

Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data

Chenlei Leng*

Department of Statistics and Applied Probability, National University of Singapore, 117546, Republic of Singapore

Abstract

Gene expression data sets hold the promise to provide cancer diagnosis on the molecular level. However, using all the gene profiles for diagnosis may be suboptimal. Detection of the molecular signatures not only reduces the number of genes needed for discrimination purposes, but may elucidate the roles they play in the biological processes. Therefore, a central part of diagnosis is to detect a small set of tumor biomarkers which can be used for accurate multiclass cancer classification. This task calls for effective multiclass classifiers with build-in biomarker selection mechanism.

We propose the sparse optimal scoring (SOS) method for multiclass cancer characterization. SOS is a simple prototype classifier based on linear discriminant analysis, in which predictive biomarkers can be automatically determined together with accurate classification. Thus, SOS differentiates itself from many other commonly used classifiers, where gene preselection must be applied before classification. We obtain satisfactory performance while applying SOS to several public data sets.

Supplementary materials for this paper are available from <http://www.stat.nus.edu.sg/~stalc/SOS>.

Key words: Microarray data analysis; Biomarker detection; Multiclass classification.

1 Background

Traditional diagnosis of cancerous malignancies relies on subjective interpretation of clinical information. However, such diagnosis may fail due to atypical

* Tel: (65) 65164462. Fax: (65) 68723919.

Email address: stalc@nus.edu.sg (Chenlei Leng).

tumors, incomplete clinical information and other difficulties associated with clinical diagnosis. The advent of large scale genomic technology enables us to study thousands of biomarkers together (Golub et al., 2002; Alizadeh et al., 2000; Pomeroy et al., 2002). In particular, gene expression profiles based on microarray technology promises to offer precise, objective and systematic human cancer classification.

On the other hand, microarray data sets also bring new challenges. Microarray experiments generate data sets with thousands of genes on up to a few hundred of arrays. Statistically speaking, such data sets are characterized by the large dimensionality compared to the sample size. Thus, dimension reduction or variable selection is needed. Biologically speaking, although the number of genes arrayed is large, it is reasonable to expect that only a small subset of biomarkers are associated with diseases. It is clear that solutions to accurate diagnosis and biomarker selection have important biomedical implications.

The problem of multiclass cancer classification using microarray data has been extensively studied. Ramaswamy et al. (2001) proposed to use the support vector machine (SVM) algorithm by using a binary SVM in a one-versus-all scheme. They concluded that their algorithm needs all the genes for accurate classification for their data set. Lee et al. (2004) and Lee and Lee (2003) proposed the multiclass SVM by treating all classes simultaneously and argued that one-versus-all comparison may be suboptimal in certain situations. However, a separate gene selection method has to be applied before applying their method. To remedy the problem, more recently, Koo et al. (2006) proposed structured polychotomous machine by introducing an analysis of variance decomposition to SVM. Tibshirani et al. (2002) proposed the nearest shrunken centroid (NSC) method which effectively shrinks class centroids toward the origin. Feature selection is achieved by shrinking the centroids of some genes to exactly zero. There are also other dimensionality reduction methods for multiclass classification. For example, the partial least squares method was studied by Nguyen and Rocke (2002). Dettling (2004) combined BagBoosting algorithm by combining bagging and boosting. Liu et al. (2005) combined genetic algorithm and all paired SVM methods for multiclass cancer categorization. Davis et al. (2006) addressed the issue of classification and gene selection by pairwise combining various biomarker selection methods and classifiers.

We propose a novel multiclass classifier termed the sparse optimal scoring (SOS) classifier for multiclass classification in this paper. The proposed SOS approach can be seen as a natural extension of the optimal scoring method (Hastie et al., 1994). Thus, SOS is based on Fisher's linear discriminant analysis (LDA), one of the most popular techniques for classification. As will be seen later, SOS is a simple prototype yet accurate classifier. It operates by extracting discriminant variables (features) which best separate

the class centroids and then assign samples to the nearest centroid, according to Mahalanobis distance. However, SOS differs from optimal scoring in one fundamental aspect. SOS seeks to extract features which are sparse linear combinations of a selected subset of all the genes. Thus, multiple tumor types can be classified according to a small set of the genes. Furthermore, SOS retains the virtue of LDA by permitting low dimensional projections of data. Operationally, SOS is realized by penalizing the norm of individual variables and thus is closely related to the Lasso (Tibshirani, 1996) for two-category classification and its grouped version Yuan and Lin (2006) for multi-category classification. Computationally, we apply an efficient path following algorithm in a stepwise manner. The optimal step is chosen by V-fold cross validation, a well studied statistical evaluation tool. We applied SOS to three public data sets and obtained satisfactory performance.

2 Results and Discussion

This section presents the classification performance of SOS on three publicly available data sets.

2.1 Datasets

The GCM datasets compiled by Ramaswamy et al. (2001) contains the expression profiles of 198 tumor samples representing 14 common human cancer classes. The raw data is available from <http://www-genome.wi.mit.edu/MPR/GCM.html>. We focus on 190 tumor samples after excluding 8 metastatic samples. Data preprocessing was done according to Ramaswamy et al. (2001). More specifically, genes were excluded if they exhibited less than 5-fold and 500 units absolute variation across the data set after a threshold of 20 units, ceiling of 16,000 units. Of 16,063 biomarkers considered, 11,319 passed this filter and were used for further classification. Base 10 logarithmic transformation was further applied to each expression. Finally, we standardize each array to mean 0 and variance 1 according to Dudoit et al. (2002).

Our second dataset is the Brown data set Munagala et al. (2004). The Brown data set is publicly available from http://microarray-pubs.stanford.edu/margin_clus/. A detailed account of the data can be found at <http://cmgm.stanford.edu/pbrown>. The data set consists of 268 primary tumor samples representing 18 tissue types and 104 normal samples. There are in total 7452 genes. For this data set, we focus on a ten class classification problem by distinguishing the normal tissues and the following 9 tumor types: breast, central nerve system, kidney, lung, ovary, pancreas, prostate, soft tissue and stomach. The other

tumor types were excluded due to their small sizes. The total number of arrays for this reduced data set is 328. Data preprocessing was done according to Munagala et al. (2004). with missing expression values imputed by 10 nearest-neighbor method. Finally arrays were standardized to mean 0 and variance 1. The Brown data set is interesting since a normal sample is also available.

The third data set is the small round blue cell tumor obtained from SRBCT study detailed in Khan et al. (2001). There are 2308 genes for a total of 83 tumor samples in four tumor classes. This dataset is available from <http://research.nhgri.nih.gov/microarray/Supplement/>. The data set is standardized to zero mean and unit variance for each array.

2.2 Comparison Methods

For comparison purposes, we include several other classifiers. They include support vector machines (SVM), random forest (RF, Breiman, 2001), diagonal linear discriminant analysis (DLDA) and diagonal quadratic discriminant analysis (DQDA) from Dudoit et al. (2002). Also included is the nearest centroid classifier (NSC) by Tibshirani et al. (2002). Among these classifiers, only NSC possesses automatic feature selection capability.

SOS incorporates multiclass feature selection and hence does not depend crucially on preliminary gene selection. However, in order to reduce computational load, we preselect the top 1000 genes according to the ratio of between classes sum of squares to within classes sum of squares (Dudoit et al., 2002). For SVM, DLDA, DQDA and RF, we also tried 200, 400, 800 preselected genes. Using fewer numbers of genes gave similar or even inferior results. Due to page limit, only the results with 1000 genes were presented.

To assess the performance of various classifiers, we randomly partitioned the data into a balanced training set comprising two-thirds of the arrays and a test set with one-third of the arrays. For each training set, gene preselection was done without using the test set. Then, the classifiers were obtained using the training data set. They were applied to the test set to obtain classification performance. For a reliable evaluation, this process was done 50 times.

2.3 GCM Dataset

The GCM dataset consists of 190 tumor samples (breast: 11, prostate: 10, lung: 11, colorectal: 11, lymphoma: 22, bladder: 11, melanoma: 10, uterus: 10, leukemia: 30, renal: 11, pancreas: 11, ovary: 11, mesothelioma: 11, CNS: 20). We applied the six multiclassifiers to this GCM data. The average

test errors were reported in Table 1. SOS outperforms other methods with the smallest average test error (20.25%) The next best classifier is random forest with an average error rate 24.32%. NSC comes next with an error rate 25.62%. SVM has an average error rate 26.67%, slightly larger than NSC. The two simple classifiers DLDA and DQDA do not perform well for this data set, with error rates 27.30% and 30.92%, respectively. Ramaswamy et al. (2001) analyzed the same dataset via the SVM and they achieved 22.0% error rate on a pre-defined training set and a test set.

Table 1 here.

Among the six classifiers compared, only SOS and NSC have build-in gene selection mechanism. The average numbers of chosen biomarkers were shown in Table 2. SOS tends to select much smaller sets of biomarkers (78.24 on the average) than NSC (985.24) and yet has smaller misclassification error. Tibshirani and Hastie (2006) studied a similar data set and concluded that NSC needed more than 4000 genes to achieve the optimal classification performance.

Table 2 here.

We then applied SOS to the whole dataset with 190 tumor samples and chose λ by 10-fold cross validation. This procedure resulted in 143 biomarkers which correspond to the minimum misclassification error zero. SOS finds $G - 1 = 13$ feature directions and this provides a geometrical view of the distributions of the samples. In Figure 1, we plotted two-dimensional projections of the samples produced by SOS. It is shown that the first two feature directions can distinguish four cancer types (Colorectal, leukemia, ovary and CNS) from the rest ten cancer types. Feature sets 5 and 6 separate six cancer types (lung, melanoma, mesothelioma, pancreas, prostate, uterus) from other cancers. Finally, feature sets 11 and 12 discriminate three cancer types (bladder, breast, renal) from other cancer types. These three plots can clearly separate 14 tumor types.

Figure 1 here.

The heatmap of the chosen 143 genes was plotted in Figure 2. A distinct pattern for each tumor class is obvious. The first thirty genes chosen by SOS for the whole data set are shown in Table 3. Some of the genes are biologically meaningful. For example, the first gene in the list (X16663) encodes the hematopoietic cell specific protein, which is an intracellular signaling protein that specifically expressed in blood-forming cells. This gene has distinctively high expressions in lymphoma and leukemia tumor samples but not other non-blood-cell relevant tumor types (Fig. 2). On the other hand, genes with little-known functions which have unique expression patterns may facilitate our understanding of important biological pathways. For example, U57911,

which is found prominently expressed in fetus brain and minimally expressed in other fetus tissue Schwart et al. (1994) has a high expression level in colorectal tumor samples. This suggests that the encoded protein may be involved in other functions besides the nervous system development.

Figure 2 here.

Table 3 here.

2.4 *Brown dataset*

The preprocessed Brown data set consists of 328 samples (breast: 22, CNS: 30, kidney: 36, lung: 47, normal: 104, ovary: 21, pancreas: 7, Prostate: 12, soft tissue: 31, stomach: 18). A similar random partition for GCM data was conducted and the average misclassification errors were summarized in Table 1. For this data set, SVM, SOS and random forest gave similar error rates. The average numbers of genes chosen by SOS and NSC are shown in Table 2. NSC seems to need all the 1000 genes for classification. In a separate study, Tibshirani and Hastie (2007) showed that for the Brown dataset, NSC needed about 4000 genes for the best classification performance. This is consistent with our results.

The SOS chose 209 genes for this data set while λ is chosen by 10-fold cross validation (the heatmap can be found in online supplementary materials). Again, we projected the samples to selected two dimensional spaces and the plots are shown in Figure 3. We see clearly that extracted feature 1 and 2 separate breast, CNS and soft tissue tumors from the others, feature 3 and 4 distinguish ovary and kidney cancer from the others, feature 5 and 6 divide lung from the other types, and finally, feature 7 and 8 separate pancreas, prostate and stomach tumors from the others. More interestingly, the normal tissues do not form a distinctive group in any of the two-dimensional projections, demonstrating the difficulty of isolating normal tissues from multiple tumor tissues. Overall, when SOS is applied to the whole data set, there is one misclassified instance.

Figure 3 here.

2.5 *SRBCT*

This dataset has four tumor classes: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). From Table 1, we saw that SOS (1.26%) and the random forest (1.29%) have

smaller average misclassification errors. In comparing the number of chosen genes, we found that SOS tends to choose fewer genes than NSC, as shown in Table 2. When applied to the whole data set, SOS chose 17 genes. The corresponding heatmap and the low-dimensional projections of the four classes were provided in the supplementary materials. An obvious separation of the 4 tumor types can be clearly seen.

3 Discussion

Accurate classification and biomarker detection remain the two main tasks of multicategory microarray data analysis. We have proposed a new multicategory classifier SOS and have demonstrated its excellent performance in classifying multiple multiple types, in addition to gene selection capability. Particularly, the proposed SOS achieves smaller or similar classification errors with fewer genes compared to other existing classifiers. Our results are in marked contrast to the SVM method described in Ramaswamy et al. (2001), where all the genes were needed to ensure the optimal performance.

A salient feature of SOS is that it treat all classes simultaneously by mapping a class label to a vector of $G - 1$ real numbers. Treating all the classes simultaneously may potentially bring computational efficacy since for the data sets we considered, one-versus-one or one-versus-all scheme incurs heavy computational burden. Furthermore, treating all the classes on equal footage may alleviate a masking problem inherited in one-versus-one (one-versus-all) scheme (Lee et al., 2004). To facilitate biomarker selection, we penalize the norm of each gene such that by appropriately choosing the tuning parameter, sparse features represented by linear combinations of a small set of genes can be extracted. Due to its similarity to the LDA, the samples can be easily visualized by projecting them to appropriate low-dimensional subspaces. As clearly demonstrated in Figure 2 and Figure 3, such low-dimensional projections may provide further insight on the separability of various tumor types. Thus, SOS offers easy interpretation of the result.

SOS selects up to n genes for a dataset with sample size n . This does not affect performance on the datasets we analyzed. However, for extremely small data sets, a few genes may not be able to give accurate classification. To handle this situation, we can add another penalty $\lambda_0 \sum_{k=1}^K \|\beta_k\|^2$ to (2), which is essentially an extension of the elastic net method proposed by Zou and Hastie (2005). In current implementation, we allowed SOS to have linear boundaries only. Future research can include the extension of SOS in handling nonlinear boundaries.

4 Methods

4.1 Linear Discriminant Analysis

To establish a connection between optimal scoring and LDA, it is necessary to review LDA first. Suppose that there are n observations, p input variables and G categories. Denote by x_i the feature vector for the i th observation and $g_i \in \mathcal{G} = \{1, \dots, G\}$ with $G \geq 2$ the response. The goal of multicategory classification is to find a function $\Delta : x \rightarrow \mathcal{G}$ such that for a new x , it uniquely assigns a class membership to x . Let n_k be the number of g_i equal to k . Denote the within-class and between-class covariances as Σ_W , Σ_B respectively:

$$\Sigma_W = \frac{1}{n} \sum_{k=1}^G \sum_{y_j=k} (x_j - x^{(k)})(x_j - x^{(k)})^T, \Sigma_B = \frac{1}{n} \sum_{k=1}^G n_k (x^{(k)} - \bar{x})(x^{(k)} - \bar{x})^T,$$

where $\bar{x} = \frac{1}{n} \sum_i x_i$; and $x^{(k)} = \sum_{y_i=k} x_i / n_k$.

The solutions of LDA $a = [a_1, \dots, a_d]$ are obtained sequentially. Given the first $k - 1$ discriminant coordinates a_1, \dots, a_{k-1} , the k th feature is given by $z_k = x^T a_k$ such that

$$a_k = \operatorname{argmax}_a a^T \Sigma_B a, \text{ s.t. } a_k^T \Sigma_W a_j = 0, j < k, \text{ and } a_k^T a_k = 1. \quad (1)$$

The solution to LDA is a generalized eigen problem and can be easily solved. The total number of discriminant variables $d \leq G$ is usually determined by prediction criteria. Geometrically speaking, the LDA in (1) maximizes the separation between classes, relative to the closeness of the observations in the same category.

LDA enjoys several advantages. Firstly, it is a simple prototype classifier with linear decision boundaries. Secondly, it permits natural low-dimensional views of the data by projecting the high dimensional data onto the extracted features. Thirdly, it usually produces the best classification results, see a discussion in Hastie et al. (1994). Thus, SOS allows one to view informative low-dimensional projections of the data via extracted features, similar to LDA. Note that LDA is feasible only for data sets with more samples than features and is not applicable for large genomic data sets.

4.2 Optimal Scoring

LDA can be re-formulated as a regression problem via optimal scoring. See Hastie et al.(1994) for more discussions. Formally, let $\theta : \mathcal{G} \rightarrow R^1$ be a function

that assigns scores to the classes. We can find $K \leq G - 1$ independent scorings for the class label $\{\theta_1, \dots, \theta_K\}$, and K corresponding linear maps $\eta_k(x) = x^T \beta_k$, $k = 1, \dots, K$. And the scores θ_k and the maps β_k are chosen to minimize the average squared residual

$$ASR = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n (\theta_k(g_i) - x_i^T \beta_k)^2.$$

The sequence of LDA vectors a_k are known to be identical to the sequence β_k up to a constant (Mardia et al. 1979). In practice, the linear operator η_k can be replaced by a regularized regression. The real benefit of expressing LDA in regression context is the ability of performing variable selection and model regularization, as pointed out by Hastie et al. (1994).

Create $n \times G$ indicator response matrix Y such that $Y_{ij} = 1$ if the class of the i th sample is $g_i = j$, and 0 otherwise. Hastie et al. (1994) proposed the following algorithm for flexible discriminant analysis:

1. Choose an initial score matrix Θ_0 satisfying $\Theta_0^T D_G \Theta_0 = I$, where $D_G = Y^T Y / n$. Let $\Theta_0^* = Y \Theta_0$;
2. Fit a multiple regression model of Θ_0^* on X , yielding fitted values $\hat{\Theta}_0^*$. Let $\hat{\eta}(X)$ be the vector of fitted functions;
3. Obtain eigenvector matrix Φ of $\Theta_0^{*T} \hat{\Theta}_0^*$; the optimal scores are $\Theta = \Theta_0 \Phi$;
4. Update the final model from step 2 via $\hat{\eta} \leftarrow \Phi^T \hat{\eta}(X)$.

Define D as a diagonal matrix with k th diagonal term

$$D_{kk} = \left\{ \frac{1}{\alpha_k^2 (1 - \alpha_k^2)} \right\}^{1/2},$$

where α_k is the k th largest eigenvalue calculated in step 3. The decision rule for a new x has the form of a weighted nearest centroid rule, which assigns x to class j that minimizes

$$\delta(x, j) = \|D(\eta(x) - \bar{\eta}^j)\|^2,$$

where $\bar{\eta}^j = \sum_{g_i=j} \eta(x_i) / n_j$ denotes the fitted centroid of the j th class.

Hastie et al. (1994) showed that by recasting LDA as a regression problem, the performance of LDA can be improved by regularized optimal scoring method.

In order to handle large scale genomic datasets, Ghosh (2003) was among the first to study a number of dimensional reduction techniques for optimal scoring, including principle component regression, partial least square regression and ridge regression. All the aforementioned methods lack a build-in variable selection mechanism. Ghosh (2003) suggested to select relevant biomarkers by

examining the fitted regression coefficients in step 2. However, large fitted coefficients does not necessarily translate into important variables. Furthermore, motivated by the ROC considerations, Ghosh and Chinnaiyan (2005) studied the two-category classification using Lasso. They commented that their method cannot handle multicategorical responses in the current context.

4.3 Sparse Optimal Scoring

SOS explores a principled variable selection via the LASSO type of penalty within the optimal scoring setting. In doing so, selecting variables and finding separating boundaries are jointly realized.

Writing $\beta = (\beta_1, \dots, \beta_K)$ and denoting rows of β as $\beta_{(j)}$, we propose to replace step 2 of the algorithm by a penalized regression

$$\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n (\theta_k(g_i) - x_i^T \beta_k)^2 + \lambda \sum_{j=1}^p \|\beta_{(j)}\|, \quad (2)$$

where $\|a\|$ is the l_2 norm of a . We see immediately that it reduces to the LASSO (Tibshirnai, 1996) for $K = 1$, and it reduces to the group LASSO (Yuan and Lin, 2006) when $K > 1$. The l_2 norm has the property that if λ is appropriately chosen, $\beta_{(j)}$ may be estimated exactly a zero vector. Thus, biomarker detection is achieved. Furthermore, this penalty is particularly attractive due to its simplicity and ease of implementation.

4.4 Algorithm

We outline here an efficient lars algorithm for solving (2), adopting a similar algorithm in Yuan and Lin (2006). For this purpose, we rewrite (2) as

$$\frac{1}{n} \sum_{i=1}^n \|\Theta^* - X\beta\|^2 + \lambda \sum_{j=1}^p \|\beta_{(j)}\|, \quad (3)$$

where Θ^* is a $n \times K$ matrix. Our algorithm can be viewed as an extension of the lasso to multivariate responses.

Starting with β as a zero matrix, we seek a variable (say X_{j_1}) that minimizes the angle with Θ^* (i.e., $\|X_{j_1} \Theta^*\|$ is the largest). The algorithm marches in the direction of the projection of Θ^* on X_{j_1} until some other variable (say X_{j_2}) has as small an angle with the current residual r , i.e.

$$\|X'_{j_1} r\| = \|X'_{j_2} r\|.$$

From here the algorithm proceeds in the direction of the projection of the residual onto the space spanned by X_{j_1} and X_{j_2} . This process is repeated until either all the variables are included or the residual r becomes a zero matrix. The algorithm can be summarized as following

SOS algorithm

1. Begin with $\beta^{[0]} = 0$, $r^{[0]} = \Theta^*$ and $s = 1$.
2. Compute the current most-correlated set

$$\mathcal{A}_s = \arg \max_j \|X'_j r^{[s-1]}\|,$$

and calculate the current direction γ with

$$\gamma_{\mathcal{A}_s} = (X'_{\mathcal{A}_k} X_{\mathcal{A}_s})^{-1} X'_{\mathcal{A}_s} r^{[s-1]}$$

and $\gamma_{\mathcal{A}^c} = 0$. Here $X_{\mathcal{A}_s}$ is the matrix with columns of X corresponding to \mathcal{A}_s .

3. For every $j \notin \mathcal{A}_s$, compute how far the algorithm will proceed in the direction γ before X_j enters the most correlated set. This can be measured by an $a_j \in [0, 1]$ such that

$$\|X'_j(r^{[k-1]} - a_j X \gamma)\|^2 = \|X'_{j'}(r^{[k-1]} - a_j X \gamma)\|,$$

where j' is arbitrarily from \mathcal{A}_s .

4. If $\mathcal{A}_s \neq \{1, \dots, p\}$, let $a = \min_{j \notin \mathcal{A}_s} \{a_j\} \equiv a_{j^*}$ and update $\mathcal{A}_{s+1} = \mathcal{A}_s \cup \{j^*\}$; otherwise set $a = 1$.
5. Update $\beta^{[s]} = \beta^{[s-1]} + a\gamma$, $r^{[s]} = \Theta^* - X\beta^{[s]}$ and $s = s + 1$. If $r^{[s]}$ is not a zero matrix, go back to step 2 until $a = 1$; otherwise stop.

Roughly speaking, SOS algorithm searches for genes in a sequential manner by locating at one step one gene which gives better separation among the classes.

4.5 Tuning

To select the tuning parameter λ , we use V fold cross validation to minimize misclassification error. More precisely, for a fixed V , we randomly split the data into V parts S_1, \dots, S_V which are roughly equal sized. The optimal λ is chosen to minimize the cross validated misclassification criterion

$$CV(\lambda) = \sum_{v=1}^V \left\{ \sum_{i \in S_v} g_i \neq \hat{g}_i^{(-v)} \right\},$$

where $\hat{g}_i^{(-v)}$ is the SOS estimate of g_i based on the data without the v th subset. In this paper, we chose to use $V = 10$.

5 Acknowledgements

Leng's research is partially supported by NUS research grant R-155-050-053-133. The author is grateful to Hui Yu for helpful discussions.

6 Figures

6.1 *Figure 1 - GCM data set*

Selected two-dimensional SOS estimated projections of the data set. File name: ram.eps.

6.2 *Figure 2 - Heatmap for GCM data*

The heatmap of the chosen 143 genes for GCM dataset. Each row corresponds to a single gene. Each column corresponds to a single array. Note that genes are clustered by hierachical clustering. File name: hm.ram.1.eps.

6.3 *Figure 3 - Brown data set*

Selected two-dimensional SOS estimated projections of the data set. File name: mun.eps.

7 Tables

7.1 *Table 1 - Miscallsification errors*

Average misclassification errors over 50 random partitions for GCM dataset, Brown dataset and SRBCT dataset. The standard errors are in parentheses.

Dataset	SOS	NSC	DLDA	DQDA	SVM	RF
	(%)	(%)	(%)	(%)	(%)	(%)
GCM	20.25	25.62	27.30	30.92	26.67	24.32
	(4.38)	(4.84)	(4.08)	(5.28)	(3.78)	(3.85)
Brown	9.83	16.22	16.15	14.28	9.56	10.86
	(2.42)	(2.68)	(2.82)	(3.03)	(2.13)	(2.41)
SRBCT	1.26	1.93	6.28	3.79	6.79	1.29
	(2.52)	(2.52)	(6.51)	(4.11)	(5.27)	(3.21)

7.2 Table 2 - The number of chosen genes

Average numbers of genes selected over 50 random partitions for GCM dataset, Brown dataset and SRBCT dataset. The standard errors are in parentheses.

	SOS	NSC
GCM	78.24	985.20
	(14.73)	(41.90)
Brown	151.88	1000
	(17.12)	(0)
SRBCT	25.88	148.16
	(9.22)	(168.56)

7.3 Table 3 - The top 30 genes for GCM data set

Gene Name	Gene Description
X16663_at	HEMATOPOIETIC LINEAGE CELL SPECIFIC PROTEIN
W56613_at	SH3 binding protein
U58658_at	Unknown protein mRNA within the p53 intron 1
M88163_at	SNF2L1 SNF2 (sucrose nonfermenting, yeast, homolog)-like 1
D86976_at	KIAA0223 gene, partial cds
U66838_at	Cyclin A1 mRNA
U57911_at	Fetal brain (239FB) mRNA, from the WAGR region
AA405288_at	EST: zt37f09.r1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 724553 5' similar to contains Alu repetitive element; contains element LTR5 repetitive element ;, mRNA sequence.
AA174173_at	EST: PTH156 HTC1DL1 Homo sapiens cDNA 5'/3', mRNA sequence. (from Genbank)
U79293_at	Clone 23948 mRNA sequence
X63522_s_at	RETINOIC ACID RECEPTOR RXR-BETA
L00354_at	PROCHOLECYSTOKININ PRECURSOR
U62531_at	SLC4A2 Solute carrier family 4, anion exchanger, member 2 (erythrocyte membrane protein band 3-like 1)
L35269_at	ZINC FINGER PROTEIN 35
X15573_at	PFKL Phosphofructokinase (liver type)
AA313786_at	EST: EST185649 Colon carcinoma (HCC) cell line Homo sapiens cDNA 5' end, mRNA sequence. (from Genbank)
S82185_at	Escherichia coli unknown mRNA
U09848_at	ZNF139 Zinc finger protein 139 (clone pHZ-37)
L05568_at	SLC6A4 Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4
M27543_at	GNAI1 Alternative guanine nucleotide-binding regulatory protein (G) alpha-inhibitory-subunit
U93205_at	Nuclear chloride ion channel protein (NCC27) mRNA
M30269_at	NID Nidogen (enactin)
U48936_at	Amiloride-sensitive epithelial sodium channel gamma subunit mRNA, 5' end, partial cds
U16811_s_at	Bak protein mRNA
U79257_at-2	Human clone 23932 mRNA sequence

References

- [1] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, G. Yang, M. Land, T. Moore, J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, L. Staudt, Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, 403 (2000), 503–511.
- [2] L. Breiman, Random forests, *Machine Learning*, 45 (2001), 5–32.
- [3] C. Davis, F. Gerick, V. Hintermair, C. Friedel, K. Fundel, R. Kffner, R. Zimmer, Reliable gene signatures for microarray classification: assessment of stability and performance, *Bioinformatics*, 22 (2006), 2356–2363.

- [4] M. Dettling, BagBoosting for tumor classification with gene expression data, *Bioinformatics*, 20 (2004), 3583–3593.
- [5] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97 (2002), 77–87.
- [6] T. Golub T, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh M, J. Downing, M. Caligiuri, C. Bloomfield, R. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286 (1999), 531–537.
- [7] D. Ghosh, Penalized discriminant methods for the classification of tumors from microarray experiments, *Biometrics*, 59 (2003), 992–1000.
- [8] D. Ghosh, A. Chinnaiyan, Classification and selection of biomarkers in genomic data using LASSO, *Journal of Biomedecine and Biotechnology*, 2 (2005), 147–154.
- [9] T. Hastie, R. Tibshirani, A. Buja A, Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association*, 89 (1994), 1255–1270.
- [10] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, P. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7 (2001), 673–679.
- [11] J. Koo, I. Sohn, S. Kim, J. Lee, Structured polychotomous machine diagnosis of multiple cancer types using gene expression, *Bioinformatics*, 22 (2006), 950–958.
- [12] Y. Lee Y, C. Lee, Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, 19 (2003), 1132–1139.
- [13] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data, *Journal of the American Statistical Association*, 99 (2004), 67–81.
- [14] J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, X. Ling, Multiclass cancer classification and biomarker discovery using GA-based algorithms, *Bioinformatics*, 21 (2005), 2691–1697.
- [15] K. Mardia, J. Kent, J. Bibby, *Multivariate Analysis*, Academic Press, 1979.
- [16] K. Munagala, R. Tibshirani, P. Brown: Cancer characterization and feature set extraction by discriminative margin clustering, *BMC Bioinformatics*, 5 (2004).
- [17] D. Nguyen, D. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, 18 (2002), 1216–1226.
- [18] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano,

- G. Stolovitzky, D. Louis, J. Mesirov, E. Lander, T. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, (415) 2002, 436–442.
- [19] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, T. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, *PNAS*, 98 (2001), 15149–54.
- [20] F. Schwart, R. Neve, R. Eisenman, M. Gessler, G. Bruns, A WAGR region gene between PAX-6 and FSHB expressed in fetal brain, *Human Genetics*, 94 (1994), 658–664.
- [21] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, 58 (1996), 267–299.
- [22] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, 99 (2002), 6567–6572.
- [23] R. Tibshirani, T. Hastie, Margin trees for high-dimensional classification, *Journal of Machine Learning Research*, 8 (2007), 637–652.
- [24] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, 68 (2006), 49–67.
- [25] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, 67 (2005), 301–320.