BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Clustering Threshold Gradient Descent Regularization: with Applications to Microarray Studies

Shuangge Ma [a] and Jian Huang [b]

[a]Department of Epidemiology and Public Health, Yale University, New Haven, CT, USA
[b]Departments of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA

ABSTRACT

Motivation An important goal of microarray studies is to discover genes that are associated with clinical outcomes such as disease status and patient survival. While a typical experiment surveys gene expressions on a global scale, there may be only a small number of genes that have significant influence on a clinical outcome. Moreover, expression data have cluster structures and the genes within a cluster have correlated expressions and coordinated functions, but the effects of individual genes in the same cluster may be different. Accordingly, we seek to build statistical models with the following properties. First, the model is sparse in the sense that only a subset of the parameter vector is non-zero. Second, the cluster structures of gene expressions are properly accounted for.

Results: For gene expression data without pathway information, we divide genes into clusters are using commonly used methods such as K-means or hierarchical approaches. The optimal number of clusters is determined using the Gap statistic. We propose a Clustering Threshold Gradient Descent Regularization (CTGDR) method, for simultaneous cluster selection and within cluster gene selection. We apply this method to binary classification and censored survival analysis. Compared to the standard TGDR and other regularization methods, the CTGDR takes into account the cluster structure and carries out feature selection at both the cluster level and within-cluster gene level. We demonstrate the CTGDR on two studies of cancer classification and two studies correlating survival of lymphoma patients with microarray expressions.

Availability: R code is available upon request.

Contact: shuangge.ma@yale.edu

1 INTRODUCTION

Microarray technology provides a way of monitoring gene expressions on a large scale. Tremendous efforts have been devoted to discovering genes that are associated with variations of clinical outcomes. Understanding of the molecular biology that underlies such variations might provide a more accurate method of diagnosis and suggest new therapeutic approaches. See for example, Alizadeh et al. (2000), Garber et al. (2001), and Rosenwald et al. (2003). Two types of clinical outcomes have been of special interest. The first type is categorical outcome, which includes the presence or absence of tumor as in Alon et al. (1999) or different types of tumors as in Alizadeh et al. (2000). The second type is survival outcome, which corresponds to the occurrence time of certain event such as cancer. See for example Rosenwald et al. (2003) and Dave et al. (2004).

Classification and survival analysis using microarray data are challenging because of the large number of genes and relatively small sample size. Various model reduction methods have been proposed, including the singular value decomposition (Golub and Van Loan 1996), partial least squares (Nguyen and Rocke 2002), principal component analysis (Ma et al. 2006), LASSO-LARS (Gui and Li 2005a), and Threshold Gradient Descent Regularization (TGDR, Gui and Li 2005b; Ma and Huang 2005) among others. The essence of the aforementioned techniques is to identify a small number of representative features–individual genes or linear combinations of genes, and build predictive models based on those representative features. In the feature selection, all genes are treated in an equal manner and the intrinsic gene correlation structures are usually ignored.

Statistically speaking, there exist genes whose expressions are highly correlated and should be put into clusters (Tamayo et al. 1999). Biologically speaking, there exist gene pathways composed of co-regulated genes with coordinated functions (Eisen et al. 1998). See for example the Gene Ontology (Harris et al. 2004). Although clusters defined based on statistical correlation and biological functions do not match perfectly, they tend to have certain correspondence (Clare and King 2002; Tavazoie et al. 1999; Yeung et al. 2001).

Cluster analysis methods have been employed in microarray studies as a dimension reduction tool (Alizadeh et al. 2000; Dave et al. 2004). With this approach, a small number of gene clusters are first constructed, using methods such as the K-means or hierarchical (Johnson and Wichern 2002). The mean expressions of genes within the same clusters are then computed and used as covariates for downstream model building. A limitation of this approach is that feature selection is carried out only at the cluster level. Once a cluster is used in the final model, all genes within that cluster are included. Although genes within the same cluster may have correlated expressions, it is not necessarily true that they will all be associated with a specific clinical outcome. Including noisy genes may lead to ill-behaved models. Gene selection within clusters is still needed to yield more reliable models. Wei and Li (2006) proposes a nonparametric pathway-based regression approach that explicitly makes use of available pathway information. They use the gradient-based boosting algorithm (GDB, Friedman 2001) for model fitting and the importance score (Breiman et al. 1984; Friedman 2001) for ranking pathways and genes. However, they do not explicitly consider variable selection at either the cluster or individual gene levels.

Regularization methods such as the LASSO and TGDR are effective methods for variable selection. Although capable of selecting a small number of important genes, these methods do not incorporate cluster structure. On the other hand, standard approaches using cluster analysis results as input explicitly take

© Oxford University Press 2005.

into account cluster structure, but cannot carry out individual gene selection.

To combine strength of the aforementioned approaches, we propose a Clustering TGDR (CTGDR) method that incorporates cluster structure into TGDR-based variable selection. The proposed CTGDR carries out feature selection at two levels: at the cluster level and the individual gene level within each cluster. Thus it takes advantages of both the cluster-based and regularized variable selection methods.

In section 2, we present the data and models that we consider. We use logistic regression for binary classification and Cox model for right censored survival analysis as examples. We select the optimal number of clusters using the Gap statistic. The CTGDR algorithm is described in section 3. Tuning parameter selection and evaluation are also discussed. We present two classification examples in section 4 and two survival analysis examples in section 5. The article ends with discussions in section 6.

2 DATA AND MODEL SETTINGS

Let Z be a length d vector of gene expressions, and let Y be the clinical outcome of interest. We assume that Y is associated with Z through model $Y \sim \phi(\beta' Z)$ with a regression function ϕ and unknown regression coefficient β. In addition, we assume there exists a smooth objective function and a proper estimate of β can be obtained by maximizing that function. We are particularly interested in the classification and survival analysis problems using microarray gene expression data due to their extensive applications in biomedical studies.

2.1 Binary classification

For the classification problems, Y is a categorical variable indicating the disease status. For simplicity, we focus on binary classification only. Suppose that $Y = 1$ denotes the presence and $Y = 0$ indicates the absence of disease. We assume the commonly used logistic regression model, where the logit of the conditional probability is $\text{logit}(P(Y = 1|Z)) = \alpha + \beta' Z$. Here β is the length d vector of unknown regression coefficient and α is the unknown intercept. Based on a random sample of n observations $X_i = (Y_i, Z_i), i = 1, \ldots, n$, the maximum likelihood estimator is defined as $(\hat{\alpha}, \hat{\beta}) = argmax_{\alpha,\beta} R_n(\alpha, \beta)$, where

$$R_n(\alpha, \beta) = \sum_{i=1}^{n} Y_i \log\left(\frac{\exp(\alpha + \beta' Z_i)}{1 + \exp(\alpha + \beta' Z_i)}\right) + (1 - Y_i) \log\left(\frac{1}{1 + \exp(\alpha + \beta' Z_i)}\right). \quad (1)$$

For simplicity, we denote $R_n(\alpha, \beta)$ as $R_n(\beta)$.

2.2 Cox survival analysis

For right censored survival data, $Y = (T, \Delta)$, where $T = min(U, V)$ and $\Delta = I(U \leq V)$. Here U and V denote the event time of interest and censoring time, respectively. The most widely used model for censored survival data is the Cox model (Cox, 1972) which assumes that the conditional hazard function $\lambda(u|Z) = \lambda_0(u) \exp(\beta' Z)$. λ_0 is the unknown baseline function and β is the regression coefficient. Based on a random sample of n observations $X_i = (Y_i, Z_i), i = 1, \ldots, n$, the partial likelihood estimator is defined as the value $\hat{\beta}$ that maximizes

$$R_n(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' Z_i)}{\sum_{j \in r_i} \exp(\beta' Z_j)} \right\}^{\delta_i},$$

where $r_i = \{j : T_j \geq T_i\}$ is the risk set at time T_i.

2.3 Cluster structure

The proposed CTGDR approach assumes the cluster structure has been well defined. Some gene expression data have well defined biological gene pathways. Such cluster structure can be obtained from web databases such as the GO (*http://www.geneontology.org*). See for example Wei and Li (2006). However it is also well known that the pathway information may be only partially or even not available for a large number of genes. In this case, we propose defining cluster structure based on statistical measurements (Tamayo et al. 1999).

Commonly used clustering methods include the hierarchical, K-means, tree-truncated vector quantization and self-organizing map methods, among many others. For a general reference, see Gordon (1999). In general there does not exist optimal clustering method. In this article, we consider the two most popular unsupervised approaches: hierarchical and K-means methods.

We use the Gap statistic (Tibshirani et al. 1999) to select the optimal number of clusters. For a chosen clustering approach, we first choose L_{max}–the largest number of clusters. Then for $L = 1, \ldots, L_{max}$:

1. Generate L clusters using the selected approach. Denote rss_L as the total within block sum of squares.
2. Create a new dataset by separately permuting each gene expression measurements. Apply the clustering method to the permuted expression data. Let $\widetilde{rss}_L$ denote the resulting within cluster sum of squares. Repeat this for a number of times and compute the average $ave(\widetilde{rss}_L)$.
3. Compute the Gap statistic as $gap(L) = ave(\widetilde{rss}_L) - rss_L$.

Choose the value L that maximizes $gap(L)$. We refer to Tibshirani et al. (1999) and McLachlan et al. (2004) for detailed discussions of the Gap statistic. We assume gene $j = 1, \ldots, d$ belongs to one of the clusters $C(j) \in \{1, \ldots, L\}$.

3 CLUSTERING TGDR

The CTGDR can be consider a generalization of the TGDR, which is introduced by Friedman and Popescu (2004) in the context of linear regression analysis and has been employed in microarray studies by Gui and Li (2005b) and Ma and Huang (2005). For completeness, we first briefly describe the TGDR algorithm.

3.1 TGDR algorithm

Denote $\Delta\nu$ as the small positive increment as in ordinary gradient descent methods (Friedman and Popescu 2004). In the implementation of this algorithm, we choose $\Delta\nu = 1 \times 10^{-4}$. Denote $\nu_k = k \times \Delta\nu$ as the index for the point along the parameter path after k steps. Let $\beta(\nu_k)$ denote the parameter estimate corresponding to ν_k. For any fixed threshold $0 \leq \tau \leq 1$, the TGDR algorithm consists of the following steps:

1. Initialize $\beta(0) = 0$ and $\nu_0 = 0$.

2. With current estimate β, compute the negative gradient $g(\nu) = -\partial R_n(\beta)/\partial\beta$. Denote the j^{th} component of $g(\nu)$ as $g_j(\nu)$. If $\max_j\{|g_j(\nu)|\} = 0$, stop the iterations.
3. Compute the threshold vector $f(\nu)$ of length d, where the j^{th} component of $f(\nu)$: $f_j(\nu) = I\{|g_j(\nu)| \geq \tau \times \max_l |g_l(\nu)|\}$.
4. Update $\beta(\nu + \Delta\nu) = \beta(\nu) - \Delta\nu \times g(\nu) \times f(\nu)$ and update ν by $\nu + \Delta\nu$, where the product of f and g is component-wise.
5. Steps 2–4 are repeated k times. The number of iterations k is determined by cross validation.

The tuning parameters τ and k jointly determine the property of β. When $\tau \approx 0$, β is dense even for small values of k. When $\tau \approx 1$, β is sparse for small k and remains so for a relatively large number of iterations, but will become dense eventually. At the extreme when $\tau = 1$, the TGDR usually increases in the direction of a single covariate in each iteration. When τ is in the middle range, the characteristics of β are between those for $\tau = 0$ and $\tau = 1$. For $\tau \neq 0$, variable selection can be achieved with cross validated, finite k, by having certain components of β exactly zero. We refer to Friedman and Popescu (2004) for more detailed discussions. The TGDR described here is capable of individual gene selection but does not account for the cluster structure.

3.2 Naive CTGDR

Naive CTGDR Algorithm I. This algorithm modifies step 3 of the TGDR as follows:

$$f_j^1(\nu) = I\left\{\sum_{m\in C(j)} |g_m(\nu)| \geq \tau_1 \times \max_{C(k)} \sum_{l\in C(k)} |g_l(\nu)|\right\}, \quad (2)$$

where $0 \leq \tau_1 \leq 1$ is the threshold tuning parameter. The other steps in the TGDR are kept unchanged.

Compared to the original TGDR, algorithm I uses *cluster gradients* to replace individual gradients. The combined effects of genes in the same clusters are considered and compared with the combined effects of other clusters. This algorithm is similar to the traditional clustering approaches in the sense that gene selection is achieved on a cluster basis, and if the combined effect of genes in a cluster is important, then all the genes within this cluster will be included in the final model. The key difference is that the naive CTGDR I estimated coefficients of genes in the same clusters may be different. So genes within the same clusters may still have different contributions in the final model, whereas in traditional cluster based methods, all genes within the same clusters have the same coefficient and hence equal contributions to the outcome.

Algorithm I does feature selection at the cluster level. If a cluster is selected, then all the genes in this cluster are selected. Thus the total number of genes in the final model can be large. Consider for an example a hypothetical study with 2000 genes and five clusters of equal sizes are constructed. Then using algorithm I, it is possible three or four clusters are selected. The total number of genes in the final model will be greater than 1000. Although the prediction performance may still be satisfactory, this makes the final estimation results hard to interpret from a gene discovery point of view. Since it is often the case that only a subset of genes within each cluster have important impact on the outcome of interest, gene selection within cluster is still needed.

Naive CTGDR Algorithm II. This algorithm partly solves the drawbacks of algorithm I. Denote $\tau_2 \in [0, 1]$ as the threshold tuning parameter. We replace f in step 3 of the TGDR with

$$f_j^2(\nu) = I\left\{|g_j(\nu)| \geq \tau_2 \times \max_{l\in C(j)} |g_l(\nu)|\right\}, \quad (3)$$

so that each gene is only compared with other genes within the same cluster and only important genes from each cluster are selected. The rationale is that genes from different clusters may not be directly comparable. So a fair comparison should be for genes within the same clusters. Within each cluster, we use the TGDR to identify important genes.

We have employed algorithm II in the examples in sections 4 and 5. We are able to identify a smaller number of genes ($\sim$200, much fewer than that from naive algorithm I) with satisfactory prediction performance. However, algorithm II has its own drawbacks. It is roughly equivalent to carrying out the TGDR in each cluster separately and the final model includes genes selected from all clusters. The underlying assumption is that all clusters are associated with the outcome of interest. Previous cluster based methods as in Dave et al. (2004) and Alizadeh et al. (2000) show that this is not necessarily true. Cluster selection is still needed.

3.3 CTGDR algorithm

The naive CTGDR algorithm I carries out cluster selection, but does not select important genes within each cluster. On the other hand, the naive CTGDR algorithm II does gene selection in each cluster separately, but does not select clusters. The advantages and drawbacks of the naive CTGDR algorithms motivate the following CTGDR algorithm.

Let $\tau_1,\ \tau_2 \in [0, 1]$ be two threshold parameters. In step 3 of the TGDR algorithm, define

$$f_j(\nu) = f_j^1(\nu) \times f_j^2(\nu), \quad (4)$$

where $f^1(\nu)$ is defined in (2) with threshold value τ_1 and $f^2(\nu)$ is defined in (3) with threshold value τ_2, respectively.

In (4), the term $f^1(\nu)$ carries out cluster selection, while $f^2(\nu)$ carries out within-cluster gene selection. So the combined f can carry out feature selection at both the cluster level and within cluster level. Further flexibility is introduced by allowing two possibly different threshold values. In this algorithm, if a gene or a cluster is known to be associated with the clinical outcome *a priori*, then it can be excluded from the thresholding step.

The three tuning parameters k, τ_1 and τ_2 jointly determine the properties of the CTGDR estimates. The τ_1 and τ_2 have similar effects as the tuning parameter τ for the standard TGDR in section 3.1. If τ_1 and τ_2 are both close to 1, then the estimate remains sparse for a relatively large k, but will become dense eventually. If τ_1 and τ_2 are both close to 0, the estimate is dense for even a very small k. τ_1 and τ_2 determine the degree of sparsity on cluster level and within cluster level, respectively, with larger thresholding values leading to more parsimonious models with fixed k. With nonzero τ_1 and τ_2, the model with small to moderate k usually has a small number of clusters and a small number of genes within each selected cluster.

3.3.1 Possible extensions In the above CTGDR algorithm, the cluster gradient is defined as the sum of absolute values of individual gradients. This is the default definition when there is no extra information on the clusters. If there exists external knowledge of

the clusters, then we can modify the indicator function in (2) as

$$I\left\{w_j \sum_{m\in C(j)} |g_m(\nu)| \geq \tau_1 \times \max_{C(k)} w_k \sum_{l\in C(k)} |g_l(\nu)|\right\}, \qquad (5)$$

where w_j is the positive weight measuring the relative importance of cluster j. A simple choice of w_j is the inverse of cluster size, so that the relative importance of clusters is not affected by cluster size. If external knowledge about the relative importance of genes within the same cluster is present, then the cluster gradient can be defined as the weighted sum of individual gradients, with more stable and more important genes having larger weights. Further flexibility can be introduced by considering weighted gradients.

3.4 Tuning parameter selection

We select the tuning parameters k and (τ_1, τ_2), which jointly determine the characteristics of the estimator, using the following two-step approach.

First we choose the tuning parameter k for any fixed (τ_1, τ_2) using V-fold cross validation (Wahba 1990) as follows. Partition the data randomly into V non-overlapping subsets of equal sizes. Choose k to maximize the cross-validated objective function

$$CV(k) = \sum_{v=1}^{V} \left[R_n(\beta^{(-v)}) - R_n^{(-v)}(\beta^{(-v)})\right], \qquad (6)$$

where $\beta^{(-v)}$ is the CTGDR estimate of β based on the data without the v^{th} subset for a fixed k and $R_n^{(-v)}$ is the objective function R_n evaluated without the v^{th} subset. In our study, we set $V = 5$.

After cross validation over k, model features for different τ_1 and τ_2 can be obtained. We choose parsimonious models with relatively large CV score. An AIC type score as in Huang et al. (2006) can be used as model selection criterion. Cross validation over τ_1 and τ_2 can also be considered, i.e, we can select the model with the largest CV score over all possible k, τ_1 and τ_2. However, this approach may lead to models with slightly larger CV scores, but a lot more genes, which may be less stable models.

3.5 Evaluation

Unlike in standard classification or survival analysis where the association between clinical outcome and covariates is of primary interest, studies given in sections 4 and 5 put more emphasis on selection of important genes and prediction. So we consider the following cross validation based approach for evaluating prediction performance, as suggested in Ma and Huang (2005).

1. We first partition the data randomly into a training set of size n_1 and a testing set of size n_2 with $n_1 + n_2 = n$. In this article, we set $n_1 \sim 2/3n$.
2. Compute the CTGDR estimate based on the training set only. Using this training set estimate, we compute a prediction index for the testing set.
3. To take into account the possibility of an extreme prediction performance due to a rare partition, we repeat this process B (for example 200) times. Each time a new partition is made and the prediction index is computed.

For classification studies, the prediction index can be the prediction error. For censored survival studies, we first create two risk groups based on dichotomizing the estimated linear risk scores $\hat{\beta}' Z_i$ at the median risk score for the testing set. We then use the Logrank statistic to assess whether the survival curves of different risk groups are different. A large value of the Logrank statistic indicates that the high and low risk groups are well separated, and suggests satisfactory prediction performance of the CTGDR estimate.

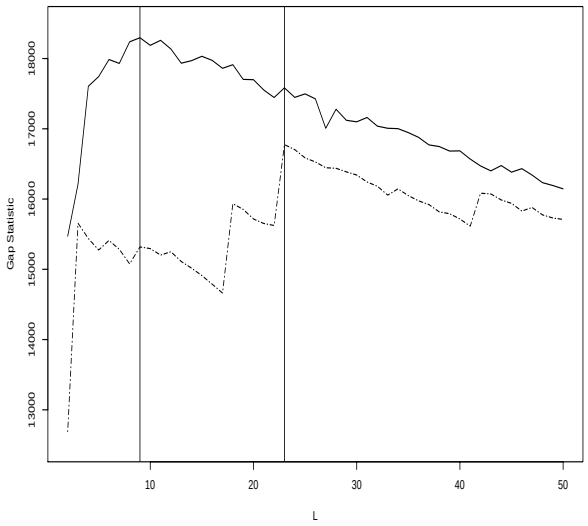

Fig. 1. Colon data: gap statistic as a function of number of clusters. Solid line: K-means clustering. Dashed line: Hierarchical clustering.

4 BINARY CLASSIFICATION

Colon data. In this dataset, expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes are measured using the Affymetrix gene chips. A selection of 2000 genes with the highest minimal intensity across the samples has been made by Alon et al. (1999), and these data are publicly available at *http://microarray.princeton.edu/oncology/*. The colon data have been analyzed in several previous studies using other statistical approaches, see for example Dettling and Buhlmann (2003), Pochet et al. (2004), Nguyen and Rocke (2002) and Ma and Huang (2005).

Nodal data. This dataset was first presented by West et al. (2001) and Spang et al. (2001). It includes expression values of 7129 genes from 49 breast tumor samples. The expression data were obtained using the Affymetrix gene chip technology and are available at *http://mgm.duke.edu/genome/dna_micro/work/*. The response describes the lymph nodal (LN) status, which is an indicator for the metastatic spread of the tumor. Among the 49 samples, 25 are positive (LN+) and 24 are negative (LN−). We threshold the raw data with a floor of 100 and a ceiling of 16000. Genes with $\max(expression)/\min(expression) < 10$ and/or $\max(expression) - \min(expression) < 1000$ are also excluded (Dudoit et al. 2002). 3332 (46.7%) genes pass the first step screening. A base 2 logarithmic transformation is then applied. This data have also been studied by Dettling and Buhlmann (2003).

Table 1. Colon and Nodal data. Tuning: values of tuning parameters; Nonzero: number of selected genes; Clus.: number of selected clusters; Error: mean prediction errors.

Approach	Tuning	Nonzero	Clus.	Error
	Colon			
K-means-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	13	5	0.111
K-means-simple	–	500	9	0.166
Hierarchical-CTGDR	$(\tau_1, \tau_2) = (0.9, 1.0)$	15	3	0.111
Hierarchical-simple	–	500	23	0.222
TGDR	$\tau = 0.9$	28	–	0.149
LASSO	$u = 1.4$	8	–	0.170
	Nodal			
K-means-CTGDR	$(\tau_1, \tau_2) = (1.0, 0.9)$	33	5	0.143
K-means-simple	–	500	10	0.228
Hierarchical-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	18	2	0.143
Hierarchical-simple	–	500	11	0.177
TGDR	$\tau = 1.0$	29	–	0.147
LASSO	$u = 8.0$	40	–	0.222

We first identify 500 genes for each dataset based on marginal significance to gain further stability as in Ma and Huang (2005). Compute the sample standard errors of the d genes $se_{(1)}, \ldots, se_{(d)}$ and denote their median as $med.se$. Compute the adjusted standard errors as $0.5(se_{(1)} + med.se), \ldots, 0.5(se_{(d)} + med.se)$. Then the genes are ranked based on the t-statistics computed with the adjusted standard errors. 500 genes with the largest absolute values of the adjusted t-statistics are used for classification.

For the Colon and Nodal data, we construct the clusters using the K-means and hierarchical methods. The number of clusters is chosen using the Gap statistic. For the Colon data, we show in Figure 1 the Gap statistic as a function of L. Since it has been suggested that the clusters should not be too small on average (Dave et al. 2004), we set $L_{max} = 50$. For the Colon data, the Gap statistic yields the optimal number of clusters 9 (K-means) and 23 (hierarchical); for the Nodal data, 10 (K-means) and 11 (hierarchical).

We apply the CTGDR to the clustered data obtained above. We consider the tuning parameters τ_1 and τ_2 taking values in the grid 0, 0.1,..., 1.0. Model features selected via cross validation are shown in Table 1. For the Colon data, 13 (K-means) and 15 (hierarchical) genes are selected in the final models, representing 5 and 3 clusters. For the Nodal data, 33 (K-means) and 18 (hierarchical) genes are selected in the final models, representing 5 and 2 clusters. The estimated coefficients and gene descriptions for the final models are available upon request.

We evaluate the prediction performance of the CTGDR using the approach discussed in section 3.5. For comparison, we also consider three alternatives: (1) a simple clustering approach, where the clusters are the same as used under the CTGDR. The within cluster median expressions are used as covariates. This mimics the approach in Dave et al. (2004). We refer to this approach as K-means-simple or Hierarchical-simple, depending on the clustering methods used; (2) Logistic model with TGDR for variable selection; and (3) Logistic model with LASSO for variable selection, where the tuning parameter u is the L_1 norm of β. For the TGDR and LASSO, the tuning parameters are also chosen via 5-fold cross validation.

For the Colon data, the CTGDR yields mean prediction errors 0.111 based on 200 random partitions under both the K-means and hierarchical clustering. Simple clustering approaches yield mean prediction errors 0.166 and 0.222; The TGDR approach has mean prediction error 0.149 and the LASSO yields prediction error 0.170. The CTGDR also has better prediction performance than the SMRC in Ma and Huang (2005, mean classification error 0.14), the boosting (Dettling and Buhlmann 2003, mean classification error: LogitBoost 0.16; AdaBoost 0.18), the classification tree (Dettling and Buhlmann 2003, mean classification error 0.15) and the SVM (Pochet, et al. 2004, mean classification error 0.18).

For the Nodal data, the mean prediction errors are 0.143 with the CTGDR under both clustering schemes. Simple clustering based approach has less optimal prediction errors 0.228 and 0.177. The TGDR and LASSO mean prediction errors are 0.147 and 0.222, respectively. Applying the SMRC approach as in Ma and Huang (2005), we get mean prediction error 0.147. The prediction performance of the CTGDR is better than the boosting based approaches (mean classification errors 0.184, 0.265 and 0.224) and 1-nearest neighbor (mean classification error 0.367) and classification tree (mean classification error 0.204). See details in Dettling and Buhlmann (2003).

5 SURVIVAL ANALYSIS

Follicular Lymphoma data. Follicular lymphoma is the second most common form of non-Hodgkin's lymphoma, accounting for about 22 percent of all cases. A study was conducted to determine whether the survival probability of patients with follicular lymphoma can be predicted by the gene-expression profiles of the tumors at diagnosis (Dave et al. 2004). Fresh-frozen tumor-biopsy specimens and clinical data from 191 untreated patients who had received a diagnosis of follicular lymphoma between 1974 and 2001 were obtained. The median age at diagnosis was 51 years (range 23 to 81), and the median follow up time was 6.6 years (range less than 1.0 to 28.2). The median follow up time among patients alive at last follow up was 8.1 years. Eight records with missing survival information are excluded from the downstream analysis. Affymetrix U133A and U133B microarray genechips were used to measure gene expression levels from RNA samples. A log2 transformation was applied to the Affymetrix measurements. We first filter the 44928 gene measurements with the following criteria: (1) the max expression value of each gene across 191 samples must be greater than 9.186 (the median of the maximums of all probes). (2) the max-min should be greater than 3.874 (the median of the max-min of all probes). (3) Compute correlation coefficients of the uncensored survival times with gene expressions. Select the genes whose correlations with survival time are greater than 0.2. There are 729 genes that pass this screening process. We normalize genes across samples to have mean 0 and variance 1.

Mantel Cell Lymphoma data. Rosenwald et al. (2003) reported a study using microarray expression analysis in mantle cell lymphoma (MCL). Among 101 untreated patients with no history of previous lymphoma included in this study, 92 were classified as having MCL, based on established morphologic and immunophenotypic criteria. Survival times of 64 patients were available and other 28 patients were censored. The median survival time was 2.8 years (range 0.02 to 14.05 years). Lymphochip DNA microarrays (Alizadeh et al., 2000) were used to quantify mRNA

Table 2. Follicular and MCL data. Tuning: values of tuning parameters; Nonzero: number of selected genes; Clus.: number of selected clusters; Logrank: median of the Logrank test statistics.

Approach	Tuning	Nonzero	Clus.	Logrank
		Follicular		
K-means-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	111	13	4.685
K-means-simple	–	729	34	1.175
Hierarchical-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	71	1	4.204
Hierarchical-simple	–	729	24	1.146
TGDR	$\tau = 1.0$	42	–	3.836
LASSO	$u = 0.8$	14	–	1.509
		MCL		
K-means-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	85	8	9.587
K-means-simple	–	834	30	2.837
Hierarchical-CTGDR	$(\tau_1, \tau_2) = (1.0, 1.0)$	65	1	9.036
Hierarchical-simple	–	834	13	7.284
TGDR	$\tau = 1.0$	39	–	9.151
LASSO	$u = 1.7$	23	–	4.671

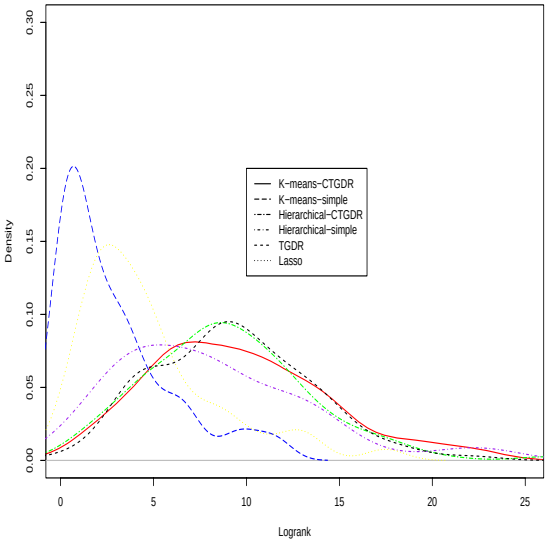

Fig. 2. MCL data. Kernel-smoothed density estimates of the predictive Logrank statistics based on 200 random partitions as described in Section 5 for different approaches.

expression in the lymphoma samples from the 92 patients. The gene expression data that contains expression values of 8810 cDNA elements is available at *http://llmpp.nih.gov/MCL*. We pre-process the data as follows to exclude noises and gain further stability: (1) Compute the variances of all gene expressions; (2) Compute correlation coefficients of the uncensored survival times with gene expressions; and (3) Select the genes with variances larger than the first quartile and with correlation coefficients larger than 0.25. 834 out of 8810 genes pass the above initial screening. We standardize these genes to have zero mean and unit variance.

For the Follicular data, the Gap statistic yields optimal number of clusters equal to 34 (K-means) and 24 (hierarchical), respectively. Plot similar to Figure 1 can be generated and is omitted here. We show model features with cross validation selected tuning parameters in Table 2. Under the hierarchical clustering, the largest cluster has 347 genes and all genes with nonzero coefficients in the CTGDR model come from this cluster. For evaluation, we compute the Logrank test statistics from random partitions. For comparison, we also consider three alternatives: the simple clustering based approach as in classification study, the simple TGDR as described in section 3.1, and the LASSO approach as in Gui and Li (2005a). We can see from Table 2 that the CTGDR coupled with K-means or hierarchical clustering have the best prediction performance. The K-means-CTGDR has slightly better prediction than the Hierarchical-CTGDR, but the difference is not dramatic.

For the MCL data, the optimal number of clusters are 30 and 13 under K-means and hierarchical clustering respectively. Model features with optimal tuning parameters are also shown in Table 2. One dominating cluster is also present under the hierarchical method, which leads to all genes identified by CTGDR coming from this cluster. In Figure 2, we show the kernel-smoothed histograms of the predictive Logrank statistics calculated from random partitions as described in Section 3.5 for different approaches. The K-means-CTGDR method has the best prediction performance in terms of the predictive Logrank statistic. However, the TGDR is slightly better than the hierarchical-CTGDR, but the difference is very small.

6 DISCUSSIONS

The proposed CTGDR approach can carry out feature selection at the cluster and individual gene levels simultaneously, and directly accounts for cluster structures in microarray gene expression data. This algorithm is quite flexible in that it can use any clustering results, including those based on gene annotation, as input in the analysis. We use logistic regression for classification and Cox model for survival data as examples to illustrate the effectiveness of the CTGDR. However, the CTGDR algorithm does not depend on the actual form of the objective function, as long as it is well defined and differentiable. So the CTGDR can be used in survival analysis with other models such as the accelerated failure time and additive hazards models, and classification analysis based other objective functions such as the SVM hinge loss and the ROC objective function.

We have demonstrated the proposed approach on four publicly available datasets. In these examples, there do not exist well defined biological clusters. So we constructed the clusters using two popular approaches: the K-means (a top-down approach) and the hierarchical (bottom-up) methods. We used the Gap statistic to determine the optimal number of clusters. For the four datasets we considered, comparing to several existing methods, the CTGDR has better prediction performance by simultaneously accounting for the cluster structure and carrying out two-level feature selection.

We note that there exist quite a few alternative clustering approaches, for example the self-organizing map. However there exists no optimal clustering method. We only demonstrate the two popular ones in this study. It is possible that other clustering approaches or other ways of determining optimal number of clusters can also lead to models with satisfactory prediction. Comparison of different clustering schemes is beyond the scope of this paper.

When there exist well defined biological clusters (e.g., biological pathways), the proposed CTGDR can utilize that information.

The essential idea of the CTGDR is to carry out feature selection simultaneously at two levels. Specifically, CTGDR is a combination of TGDR at the cluster level and TGDR at the gene level within cluster. With the same spirit, combinations of different regularization approaches can in fact be considered. For example, we can use TGDR for within cluster selection and LASSO for cluster selection. Considering that there are many variable selection methods, the combinations will have a very long list and it is beyond the scope of the current manuscript to conduct a thorough investigation.

We have only considered classification and survival models in which the outcome variable depends on a simple linear combination of the gene expression data. The CTGDR is applicable to more complicated models which may include nonparametric and nonlinear components. It is also applicable to models with interactions at both the cluster and individual gene levels. Such models would probably be more realistic from a biological standpoint. We plan to consider such issues in future studies.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and two referees for their helpful comments that led to significant improvement of this article. The work of JH is supported in part by grant NIH P30 CA 086862-06.

REFERENCES

ALON, U., BARKAI, N., NOTTERMAN, D., GISH, K., MACK, S. and LEVINE, J. (1999) Broad Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* **96**, 6745–6750.

ALIZADEH, A.A., EISEN M.B., DAVIS R.E., MA C., ET AL. (2000) Distinct types of diffuse large B-Cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

BREIMAN, L., FIREDMAN, J.H., OLSHEN, R.A., and STONE, C. (1984): Classification and Regression Trees. Wadsworth Inc. Monterey, California, U.S.A.

CLARE, A. and KING, R.D. (2002). How well do we understand the clusters found in microarray data? *In Silico Biology* **0046**.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society, Series B*, **34**: 187–220.

DAVE, S.S., WRIGHT, G., TAN, B. ET AL. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England Journal of Medicine* **351** 2159–2169.

DETTLING, M. and BUHLMANN, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* **9**, 1061–1069.

DUDOIT, S., FRIDYLAND, J.F. and SPEED, T.P. (2002) Comparison of discrimination methods for tumor classification based on microarray data. *JASA* **97**, 77–87.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**: 14863–14868.

FRIEDMAN, J.H. (2001): Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**: 1189–1232.

FRIEDMAN, J.H. and POPESCU, B.E. (2004). Gradient directed regularization for linear regression and classification. *Technical report, Department of Statistics, Stanford University.*.

GARBER ME, TROYANSKAYA OG, SCHLUENS K, PETERSEN S, THAESLER Z, PACYNA-GENGELBACH M, VAN DE RIJN M, ROSEN GD, PEROU CM, WHYTE RI, ALTMAN RB, BROWN PO, BOTSTEIN D and PETERSEN I. (2001). Diversity of gene expression in adenocarcinoma of the lung. *PNAS* **98**, 13784–13789.

GOLUB, G. and VAN LOAN, C. (1996). *Matrix Computations.* Johns Hopkins Univ Press, Baltimore.

GORDON, A. (1999) *Classification*. Chapman and Hall.

GUI, J. and LI, H. (2005a) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21** 3001–3008.

GUI, J. and LI, H. (2005b) Threshold gradient descent method for censored data regression with applications in pharmacogenomics. *Proceedings of Pacific Symposium on Biocomputing 2005*.

HARRIS, M.A., CLARK, J. IRELAND, A. ET AL. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32** 258-261.

HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics*, **62** 813-820.

JOHNSON, R.A. and WICHERN, D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice-Hall.

MA, S. and HUANG, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, **21** 4356–4362.

MA, S., KOSOROK, M. R. and FINE, J.P. (2006). Additive risk models for survival data with high dimensional covariates. *Biometrics*, **62** 202–210.

MCLACHLAN, G.J., DO, K. and AMBROISE, C. (2004) *Analyzing Microarray Gene Expression Data*. Wiley.

NGUYEN, D. and ROCKE, D.M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**: 39–50.

POCHET, N., DE SMET, F., SUYKENS, J. and DE MOOR, B. (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **17**, 3185–3195.

ROSENWALD, A. WRIGHT, G., WIESTNER, A., CHAN, W. C., ET AL. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185–197.

SPANG, R. BLANCHETTE, C., ZUZAN, H., MARKS, J., NEVINS, J. and WEST, M. (2001) Prediction and uncertainty in the analysis of gene expression profiles. *Proceedings of the German Conference on Bioinformatics GCB 2001*.

TAVAZOIE, S., HUGHES, J., CAMPBELL, M., CHO, R. and CHURCH, G. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22** 281-285.

TAMAYO, P., SLONIM, T., MESIROV, J., ZHU, Q., KITAREEWAN, S. and DMITROVSKY, E. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *PNAS*, **96** 2907-2912.

TIBSHIRANI, R., HASTIE, T., EISEN, M, ROSS, D., BOTSTEIN, D. and BROWN, P. (1999) Clustering methods for the analysis of DNA microarray data. *Manuscript.*

WAHBA, G. (1990) *Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

WEI, Z. and Li, H. (2006). Nonparametric pathway-based regression models for analysis of genomic data. *University of Pennsylvania Biostatistics Working Papers, Year 2006, Paper 6.*

WEST, M. BLANCHETTE, C., DRESSMNA, H., HUANG, E., ISHIDA., S., SPANG, R., ZUZAN, H., OLSON, J., MARKS, J. and NEVINS, J. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS* **98**, 11562–11467.

YEUNG, K.Y., HAYNOR, D. and RUZZO, W. (2001) Validating clustering for gene expression data. *Bioinformatics* **17** 309–31.