Borrowing information from relevant microarray studies for sample classification using weighted partial least squares

Xiaohong Huang

Eli Lilly and Company

Indianapolis, IN 46285

Outline

- Introduction
- Statistical classification methods for microarray data
  - Partial least squares (PLS)
  - Penalized partial least squares (PPLS)
  - Applications of PPLS to Minnesota data and PGA data
- Classification with combined data of multiple studies
  - PLS with a conjugate gradient path
  - Weighted PLS/PPLS
  - Experiment: Combined data
- Summary and discussion

Introduction

- Traditional medical diagnosis/classification method is very subjective
  - Based on morphological characteristics, pathological features
  - Depends on highly trained pathologists

Limitation: Hard to diagnose disease subtypes that are morphologically similar but follow different clinical courses.

- New classification method is objective
  - Based on microarray gene expression data.
  - Can be highly accurate.

Potentials: diagnose disease subtypes; predict clinical outcomes...

Example: Two-class microarray

- Notations:

$$\begin{array}{c|cccccc}
 & \multicolumn{3}{c}{\text{Class 1}} & \multicolumn{3}{c}{\text{Class 2}} \\
 & 1 & \dots & n_1 & n_1+1 & \dots & n_1+n_2=n \\
\hline
\text{gene } 1 & X_{1,1} & \dots & X_{1,n_1} & X_{1,n_1+1} & \dots & X_{1,n} \\
\text{gene } 2 & X_{2,1} & \dots & X_{2,n_1} & X_{2,n_1+1} & \dots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\text{gene } p & X_{p,1} & \dots & X_{p,n_1} & X_{p,n_1+1} & \dots & X_{p,n}
\end{array}$$

- Outcome $Y = (y_1, y_2, \dots, y_n)'$.
- Covariates $X_i = (x_{i1}, x_{i2}, \dots, x_{in})'$, $i = 1, \dots, p$.
  Covariates are often standardized $var(x_i) = 1$.

- A special feature of microarray data:

  Small n, large p

- A simple prediction problem:
  - Our goal is to predict Y from $X_1, X_2, ..., X_p$ by a linear model.
  - Especially interested in problems where $p \gg n$.
- Many new methods have appeared
  - Weighted voting, Compound covariate,...
  - Penalized regression: Shrunken centroids, LASSO,...
  - Machine learning: SVM, Bagging/boosting trees,...

Penalized Partial Least Squares

- Partial Least Squares (PLS)
  - Particularly suited for constructing linear models when there are more variables than observations.
  - Robust to the collinearity between covariates.
  - Suited for fitting linear models with microarray data.
- Penalized Partial Least Squares (PPLS)
  - A penalized regression method built on the framework of PLS.

Ref: **Huang, X.** and Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics* **19**, 2072-2078.

Application to the Minnesota data: PPLS

- Minnesota data
  - Oligonucleotide microarray data obtained by Hall et al. (2003) in a heart failure study conducted at the Medical School of UMN.
  - Contain 30 samples: 10 ischemic, 7 ischemic with acute MI and 13 idiopathic.
  - Affy HG-U133A chips: Contain 22,283 genes.
  - Initially processed in MAS 5.0.

Goal: Distinguish between the ischemic and the idiopathic etiology classes.

Initial gene ranking

- Given gene i

$$F_i = \frac{MS_{class}}{MS_{error}} = \frac{(\sum_{c=1}^{C} n_c(\bar{x}_{i_c} - \bar{x}_i)^2)/(C-1)}{(\sum_{c=1}^{C}\sum_{j \in c}(x_{ij} - \bar{x}_{i_c})^2)/(n-C)}.$$

x_{ij}: gene expression intensity of gene i and sample j.

n_c: number of samples in class c, $n = \sum_{c=1}^{C} n_c$: total sample size.

C: number of classes.

$\bar{x}_{i_c}$: mean gene expression of class c.

$\bar{x}_i$: overall mean gene expression.

- Genes with larger F-statistics -> higher rank.

Experiment: Minnesota data

- LOOCV error (Isch vs Idio, $n = 23$)

# of top genes	PPLS	SC	LASSO
50	10	5	6
100	7	5	7
200	9	7	11
800	6	8	4
1600	5	8	8
9600	9	8	7
16000	8	7	8
22283	9	5	7

Application to the PGA data: PPLS

- PGA data
  - Oligonucleotide microarray data obtained in a heart failure study conducted at the PGA Medical School.
  - Contain 36 samples: 11 normal, 11 ischemic, and 14 idiopathic.
  - Affy HG-U133 plus 2 chips: Contain $\sim 54,000$ genes.
  - Initially processed in MAS 5.0.

Goal: Distinguish between the ischemic and the idiopathic etiology classes.

Experiment: PGA data

- LOOCV error (Isch vs Idio, $n = 25$)

# of top genes	PPLS	SC	LASSO
50	3	2	2
100	1	2	1
200	1	2	1
800	1	1	3
1600	1	1	2
9600	1	1	1
16000	1	1	1
22277	1	1	1

A summary on the above experiments

- The LOOCV misclassification error
  - ranges from 5 to 11 for the Minnesota data (23 samples).
  - ranges from 1 to 3 for the PGA data (25 samples).
- These highlight some existing differences underlying the two datasets.
- **A question: Is there any signal/predictive information in the data?**

Permutation test: Minnesota data

- LOOCV error (Isch vs Idio, $n = 23$)

# of	Original data		Permutated data				
top genes	CV errors	P-value	0%	25%	50%	75%	100%
50	5	.00	5	10	11	12.75	19
100	5	.00	5	10	11	12	19
400	7	.06	4	9	11	13	17
1600	8	.08	5	9.25	11.5	13	17
6400	9	.12	6	10	11	13	21

Borrow information from relevant studies

- To increase the statistical power, borrow information from other relevant studies.
- A key difference from meta-analysis:
  - Not assuming current study shares a common set of parameters with other studies.
- Example: Identifying genes associated with ventilator-associated lung injury (VALI) based on a human study.
  - Meta-analysis: Only interested in the genes associated with VALI which are conserved across the species over the evolutionary history (Grigoryev et al. 2004).
  - Our analysis: Interested in inference on a set of parameters specific for humans.

Combining Minnesota and PGA data

- Classification with the combined data.
    - Goal: Distinguish etiologies of heart failure for Minnesota patients while treat the PGA data as secondary.
    - Problem: Unobserved differences in patient characteristics.
    - Solution: Treat samples in different studies unequally, e.g, assign different weights.
- Combining Minnesota data and PGA data.
    - Technically easy: all probe sets present on U133A chip are identically replicated on U133 Plus 2 chip.
    - Data mapped by probe set ID (6 could not be found).

Notations

- Given X, to predict Y with a linear model $F(X, \mathbf{a}) = a_0 + \sum_{i=1}^{p} a_i x_i$, the goal is to minimize the expected loss (risk):

$$R(\mathbf{a}) = E_Y L(Y, F(X, \mathbf{a})).$$

  - $L(Y, F(X, \mathbf{a}))$: loss criterion.

- An empirical estimate of the expected loss:

$$\widehat{R}(\mathbf{a}) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}).$$

- The optimal values of $\mathbf{a}$:

$$\hat{\mathbf{a}} = arg \min_{a} \frac{1}{n} \sum_{j=1}^{n} L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}).$$

Partial Least Squares (PLS)

- Conjugate gradient procedure under squared error loss:

$$\begin{aligned} \hat{\mathbf{a}}_{k+1} &= \hat{\mathbf{a}}_k + \rho_k \mathbf{s}_k \\ \mathbf{s}_k &= \mathbf{g}_k + \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}} \mathbf{s}_{k-1} \end{aligned}$$

  - $\mathbf{g}_k$: negative gradient at $\hat{\mathbf{a}}_k$

$$\mathbf{g}_k = -\frac{\partial}{\partial \mathbf{a}} \widehat{R}(\mathbf{a}) \Big|_{\mathbf{a}=\hat{\mathbf{a}}_k}$$

  - ρ_k: step size

$$\rho_k = argmin_\rho \widehat{R}(\hat{\mathbf{a}}_k + \rho \mathbf{s}_k)$$

  - k: number of PLS components.
  - Squared error loss: $L(Y, F(X, \mathbf{a})) = (Y - F(X, \mathbf{a}))^2/2$.

Propose: Weighted PLS

- Expected loss estimated by a weighted average loss:

$$\widehat{R}_w(\mathbf{a}) = \sum_{j=1}^{n} w_j L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}).$$

- Conjugate gradient procedure under squared error loss:

$$\mathbf{g}_k = -\frac{\partial}{\partial \mathbf{a}} \widehat{R}_w(\mathbf{a}) \Big|_{\mathbf{a}=\hat{\mathbf{a}}_k} = Z^T W(Y - Z\hat{\mathbf{a}}_k)$$

$$\rho_k = argmin_\rho \widehat{R}_w(\hat{\mathbf{a}}_k + \rho \mathbf{s}_k) = \begin{cases} 1 & if\ Z\mathbf{s}_k = 0 \\ \frac{(Z\mathbf{s}_k)^T W(Y - Z\hat{\mathbf{a}}_k)}{(Z\mathbf{s}_k)^T W(Z\mathbf{s}_k)} & if\ Z\mathbf{s}_k \neq 0 \end{cases}$$

  - $W = diag(w_1, \cdots, w_n)$: diagonal matrix with weights, $\sum_{j=1}^{n} w_j = 1$.
  - $Z = \begin{pmatrix} \mathbf{1} & X_1 & X_2 & \cdots & X_p \end{pmatrix}_{n \times (p+1)}$: covariate matrix.

Propose: Weighted PPLS

- Weighted PPLS
  - A weighted PLS model with a conjugate gradient path.
  - Penalized regression in the framework of weighted PLS.
  - Similar to the PPLS construction.
- Weighted PLS model: $Y = b_0 + \sum_{i=1}^{p} b_i(X_i - \bar{x}_i \mathbf{1})$
- Penalize b_i by soft-thresholding:

$$b_i' = arg\min_{\beta_i}(\beta_i - b_i)^2 + \lambda\big|\beta_i\big|$$

  - $b_i' = sign(b_i)(|b_i| - \lambda)_+$.
  - $f_+ = max(f, 0)$.
  - λ: shrinkage parameter.

Experiment: Combined data

- LOOCV error of predicting Minnesota samples
  - Weighted PPLS classifiers.
  - Top 200 genes

$w = PGA$	Shrink 0%					Shrink 40%					Shrink 80%				
	PLS components k					PLS components k					PLS components k				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0	5	9	6	7	7	5	8	8	6	5	5	8	7	6	7
1/4	6	8	6	6	5	5	7	4	3	5	8	7	2	4	3
1/2	5	8	6	4	6	6	7	3	4	5	7	7	2	3	2
3/4	4	7	7	4	6	6	7	4	4	6	8	6	2	2	2
1	5	7	7	4	7	6	7	4	4	6	8	5	3	2	2

Summary

- Weighted PPLS:
  - Penalized regression in a framework of weighted PLS.
  - Penalization/shrinking can improve over weighted PLS.
- Weighted PPLS methods with combined data:
  - Account for possible different relevances of the other studies by weighting.
  - Improve the performance of the classifier using data from a single study.

General application

- Broad scope of the weighting scheme:
  - Applicable when the PGA data only contain ischemic and normal groups.
- Further extendable:
  - The primary and secondary experiments were conducted under different (but relevant) conditions, or on different organisms.
  - Microarray data with a survival end point.
  - Other loss functions.

Reference

- **Huang, X.**, Pan, W., Grindle, S., Han, X., Chen, Y., Park, S.J., Miller, W.L., Hall, J. (2005). A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics* **6**:205.
- **Huang, X.**, Pan, W., Han, X., Chen, Y., Miller, W.L., Hall, J. (2005). Borrowing information from relevant microarray studies for sample classification using weighted partial least squares. *Computational Biology and Chemistry* **29(3)**, 204-211.