \) by orthogonalizing them with respect to \( t_0 \)
4. Iterate these 3 steps \( k=1,...,g \) (\( g \) = number of components) and calculate the PLS coefficients
5. From logistic regression of \( Y \) on components \( t_k \)

### PLS-FDR: Shrinking PLS coefficients using FDR corrected p-values

- PLS components can be expressed as linear combinations of the original predictor variables
- PLS-FDR: Shrinking PLS coefficients using FDR corrected p-values
- The coefficients of the linear combinations are slopes from individual linear regressions. If \( Y \) and \( X \) are normalized, the coefficients are the sample correlation coefficients between \( Y \) and \( X \)
- Conditioningly on the sample size and assuming normal errors, we express the PLS coefficients as functions of \( p \)-values, for example
  \[ t_k = \sum w_k^0 \cdot X_k^0 \cdot h(p_i) \cdot X_i^0, \]  
  \[ h(p_i) = 1 - A(n) - B(n)/\log(p_i) + C(n)/p_i/\log^2(p_i) \]

### PLS-FDR: Shrinking PLS coefficients using FDR corrected p-values

- For the t-distribution which is used for the linear regression coefficients and in the t-test of two-group comparison, the normalizing transformation:
  \[ z = \left[ \frac{n \cdot \ln(1 + t^2/n)}{2} \right]^{1/2} \]  
- Using the quantile approximation of \( z \) through \( p \)
  \[ z = \frac{[n \cdot \ln(1 + t^2/n)]^{1/2}}{b}, \quad a = 1.48 - 0.93 \frac{n}{n-1}, \quad b = 0.108 \]  
- Get the following approximation for the t value
  \[ t_{p,q} = \left[ [1 + (1 + 2 \cdot p \cdot (1 - p))^{1/2}] - 1 \right]^{1/2} \]  
- Using the Taylor series expansion, we can derive the following expression:
  \[ t_{p,q} = \frac{1}{\sqrt{\text{B}(n)/\text{B}(p)}} - b \left( a \left( \frac{1}{\log(p)} - \frac{1}{\log(p)} \right) \right) \]  
- A crude and simple approximation could then be the first term
  \[ t_{p,q} = \frac{1}{\sqrt{\text{B}(n)/\text{B}(p)}} \]  
- The similar expression was obtained when looking at the correlation coefficient using Fisher transformation
  \[ \tau = \frac{1}{2} \ln \frac{1 + c}{1 - c} \]
4. Repeat 1-3 N times and collect an ensemble of N predictions.

3. Use the Classifier to predict out-of-bag data to form majority votes.

- Plots of approximation versus logistic regression coefficients
  - binary response is generated from a logistic distribution
  - binary response is modeled as two groups where mean difference is introduced for a predictor

Classification: Ensemble Classifiers
1. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
2. Generate m randomly selected features according to the weights \( W \) and use them together with the bootstrap sample to construct a classifier using method "A".
3. Use the Classifier to predict out-of-bag data to form majority votes.
4. Repeat 1-3 N times and collect an ensemble of N predictions according to method "A". Prediction of test data is done by majority votes from predictions from the ensemble of Classifiers.
5. Classifier "A" can be any standard classifier: LDA, ANN, PLS, SVM, KNN, LASSO,…. 

PLS-FDR: Shrinking PLS coefficients using FDR corrected p-values

- Plot of approximations versus linear regression coefficients

Enriched Partial Logistic Regression example

Partial Logistic Regression

FDR corrected PLS
How do use these weights?

- Indirectly as part of the ensemble algorithm:
- Directly:

  Histogram of null hypothesis for p-values generated by permutations

<table>
<thead>
<tr>
<th>p-values from sample 1</th>
<th>p-values from sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram of p-values</td>
<td>Histogram of p-values</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

General FDR Weights

- Given any procedure that produces individual p-values for each predictor, we can assign weights as 
  \[ w = -\log(p_{FDR}) \]

- How do use these weights?
  - Directly:
  - Weighted – Random Forest, Trees, LDA, SVM
  - As individual shrinkage parameters – Lasso, Elastic Net, Ridge
  - Indirectly as part of the ensemble algorithm:
    1. Draw a bootstrap sample from the data. Call the observations which are not in the bootstrap sample the “out-of-bag” data.
    2. Generate m randomly selected features according to the weights (w) and use them together with the bootstrap sample to construct a classifier.
    3. Use the classifier to predict out-of-bag data to form majority votes.
    4. Repeat steps 1-3 N times and collect an ensemble of N rules. Prediction of test data is done by majority votes from predictions from the ensemble of rules.

Enriched Clustering: ABC clustering.

1. A Bootstrap approach called ABC
   Refers to the Bagging of genes and samples from Microarray data. Genes are bagged using weights proportional to their variances.

2. By creating new datasets out of subsets of columns and genes we are able to create estimates of the class response several hundred times.

3. These estimates are then used to obtain a dissimilarity (distance) measure between the samples of the original data.

4. This dissimilarity matrix is then adopted to cluster the data.

Ensemble Classifiers

<table>
<thead>
<tr>
<th>KNN (Bag)</th>
<th>KNN (Sim)</th>
<th>KNN (Enr)</th>
<th>KNN (Sim)</th>
<th>KNN (Enr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-LDA</td>
<td>DLDA</td>
<td>ElNet</td>
<td>RF</td>
<td>SVM</td>
</tr>
</tbody>
</table>

Out of Bag Error Rates

<table>
<thead>
<tr>
<th>Slc17A5 Day 0</th>
<th>Slc17A5 Day 18</th>
<th>Slc17A5 Day 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.833 0.167 0.000 0.314 0.029 0.029 0.000 0.029 0.029</td>
<td>0.583 0.667 0.667 0.583 0.583 0.583 0.750 0.750 0.750</td>
<td>0.111 0.083 0.056 0.139 0.139 0.111 0.028 0.083 0.028</td>
</tr>
<tr>
<td>0.143 0.179 0.179 0.286 0.286 0.286 0.179 0.214 0.179</td>
<td>0.417 0.583 0.000 0.500 0.583 0.250 0.000 0.500 0.000</td>
<td>0.083 0.083 0.056 0.139 0.139 0.111 0.028 0.083 0.028</td>
</tr>
<tr>
<td>0.143 0.143 0.071 0.429 0.429 0.214 0.214 0.143 0.214</td>
<td>0.143 0.179 0.179 0.286 0.286 0.286 0.179 0.214 0.179</td>
<td>0.083 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>0.333 0.833 0.167 0.333 0.417 0.250 0.500 0.583 0.500</td>
<td>0.333 0.667 0.667 0.583 0.667 0.667 0.833 0.833 0.833</td>
<td>0.143 0.143 0.071 0.429 0.429 0.214 0.214 0.143 0.214</td>
</tr>
<tr>
<td>0.333 0.833 0.167 0.333 0.417 0.250 0.500 0.583 0.500</td>
<td>0.333 0.667 0.667 0.583 0.667 0.667 0.833 0.833 0.833</td>
<td>0.143 0.143 0.071 0.429 0.429 0.214 0.214 0.143 0.214</td>
</tr>
<tr>
<td>0.333 0.833 0.167 0.333 0.417 0.250 0.500 0.583 0.500</td>
<td>0.333 0.667 0.667 0.583 0.667 0.667 0.833 0.833 0.833</td>
<td>0.143 0.143 0.071 0.429 0.429 0.214 0.214 0.143 0.214</td>
</tr>
<tr>
<td>0.333 0.833 0.167 0.333 0.417 0.250 0.500 0.583 0.500</td>
<td>0.333 0.667 0.667 0.583 0.667 0.667 0.833 0.833 0.833</td>
<td>0.143 0.143 0.071 0.429 0.429 0.214 0.214 0.143 0.214</td>
</tr>
</tbody>
</table>
**Similarity**

Data

<table>
<thead>
<tr>
<th>Similarity</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>S2</td>
<td>0.667</td>
<td>1</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>S3</td>
<td>0.667</td>
<td>0.667</td>
<td>1</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>S4</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>1</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>S5</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>1</td>
<td>0.667</td>
</tr>
<tr>
<td>S6</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>1</td>
</tr>
</tbody>
</table>

**Out of Bag Error Rates**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>RF</th>
<th>RF(p)</th>
<th>RF(p)</th>
<th>RF(p)</th>
<th>RF(p)</th>
<th>RF(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>0.028</td>
<td>0.024</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Gene expression</td>
<td>0.047</td>
<td>0.047</td>
<td>0.047</td>
<td>0.047</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td>METAB[1]</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>METAB[2]</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>METAB[3]</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>METAB[4]</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
</tbody>
</table>

**References**


D. Amaratunga and J. Cabrera (2007). A conditional t suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication, in advance access publication in *Statistics in Biopharmaceutical Research*.


**Website**

www.rutgers.edu/~cabrera/DNAMR