Journal of Machine Learning Research 11 (2010) 2141-2167 Submitted 11/09; Revised 4/10; Published 8/10

Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases

Guoqiang Yu YUG@VT.EDU
Yuanjian Feng YJFENG@VT.EDU
Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Arlington, VA 22203, USA

David J. Miller DJMILLER@ENGR.PSU.EDU
Department of Electrical Engineering
The Pennsylvania State University
University Park, PA 16802, U.S.A

Jianhua Xuan XUAN@VT.EDU
Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Arlington, VA 22203, USA

Eric P. Hoffman EHOFFMAN@CNMCRESEARCH.ORG
Research Center for Genetic Medicine
Children's National Medical Center
Washington, DC 20010, USA

Robert Clarke CLARKEL@GEORGETOWN.EDU
Department of Oncology, Physiology and Biophysics
Georgetown University
Washington, DC 20057, USA

Ben Davidson BEN.DAVIDSON@RADIUMHOSPITALET.NO
Department of Pathology
Radiumhospitalet-Rikshospitalet Medical Center
Montebello N-0310 Oslo, Norway

Ie-Ming Shih ISHIH@JHMI.EDU
Departments of Pathology and Gynecology and Oncology
Johns Hopkins Medical Institutions
Baltimore, MD 21231, USA

Yue Wang YUEWANG@VT.EDU
Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Arlington, VA 22203, USA

Editor: Donald Geman

Abstract

Microarray gene expressions provide new opportunities for molecular classification of heterogeneous diseases. Although various reported classification schemes show impressive performance, most existing gene selection methods are suboptimal and are not well-matched to the unique characteristics of the multicategory classification problem. Matched design of the gene selection method and a committee classifier is needed for identifying a small set of gene markers that achieve accurate multicategory classification while being both statistically reproducible and biologically plausible. We report a simpler and yet more accurate strategy than previous works for multicategory classification of heterogeneous diseases. Our method selects the union of one-versus-everyone (OVE) *phenotypic up-regulated* genes (PUGs) and matches this gene selection with a one-versus-rest support vector machine (OVRSVM). Our approach provides even-handed gene resources for discriminating both neighboring and well-separated classes. Consistent with the OVRSVM structure, we evaluated the fold changes of OVE gene expressions and found that only a small number of high-ranked genes were required to achieve superior accuracy for multicategory classification. We tested the proposed PUG-OVRSVM method on six real microarray gene expression data sets (five public benchmarks and one in-house data set) and two simulation data sets, observing significantly improved performance with lower error rates, fewer marker genes, and higher performance sustainability, as compared to several widely-adopted gene selection and classification methods. The MATLAB toolbox, experiment data and supplement files are available at http://www.cbil.ece.vt.edu/software.htm.

Keywords: microarray gene expression, multiclass gene selection, phenotypic up-regulated gene, multicategory classification

1. Background

The rapid development of gene expression microarrays provides an opportunity to take a genome-wide approach for disease diagnosis, prognosis, and prediction of therapeutic responsiveness (Clarke et al., 2008; Wang et al., 2008). When the molecular signature is analyzed with pattern recognition algorithms, new classes of disease are identified and new insights into disease mechanisms and diagnostic or therapeutic targets emerge (Clarke et al., 2008). For example, many studies demonstrate that global gene expression profiling of human tumors can provide molecular classifications that reveal distinct tumor subtypes not evident by traditional histopathological methods (Golub et al., 1999; Ramaswamy et al., 2001; Shedden et al., 2003; Wang et al., 2006).

While molecular classification falls neatly within supervised pattern recognition, high gene dimensionality and paucity of microarray samples pose challenges for, and inspire novel developments in classifier design and gene selection methodologies (Wang et al., 2008). For multicategory classification using gene expression data, various classifiers have been proposed and have achieved promising performance, including k-Nearest Neighbor Rule (kNN) (Golub et al., 1999), artificial neural networks (Wang et al., 2006), Support Vector Machine (SVM) (Ramaswamy et al., 2001), Naïve Bayes Classifier (NBC) (Liu et al., 2002), Weighted Votes (Tibshirani et al., 2002), and Linear Regression (Fort and Lambert-Lacroix, 2005). Many comparative studies show that SVM based classifiers outperform other methods on most bench-mark microarray data sets (Li et al., 2004; Statnikov et al., 2005).

An integral part of classifier design is gene selection, which can improve both classification accuracy and diagnostic economy (Liu et al., 2002; Shi et al., 2008; Wang et al., 2008). Many microarray-based studies suggest that, irrespective of the classification method, gene selection is vital for achieving good generalization performance (Statnikov et al., 2005). For multicategory classification using gene expression data, the criterion function for gene selection should possess high sensitivity and specificity, well match the specific classifiers used, and identify gene markers that are both statistically reproducible and biologically plausible (Shi et al., 2008; Wang et al., 2008). There are limitations associated with existing gene selection methods (Li et al., 2004; Statnikov et al., 2005). While wrapper methods consider joint discrimination power of a gene subset, complex classifiers used in wrapper algorithms for small sample size may overfit, producing non-reproducible gene subsets (Li et al., 2004; Shi et al., 2008). Moreover, discernment of the (biologically plausible) gene interactions retained by wrapper methods is often difficult due to the black-box nature of most classifiers (Shedden et al., 2003).

Conversely, most filtering methods for multicategory classification are straightforward extensions of binary discriminant analysis. These methods are devised without well matching to the classifier that is used, which typically leads to suboptimal classification performance (Statnikov et al., 2005). Popular multicategory filtering methods (which are extensions of two-class methods) include Signal-to-Noise Ratio (SNR) (Dudoit et al., 2002; Golub et al., 1999), Student's t-statistics (Dudoit et al., 2002; Liu et al., 2002), the ratio of Between-groups to Within-groups sum of squares (BW) (Dudoit et al., 2002), and SVM based Recursive Feature Elimination (RFE) (Li and Yang, 2005; Ramaswamy et al., 2001; Zhou and Tuck, 2007). However, as pointed out by Loog et al. (2001) in proposing their weighted Fisher criterion (wFC), simple extensions of binary discriminant analysis to multicategory gene selection are suboptimal because they overemphasize large between-class distances, that is, these methods choose gene subsets that preserve the distances of (already) well-separated classes, without reducing (and possibly with increase in) the large overlap between neighboring classes. This observation and the application of wFC to multicategory classification are further evaluated experimentally by Wang et al. (2006) and Xuan et al. (2007).

The work most closely related to our gene selection scheme is that of Shedden et al. (2003). These investigators focused on marker genes that are highly expressed in one phenotype relative to one or more different phenotypes and proposed a tree-based *one-versus-rest* (OVR) fold change evaluation between mean expression levels. The potential limitation here is that the criterion function considers the “rest of the classes” as a “super class”, and thus may select genes that can distinguish a single class from the remaining super class, yet without giving any benefit in discriminating between classes within the super class. Such genes may compromise multicategory classification accuracy, especially when a small gene subset is chosen.

It is also important to note that, while univariate or multivariate analysis methods using complex criterion functions may reveal subtle marker effects (Cai et al., 2007; Liu et al., 2005; Xuan et al., 2007; Zhou and Tuck, 2007), they are also prone to overfitting. Recent studies have found that for small sample sizes, univariate methods fared comparably to multivariate methods (Lai et al., 2006; Shedden et al., 2003) and simple fold change analysis produced more reproducible marker genes than significance analysis of variance-incorporated t-tests (Shi et al., 2008).

In this paper, we propose matched design of the gene selection mechanism and a committee classifier for multicategory molecular classification using microarray gene expression data. A key feature of our approach is to match a simple *one-versus-everyone* (OVE) gene selection scheme to the OVRSVM committee classifier (Ramaswamy et al., 2001). We focus on marker genes that are highly expressed in one phenotype relative to each of the remaining phenotypes, namely Phenotypic Up-regulated Genes (PUGs). PUGs are identified using the fold change ratio computed between the specified phenotype mean and each of the remaining phenotype means. Thus, we consider a gene to be a marker for the specified phenotype if the average expression associated with this phenotype is high relative to the average expressions in each of the other phenotypes. To assure evenhanded resources for discriminating both neighboring and well-separated classes, we use a fixed number of PUGs for each phenotypic class and pool all phenotype-specific PUGs together to form a gene marker subset used by the OVRSVM committee classifier. All PUGs referenced by the committee classifier are individually interpretable as potential markers for phenotypic classes, allowing each gene to inform the classifier in a way that is consistent with its mechanistic role (Shedden, et al., 2003). Since PUGs are the union of subPUGs selected by simple univariate OVE fold change analysis, they are expected to be statistically reproducible (Lai et al., 2006; Shedden et al., 2003; Shi et al., 2008).

We tested PUG-OVRSVM on five publicly available benchmarks and one in-house microarray gene expression data set and on two simulation data sets, observing significantly improved performance with lower error rates, fewer marker genes, and higher performance stability, as compared to several widely-adopted gene selection and classification methods. The reference gene selection methods are OVRSNR (Golub et al., 1999), OVRt-stat (Liu et al., 2002), pooled BW (Dudoit et al., 2002), and OVRSVM-RFE (Guyon et al., 2002), and the reference classifiers are kNN, NBC, and one-versus-one (OVO) SVM. With accuracy estimated by leave-one-out cross-validation (LOOCV) (Hastie et al., 2001), our experimental results show that PUG-OVRSVM outperforms all combinations of the above referenced gene selection and classification methods in the two simulation data sets and 5 out of the 6 real microarray gene expression data sets, and produces comparable performance on the one remaining data set. Specifically, tested on the widely-used benchmark microarray gene expression data set "multicategory human cancers data" (GCM) (Ramaswamy et al., 2001; Statnikov et al., 2005), PUG-OVRSVM produces a lower error rate of 11.05% (88.95% correct classification rate) than the best known benchmark error rate of 16.72% (83.28% correct classification rate) (Cai et al., 2007; Zhou and Tuck, 2007).

2. Methods

In this section, we first discuss multicategory classification and associated feature selection, with an emphasis on OVRSVM and application to gene selection for the microarray domain. This discussion then naturally leads to our proposed PUG-OVRSVM scheme.

2.1 Maximum a Posteriori Decision Rule

Classification of heterogeneous diseases using gene expression data can be considered a Bayesian hypothesis testing problem (Hastie et al., 2001). Let $\mathbf{x}_i = [x_{i1}, ..., x_{ij}, ..., x_{id}]$ be the real-valued gene expression profile associated with sample i across d genes for $i = 1, \ldots, N$ and $j = 1, \ldots, d$. Assume that the sample points $\mathbf{x}_i$ come from M classes, and denote the class conditional probability density function and class prior probability by $p(\mathbf{x}_i \mid \omega_k)$ and $P(\omega_k)$, respectively, for $k = 1, \ldots, M$. To minimize the Bayes risk averaged over all classes, the optimum classifier uses the well-known maximum *a* posteriori (MAP) decision rule (Hastie et al., 2001). Based on Bayes' rule, the class posterior probability for a given sample $\mathbf{x}_i$ is

$$P(\omega_k \mid \mathbf{x}_i) = \frac{P(\omega_k)\, p(\mathbf{x}_i \mid \omega_k)}{\sum_{k'=1}^{M} P(\omega_{k'})\, p(\mathbf{x}_i \mid \omega_{k'})}$$

and is used to (MAP) classify $\mathbf{x}_i$ to ω_k when

$$P(\omega_k \mid \mathbf{x}_i) > P(\omega_l \mid \mathbf{x}_i) \tag{1}$$

for all $l \neq k$.

2.2 Supervised Learning and Committee Classifiers

Practically, multicategory classification using the MAP decision rule can be approximated using parameterized discriminant functions that are trained by supervised learning. Let $f_k(\mathbf{x}_i,\theta)$, $k = 1,2,\ldots,M$, be the M outputs of a machine classifier designed to discriminate between M classes ($>$2), where θ represents the set of parameters that fully specify the classifier, and with the output values assumed to be in the range $[0,1]$. The desired output of the classifier will be "1" for the class to which the sample belongs and "0" for all other classes. Suppose that the classifier parameters are selected based on a training set so as to minimize the mean squared error (MSE) between the outputs of the classifier and the desired (class target) outputs,

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{M}\sum_{\mathbf{x}_i\in\omega_k}\left[\left[f_k(\mathbf{x}_i,\theta)-1\right]^2+\sum\nolimits_{l\neq k} f_l^2(\mathbf{x}_i,\theta)\right]. \tag{2}$$

Then, it can be shown that the classifier is being trained to approximate the posterior probability for class ω_k given the observed $\mathbf{x}_i$, that is, the classifier outputs will converge to the true posterior class probabilities

$$f_k(\mathbf{x}_i,\theta) \to P(\omega_k \mid \mathbf{x}_i)$$

if we allow the classifier to be arbitrarily complex and if N is made sufficiently large. This result is valid for any classifier trained with the MSE criterion, where the parameters of the classifier are adjusted to simultaneously approximate M discriminant functions $f_k(\mathbf{x}_i,\theta)$ (Gish, 1990).

While there are numerous machine classifiers that can be used to implement the MAP decision rule (1) (Hastie et al., 2001), a simple yet elegant way of discriminating between M classes, and which we adopt here, is based on an OVRSVM committee classifier (Ramaswamy et al., 2001; Rifkin and Klautau, 2002; Statnikov et al., 2005). Intuitively, each term within the sum over k in (2) corresponds to an OVR binary classification problem and can be effectively minimized by suitable training of a binary classifier (discriminating class k from all other classes). By separately minimizing the MSE associated with each term in (2) via binary classifier training and, thus, effectively minimizing the total MSE, a set of discriminant functions $\{f_k(\mathbf{x}_i,\theta_k\subseteq\theta)\}$ can be constructed which, given a new sample point, apply the decision rule (1), but with $f_k(\mathbf{x}_i,\theta)$ playing the role of the posterior probability.

Among the great variety of binary classifiers that use regularization to control the capacity of the function spaces they operate in, the best known example is the SVM (Hastie et al., 2001; Vapnik, 1998). To carry over the advantages of regularization approaches for binary classification tasks to multicategory classification, the OVRSVM committee classifier uses M different SVM binary classifiers, each one separately trained to distinguish the samples in a single class from the samples in all remaining classes. For classifying a new sample point, the M SVMs are run, and the SVM that produces the largest (most positive) output value is chosen as the "winner" (Ramaswamy et al., 2001). For more detailed discussion, see the critical review and experimental comparison by Rifkin and Klautau (2002). Figure 1 shows an illustrative OVRSVM committee classifier for three classes. The OVRSVM committee classifier has proved highly successful at multicategory classification tasks involving finite or limited amounts of high dimensional data in real-world applications. OVRSVM produces results that are often at least as accurate as other more complicated methods including single machine multicategory schemes (Statnikov et al., 2005). Perhaps more importantly for our purposes, the OVR scheme can be matched with an OVE gene selection method, as we elaborate next.

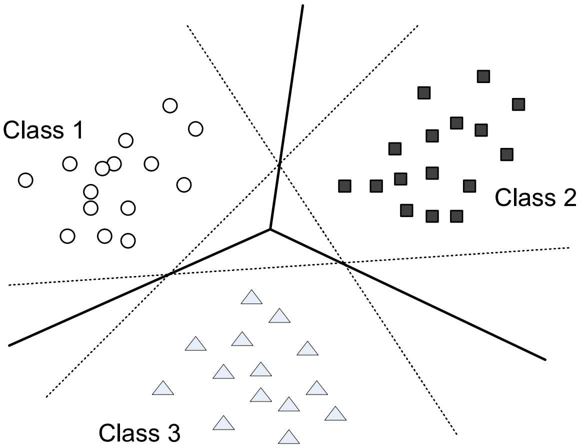

Figure 1: Conceptual illustration of OVR committee classifier for multicategory classification (three classes, in this case). The dotted lines are the decision hyperplanes associated with each of the component binary SVMs and the bold line-set represents the final decision boundary after the winner-take-all classification rule is applied.

2.3 One-Versus-Everyone Fold-change Gene Selection

While gene selection is vital for achieving good generalization performance (Guyon et al., 2002; Statnikov et al., 2005), perhaps even more importantly, the identified genes, if statistically reproducible and biologically plausible, are “markers”, carrying information about the disease phenotype (Wang et al., 2008). We will propose two novel, effective gene selection methods for multicategory classification that are well-matched to OVRSVM committee classifiers, namely, OVR and OVE fold-change analyses.

OVR fold-change based PUG selection follows directly from the OVRSVM scheme. Let N_k be the number of sample points belonging to phenotype k; the geometric mean of the expression levels (on the untransformed scale) for gene j under phenotype k is

$$\mu_j(k) = \sqrt[N_k]{\prod\nolimits_{i\in\omega_k} x_{ij}}$$

$j = 1,\ldots,d$; $k = 1,\ldots,M$. Then, we define the OVRPUGs as:

$$\mathbb{J}_{\text{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\mu_j(k)}{\sqrt[M-1]{\prod\nolimits_{l\neq k}\mu_j(l)}} \geqslant \tau_k \right\} \tag{3}$$

where $\{\tau_k\}$ are pre-defined thresholds chosen so as to select a fixed (equal) number of PUGs for each phenotype k. This PUG selection scheme (3) is similar to what has been previously proposed by Shedden et al. (2003):

$$\mathbb{J}_{\text{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\mu_j(k)}{\sqrt[N-N_k]{\prod\nolimits_{i\notin\omega_k} x_{ij}}} \geqslant \tau_k \right\}. \tag{4}$$

The critical difference between (3) and (4) is that the denominator term in (3) is the overall geometric center of the “geometric centers” associated with each of the remaining phenotypes while the denominator term in (4) is the geometric center of all sample points belonging to the remaining phenotypes. When $\{N_k\}$ are significantly imbalanced for different k, the denominator term in (4) will be biased toward the dominant phenotype(s).

However, a problem associated with both PUG selection schemes specified by (3) and (4) (and with the OVRSNR criterion Golub et al., 1999) is that the criterion function considers the remaining classes as a single super class, which is suboptimal because it ignores a gene's ability to discriminate between classes *within* the super class.

We therefore propose OVE fold-change based PUG selection to fully support the objective of multicategory classification. Specifically, the OVEPUGs are defined as:

$$\mathbb{J}_{\mathrm{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\mathrm{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\mu_j(k)}{\max_{l \neq k}\{\mu_j(l)\}} \geqslant \tau_k \right\} \tag{5}$$

where the denominator term is the maximum phenotypic mean expression level over the remaining phenotype classes. This seemingly technical modification turns out to have important consequences since it assures that the selected PUGs are highly expressed in one phenotype relative to *each* of the remaining phenotypes, that is, “high” (up-regulated) in phenotype k and “low” (down-regulated) in *all* phenotypes $l \neq k$. In our experimental results, we will demonstrate that (5) leads to better classification accuracy than (4) on a well-known multi-class cancer domain.

Adopting the same strategy as in Shedden et al. (2003), to assure even-handed gene resources for discriminating both neighboring and well-separated classes, we select a fixed (common) number of top-ranked phenotype-specific subPUGs for each phenotype, that is, $\|\mathbb{J}_{\mathrm{PUG}}(k)\| = N_{\mathrm{subPUG}}$ for all k, and pool all these subPUGs together to form the final gene marker subset $\mathbb{J}_{\mathrm{PUG}}$ for the OVRSVM committee classifier. In our experiments, the optimum number of PUGs per phenotype, N_{subPUG}, is determined by surveying the curve of classification accuracy versus N_{subPUG} and selecting the number that achieves the best classification performance. More generally, in practice, N_{subPUG} can be chosen via a cross validation procedure. Figure 2 shows the geometric distribution of the selected PUGs specified by (5), where the PUGs (highlighted data points) constitute the lateral-edge points of the convex pyramid defined by the scatter plot of the phenotypic mean expressions (Zhang et al., 2008). Different from the PUG selection schemes given by (3) and (4), the PUGs selected based on (5) are most compact yet informative, since the down-regulated genes that are not differentially expressed between the remaining phenotypes (the genes on the lateral faces of the scatter plot convex pyramid) are excluded. From a statistical point of view, extensive studies on the normalized scatter plot of microarray gene expression data by many groups including our own indicate that the PUGs selected by (5) approximately follow an independent multivariate super-Gaussian distribution (Zhao et al., 2005) where subPUGs are mutually exclusive and phenotypic gene expression patterns defined over the PUGs are statistically independent (Wang et al., 2003).

It is worth noting that the PUG selection by (5) also adopts a univariate fold-change evaluation that does not require calculation of either expression variance or of correlation between genes (Shi et al., 2008). For the small sample size case typical of microarray data, multivariate gene selection schemes may introduce additional uncertainty in estimating the correlation structure (Lai et al., 2006; Shedden et al., 2003) and thus may fail to identify true gene markers (Wang et al., 2008). The exclusion of the variance in our criterion is also supported by the variance stabilization theory (Durbin et al., 2002; Huber et al., 2002), because the geometric mean in (5) is equivalent to the arithmetic mean after logarithmic transformation and the gene expression after logarithmic transfor-

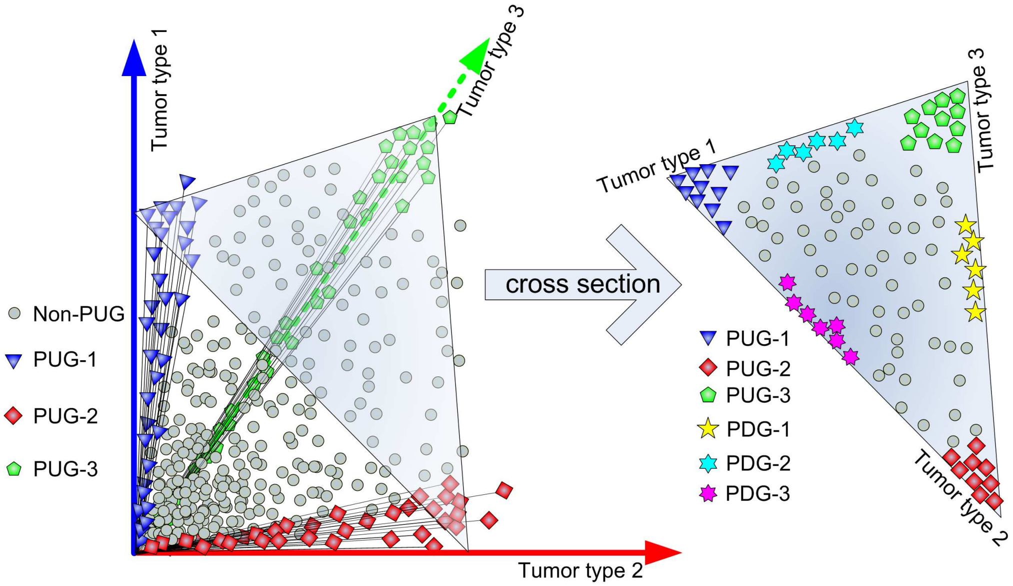

Figure 2: Geometric illustration of the selected one-versus-everyone phenotypic upregulated genes (OVEPUGs) associated with three phenotypic classes. Three-dimensional geometric distribution (on the untransformed scale) of the selected OVEPUGs, which reside around the lateral-edges of the phenotypic gene expression scatter plot convex pyramid, is shown in the left subfigure. A projected distribution of the selected OVEPUGs together with OVEPDGs is shown in the right cross-sectional plot, where OVEPDGs reside along the face-edges of the cross-sectional triangle.

mation approximately has the equal variance across different genes, especially for the up-regulated genes.

Corresponding to the definition of OVEPUGs, the OVEPDGs (which are down-regulated in one class while being up-regulated in all other classes) can be defined by the following criterion:

$$\mathbb{J}_{\text{PDG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PDG}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\min_{l \neq k}\{\mu_j(l)\}}{\mu_j(k)} \geqslant \tau_k \right\}. \qquad (6)$$

Furthermore, the combination of PUGs and PDGs can be defined as:

$$\mathbb{J}_{\text{PUG+PDG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG+PDG}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \max\left\{ \frac{\mu_j(k)}{\max_{l \neq k}\{\mu_j(l)\}},\ \frac{\min_{l \neq k}\{\mu_j(l)\}}{\mu_j(k)} \right\} \geqslant \tau_k \right\}. \qquad (7)$$

Purely from the machine learning view, PDGs have the theoretical capability of being as discriminating as PUGs. Thus, PDGs merit consideration as candidate genes. However, there are several critical differences, with consequential implications, between lowly-expressed genes and highly-expressed genes, such as the extraordinarily large proportion and relatively large noise of the lowly-expressed genes. We have evaluated the classification performance of PUGs, PDGs, and

PUGs+PDGs, respectively. Experimental results show that PDGs have less discriminatory power than PUGs and the inclusion of PDGs actually worsens classification accuracy, compared to just using PUGs. Experiments and further discussion will be given in the results section.

2.4 Review of Relevant Gene Selection Methods

Here we briefly review four benchmark gene selection methods that have been previously proposed for multicategory classification, namely, OVRSNR (Golub et al., 1999), OVR t-statistic (OVRt-stat) (Liu et al., 2002), BW (Dudoit et al., 2002), and SVMRFE (Guyon et al., 2002).

Let $\mu_{j,k}$ and $\mu_{j,\text{-}k}$ be the arithmetic means of the expression levels of gene j associated with phenotype k and associated with the super class of remaining phenotypes, respectively, on the log-transformed scale, with $\sigma_{j,k}$ and $\sigma_{j,\text{-}k}$ the corresponding standard deviations. OVRSNR gene selection for multicategory classification is given by:

$$\mathbb{J}_{\text{OVRSNR}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{OVRSNR}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\left|\mu_{j,k} - \mu_{j,\text{-}k}\right|}{\sigma_{j,k} + \sigma_{j,\text{-}k}} \geqslant \tau_k \right\}, \tag{8}$$

where τ_k is a pre-defined threshold (Golub et al., 1999). To assess the statistical significance of the difference between $\mu_{j,k}$ and $\mu_{j,\text{-}k}$, OVRt-stat applies a test of the null hypothesis that the means of two assumed normally distributed measurements are equal. Accordingly, OVRt-stat gene selection is given by Liu et al. (2002):

$$\mathbb{J}_{\text{OVR}t\text{-stat}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{OVR}t\text{-stat}}(k) = \bigcup_{k=1}^{M} \left\{ j \,\middle|\, \frac{\left|\mu_{j,k} - \mu_{j,\text{-}k}\right|}{\sqrt{\sigma_{j,k}^2 \big/ N_k + \sigma_{j,\text{-}k}^2 \big/ (N - N_k)}} \geqslant \tau_k \right\}, \tag{9}$$

where the p-values associated with each gene may be estimated. As aforementioned, one limitation of the gene selection schemes (8) and (9) is that the criterion function considers the remaining classes as a single group. Another is that they both require variance estimation.

Dudoit et al. (2002) proposed a pooled OVO gene selection method based on the BW sum of squares across all paired classes. Specifically, BW gene selection is specified by

$$\mathbb{J}_{\text{BW}} = \left\{ j \,\middle|\, \frac{\sum_{i=1}^{N} \sum_{k=1}^{M} \mathbf{1}_{\omega_k}(i) \left(\mu_{j,k} - \mu_j\right)^2}{\sum_{i=1}^{N} \sum_{k=1}^{M} \mathbf{1}_{\omega_k}(i) \left(x_{ij} - \mu_{j,k}\right)^2} \geqslant \tau \right\}, \tag{10}$$

where μ_j is the global arithmetic center of gene j over all sample points and $\mathbf{1}_{\omega_k}(i)$ is the indicator function reflecting membership of sample i in class k. As pointed out by Loog et al. (2001), BW gene selection may only preserve the distances of already well-separated classes rather than neighboring classes.

From a dimensionality reduction point of view, Guyon et al. (2002) proposed a feature subset ranking criterion for linear SVMs, dubbed the SVMRFE. Here, one first trains a linear SVM classifier on the full feature space. Features are then ranked based on the magnitude of their weights and are eliminated in the order of increasing weight magnitude. A widely adopted reduction strategy is to eliminate a fixed or decreasing percentage of features corresponding to the bottom portion of the ranked weights and then to retrain the SVM on the reduced feature space. Application to microarray gene expression data shows that the genes selected matter more than the classifiers with which they are paired (Guyon et al., 2002).

3. Results

We tested PUG-OVRSVM on five benchmarks and one in-house real microarray data set, and compared the performance to several widely-adopted gene selection and classification methods.

3.1 Description of the Real Data Sets

The numbers of samples, phenotypes, and genes, as well as the microarray platforms used to generate these gene expression data sets, are briefly summarized in Supplementary Tables 1$\sim$7. The six data sets are the MIT 14 Global Cancer Map data set (GCM) (Ramaswamy et al., 2001), the NCI 60 cancer cell lines data set (NCI60) (Staunton et al., 2001), the University of Michigan cancer data set (UMich) (Shedden et al., 2003), the Central Nervous System tumors data set (CNS) (Pomeroy et al., 2002), the Muscular Dystrophy data set (MD) (Bakay et al., 2006), and the Norway Ascites data set (NAS). To assure a meaningful and well-grounded comparison, we emphasized data quality and suitability in choosing these test data sets. For example, the data sets cannot be too "simple" (if the classes are well-separated, all methods perform equally well) or too "complex" (no method will then perform reasonably well), and each class should contain sufficient samples to support some form of cross-validation assessment.

We also performed several important pre-processing steps widely adopted by other researchers (Guyon et al., 2002; Ramaswamy et al., 2001; Shedden et al., 2003; Statnikov et al., 2005). When the expression levels in the raw data take negative values, probably due to global probe-set calls and/or data normalization procedures, these negative values are replaced by a fixed small quantity (Shedden et al., 2003). On the log-transformed scale, we further conducted a variance-based unsupervised gene filtering operation to remove the genes whose expression standard deviations (across all samples) were less than a pre-determined small threshold; this effectively reduces the number of genes by half (Guyon et al., 2002; Shedden et al., 2003).

3.2 Experiment Design

We decoupled the two key steps of multicategory classification: 1) selecting an informative subset of marker genes and then 2) finding an accurate decision function. For the crucial first step we implemented five gene selection methods, including OVEPUG specified by (5), OVRSNR specified by (8), OVRt-stat specified by (9), pooled BW specified by (10), and SVMRFE described in Ramaswamy et al. (2001). We applied these methods to the six data sets, and for each data set, we selected a sequence of gene subsets with varying sizes, indexed by N_{subPUG}, the number of genes per class. In our experiments, this number was increased from 2 up to 100. There are several reasons why we do not go beyond 100 subPUGs per class. First, classification accuracy may be either flat or monotonically decreasing as the number of features increases beyond a certain point, due to the theoretical bias-variance dilemma. Second, even in some cases where best performance is achieved using all the gene features, the idea of feature selection is to find the minimum number of features needed to achieve good (near-optimal) classification accuracy. Third, when $N_{\text{subPUG}} = 100$, the total number of genes used for classification is already quite large (this number is maximized if the sets $\mathbb{J}_{\text{PUG}}(k)$ are mutually exclusive, in which case it is N_{subPUG} times the number of classes). Fourth, but not least important, a large feature reduction may be necessary not only complexity-wise, but also for interpreting the biological functions and pathway involvement when the selected PUGs are most relevant and statistically reproducible.

The quality of the marker gene subsets was then assessed by prediction performance on four subsequently trained classifiers, including OVRSVM, kNN, NBC, and OVOSVM. In relation to the proposed PUG-OVRSVM approach, we evaluated all combinations of these four different gene selection methods and three different classifiers on all six benchmark microarray gene expression data sets.

To properly estimate the accuracy of predictive classification, a validation procedure must be carefully designed, recognizing limits on the accuracy of estimated performance, in particular for small sample size. Clearly, classification accuracy must be assessed on labelled samples ‘unseen’ during training. However, for multicategory classification based on small, class-imbalanced data sets, single batch held-out test data may be precluded, as there will be insufficient samples for both accurate classifier training and accurate validation (Hastie et al., 2001). A practical alternative is a sound cross-validation procedure, wherein all the data are used for both training and testing, but with held-out samples in a testing fold not used for any phase of classifier training, including gene selection and classifier design (Wang et al., 2008). In our experiments, we chose LOOCV, wherein a test fold consists of a single sample; the rest of the samples are placed in the training set. Using only the training set, the informative genes are selected and the weights of the linear OVRSVM are fit to the data (Liu et al., 2005; Shedden et al., 2003; Yeang et al., 2001). It is worth noting that LOOCV is approximately unbiased, lessening the likelihood of misestimating the prediction error due to small sample size; however, LOOCV estimates do have considerable variance (Braga-Neto and Dougherty, 2004; Hastie et al., 2001). We evaluated both the lowest “sustainable” prediction error rate and the lowest prediction error rate, where the sequence of sustainable prediction error rates were determined based on a moving-average of error rates along the survey axis of the number of genes used for each class, N_{subPUG}, with a moving window of width 5. We also report the number of genes per class at which the best sustainable performance was obtained.

While the error rate is estimated through LOOCV and the optimum number of PUGs used per class is obtained by the aforementioned surveying strategy, we should point out that a two-level LOOCV could be applied to jointly determine the optimum N_{subPUG} and estimate the associated error rate; however, such an approach is computationally expensive (Statnikov et al., 2005). For the settings of structural parameters in the classifiers, we used $C = 1.0$ in the SVMs for all experiments (Vapnik, 1998), and chose $k = 1, 2, 3$ in kNNs under different training sample sizes per class, as recommended by Duda et al. (2001).

3.3 Experimental Results

Our first comparative study focused on the GCM data widely used for evaluating multicategory classification algorithms (Cai et al., 2007; Ramaswamy et al., 2001; Shedden et al., 2003; Zhou and Tuck, 2007). The performance curves of OVRSVM committee classifiers trained on the commonly pre-processed GCM data using the five different gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) are detailed in Figure 3. It can be seen that our proposed OVEPUG selection significantly improved the overall multicategory classification when using different numbers of marker genes, as compared to the results produced by the four competing gene selection methods. For example, using as few as 9 genes per phenotypic class (with 126 distinct genes in total, that is, mutually exclusive PUGs for each class), we classified 164 of 190 (86.32%) of the tumors correctly. Furthermore, using LOOCV on the GCM data set of 190 primary malignant tumors, and using the optimal number of genes (61 genes per phenotypic class or 769 unique genes in total), we achieved the best (88.95% or 169 of 190 tumors) sustainable correct predictions. In contrast, at its optimum performance, OVRSNR gene selection achieved 85.37% sustainable correct predictions using 25 genes per phenotypic class, OVRt-stat gene selection achieved 84.53% sustainable correct predictions using 71 genes per phenotypic class, BW gene selection achieved 80.53% sustainable correct predictions using 94 genes per phenotypic class, and SVMRFE gene selection achieved 84.74% sustainable correct predictions using 96 genes per phenotypic class. In our comparative study, instead of solely comparing the lowest error rates achieved by different gene selection methods, we also emphasized the sustainable correct prediction rates, as potential overfitting to the data may produce an (unsustainably) good prediction performance. For our experiments in Figure 3, based on the realistic assumption that the probability of good predictions purely "by chance" over a sequence of consecutive gene numbers is low, we defined the sustainable prediction/error rates based on the moving-averaged prediction/error rates over $\delta = 5$ consecutive gene numbers. Here, δ gives the sustainability requirement.

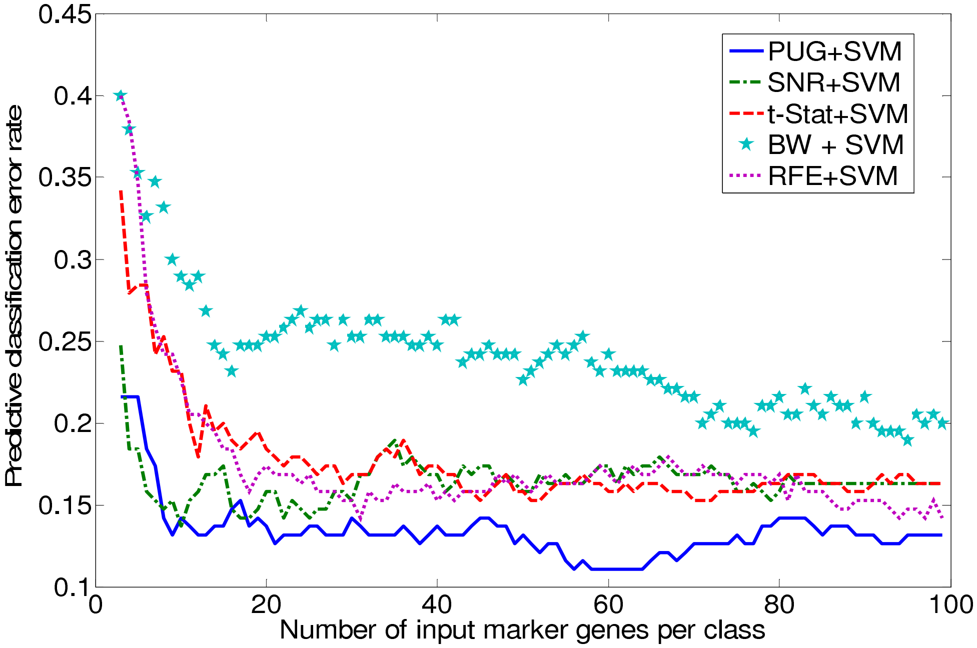

Figure 3: Comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) using the GCM benchmark data set. The curves of classification error rates were generated by using OVRSVM committee classifiers with varying size of the input gene subset.

For the purpose of information sharing with readers, based on publicly reported optimal results for different methods, we have summarized in Table 1 the comparative performance achieved by PUG-OVRSVM and eight existing/competing methods on the benchmark GCM data set, along with the gene selection methods used, the chosen classifiers, sample sizes, and the chosen cross-validation schemes. Obviously, since the reported prediction error rates were generated by different algorithms and under different conditions, any conclusions based on simple/direct comparisons of the reported results must be carefully drawn. We have chosen not to independently reproduce results by re-implementing the methods listed in Table 1, firstly because we typically do not have access to public domain code implementing other authors' methods and secondly because we feel that high reproducibility of previously published results may not be expected without knowing some likely undeclared optimization steps and/or additional control parameters used in the actual computer codes. Nevertheless, many reported prediction error rates on the GCM data set were actually based on the same/similar training sample set (144 $\sim$ 190 primary tumors) and the LOOCV scheme used in our PUG-OVRSVM experiments; furthermore, it was reported that the prediction error rates estimated by LOOCV and 144/54 split/held-out test were very similar (Ramaswamy et al., 2001). Specifically, the initial work on GCM by Ramaswamy et al. (2001) reported an achieved 77.78% prediction rate, and some improved performance was later reported by Yeang et al. (2001) and Liu et al. (2002), achieving 81.75% and 79.99% prediction rates, respectively. In the work most closely related to our gene selection scheme by Shedden et al. (2003), using a kNN tree classifier and using OVR fold-change based gene selection specified by (4), a prediction rate of 82.63% was achieved. In relation to these reported results on GCM, as indicated in Table 1, our proposed PUG-OVRSVM method produced the best sustainable prediction rate of 88.95%.

References	Gene-select	Classifier	Sample	CV scheme	Error rate
Ramaswamy et al. (2001)	OVRSVM RFE	OVRSVM	144&198	LOOCV 144/54	22.22%
Yeang et al. (2001)	N/A	OVRSVM	144	LOOCV	18.75%
Ooi and Tan (2003)	Genetic algorithm	MLHD	198	144/54	18.00%
Shedden et al. (2003)	OVR fold-change	kNN Tree	190	LOOCV	17.37%
Liu et al. (2005)	Genetic algorithm	OVOSVM	N/A	LOOCV	20.01%
Statnikov et al. (2005)	No gene selection	CS-SVM	308	10-fold	23.40%
Zhou and Tuck (2007)	CS-SVM RFE	OVRSVM	198	4-fold	16.72%
Cai et al. (2007)	DISC-GS	kNN	190	144/46	21.74%
PUG-OVRSVM	PUG	OVRSVM	190	LOOCV	**11.05%**

Table 1: Summary of comparative performances by OVEPUG-OVRSVM and eight competing methods (based on publicly reported optimum results) on the GCM benchmark data set.

A more stringent evaluation of the robustness of a classification method is to carry out the predictions on multiple data sets and then assess the overall performance (Statnikov et al., 2005). Our second comparative study evaluated the aforementioned five gene selection methods using the six benchmark microarray gene expression data sets. To determine whether the genes selected matter more than the classifiers used (Guyon et al., 2002), we used a common OVRSVM committee classifier and LOOCV scheme in all the experiments, and summarized the corresponding results in Table 2. For each experiment that used a distinct gene selection scheme applied to a distinct data set, we reported both sustainable (with sustainability requirement $\delta = 5$) and lowest (within parentheses) prediction error rates, as well as the number of genes per class that were used to produce these results. Clearly, the selected PUGs based on (5) produced the highest overall sustainable prediction rates as compared to the other four competing gene selection methods. Specifically, PUG is the consistent winner in 22 of 24 competing experiments (combinations of four gene selection schemes and six testing data sets). It should be noted that although BW and OVRSNR achieved comparably low prediction error rates on the CNS data set (with relatively balanced mixture distributions), they also produced high prediction error rates on the other testing data sets; the other competing gene selection methods also show some level of performance instability across data sets.

Gene-select	GCM	NCI60	UMich	CNS	MD	NAS
OVE PUG	**11.05%** (11.05%) [61 g/class]	**27.33%** (26.67%) [52 g/class]	**1.08%** (0.85%) [26 g/class]	**7.14%** (7.14%) [71 g/class]	**19.67%** (19.01%) [46 g/class]	**13.16%** (13.16%) [42 g/class]
OVR SNR	14.63% (13.68%) [25 g/class]	31.67% (31.67%) [58 g/class]	1.42% (1.42%) [62 g/class]	**7.14%** (7.14%) [57 g/class]	23.97% (23.97%) [85 g/class]	16.32% (15.79%) [54 g/class]
OVR t-stat	15.47% (15.26%) [71 g/class]	31.67% (31.67%) [56 g/class]	1.70% (1.70%) [45 g/class]	7.62% (7.14%) [92 g/class]	23.47% (22.31%) [56 g/class]	15.79% (15.79%) [74 g/class]
BW	19.47% (18.95%) [94 g/class]	31.67% (31.67%) [55 g/class]	1.30% (1.13%) [92 g/class]	**7.14%** (7.14%) [56 g/class]	19.83% (19.01%) [71 g/class]	21.05% (21.05%) [65 g/class]
SVM RFE	15.26% (14.21%) [96 g/class]	29.00% (28.33%) [81 g/class]	1.13% (1.13%) [58 g/class]	14.29% (14.29%) [53 g/class]	29.09% (28.10%) [73 g/class]	32.11% (31.58%) [94 g/class]

Table 2: Performance comparison between five different gene selection methods tested on six benchmark microarray gene expression data sets, where the predictive classification error rates for all methods were generated based on OVRSVM committee classification and an LOOCV scheme. Both sustainable and lowest (within parentheses) error rates are reported together with number of genes used per class.

To give more complete comparisons that also involved different classifiers (Statnikov et al., 2005), we further illustrate the superior prediction performance of the matched OVEPUG selection and OVRSVM classifier as compared to the best results produced by combinations of three different classifiers (OVOSVM, kNN, NBC) and four gene selection methods (PUG, OVRSNR, OVRt-stat, pooled BW). The optimum experimental results achieved over all combinations of these methods on the six data sets are summarized in Table 3, where we report both sustainable prediction error rates and the corresponding gene selection methods. Again, PUG-OVRSVM outperformed all other methods on all six data sets and was a clear winner in all 15 competing experiments. Our comparative studies also reveal that although gene selection is a critical step of multi-category classification, the classifiers used do indeed play an important role in achieving good prediction performance.

3.4 Comparison Results on the Realistic Simulation Data Sets

To more reliably validate and compare the performance of the different gene selection methods, we have conducted additional experiments involving realistic simulations. The advantage of using synthetic data is that, unlike the real data sets often with small sample size and with LOOCV as the only applicable validation method, large testing samples can be generated to allow an accurate and reliable assessment of a classifier's generalization performance. Two different simulation approaches were implemented. In both, we modeled the joint distribution for microarray data under each class and generated *i.i.d.* synthetic data sets consistent both with these distributions and with assumed class priors. In the first approach, we chose the class-conditional models consistent with commonly

	GCM	NCI60	UMich	CNS	MD	NAS
OVR SVM	**11.05%** (OVEPUG)	**27.33%** (OVEPUG)	**1.08%** (OVEPUG)	**7.14%** (OVEPUG)	**19.67%** (OVEPUG)	**13.16%** (OVEPUG)
OVO SVM	14.74% (OVEPUG)	33.33% (OVRSNR)	1.70% (OVEPUG)	9.52% (BW)	19.83% (BW)	16.32% (OVRSNR)
kNN	21.05% (OVEPUG)	31.67% (OVRt-stat)	2.27% (OVEPUG)	13.33% (OVEPUG)	21.81% (BW)	13.68% (OVRt-stat)
NBC	36.00% (OVRSNR)	51.67% (OVRSNR)	2.83% (OVRt-stat)	37.62% (BW)	37.69% (BW)	34.21% (OVEPUG)

Table 3: Performance comparison based on the lowest predictive classification error rates produced by OVEPUG-OVRSVM and the optimum combinations of five different gene selection methods and three different classifiers, tested on six benchmark microarray gene expression data sets and assessed via the LOOCV scheme.

accepted properties of microarray data (few discriminating features, many non-discriminating features, and with small sample size) (Hanczar and Dougherty, 2010; Wang et al., 2002). In the second approach, we directly estimated the class-conditional models based on a real microarray data set and then generated the *i.i.d.* samples according to the learned models.

3.4.1 Design I

We simulated 5000 genes, with 90 “relevant” and 4910 “irrelevant” genes. Inspired by gene clustering concept in modelling local correlations, we divided the genes into 1000 blocks of size five, each containing exclusively either relevant or irrelevant genes. Within each block the correlation coefficient is 0.9, with zero correlation across blocks. Irrelevant genes are assumed to follow a (univariate) standard normal distribution, for all classes. Relevant genes also follow a normal distribution with variance 1 for all classes. There are three equally likely classes, A, B and C. The mean vectors of the 90 relevant genes under each class are shown in Table 4. The means were chosen to make the classification task neither too easy nor too difficult and to simulate unequal distances between the classes—A and B are relatively close, with C more distant from both A and B.

	The mean vector μ for each class
μ_A	[2.8 2.8 2.8 2.8 2.8 1 1 1 1 1 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 1 1 1 1 1 3 3 3 3 3 0.1 0.1 0.1 0.1 0.1]
μ_B	[1 1 1 1 1 2.8 2.8 2.8 2.8 2.8 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 1 1 1 1 1 3 3 3 3 3 0.1 0.1 0.1 0.1 0.1]
μ_C	[1 1 1 1 1 1 1 1 1 1 14.4 14.4 14.4 14.4 14.4 8.5 8.5 8.5 8.5 8.5 8 8 8 8 8 10 10 10 10 10 10 10 10 10 10 10 9 9 9 9 9 10 10 10 10 10 3 3 3 3 3 10 10 10 10 10 10 10 10 10 10 10 10 10 10 8.5 8.5 8.5 8.5 8.5 8 8 8 8 8 9 9 9 9 9 11 11 11 11 11 7.1 7.1 7.1 7.1 7.1]

Table 4: The mean vectors of the 90 relevant genes under each of the three classes.

We randomly generated 100 synthetic data sets, each partitioned into a small training set of 60 samples (20 per class) and a large testing set of 6000 samples.

3.4.2 DESIGN II

The second approach models each class as a more realistic multivariate normal distribution $N(\mu,\Sigma)$, with the class's mean vector μ and covariance matrix Σ directly learned from the real microarray data set GCM. Estimation of a covariance matrix is certainly a challenging task, specifically due to the very high dimensionality of the gene space ($p = 15,927$ genes in the GCM data) and only a few dozen samples available for estimating $p(p-1)/2$ free covariate parameters per class. It is also computationally prohibitive to generate random vectors based on full covariances on a general desktop computer. To address both of these problems, we applied a factor model (McLachlan and Krishnan, 2008), which can significantly reduces the number of free parameters to be estimated while capturing the main correlation structure in the data.

In factor analysis, the observed $p \times 1$ vector $\mathbf{t}$ is modeled as

$$\mathbf{t} = \mu + \mathbf{W}\mathbf{x} + \varepsilon,$$

where μ is the mean vector of observation $\mathbf{t}$, $\mathbf{W}$ is a $p \times q$ matrix of factor loadings, $\mathbf{x}$ is the $q \times 1$ latent variable vector with standard normal distribution $N(\mathbf{0},\mathbf{I})$ and ε is noise with independent multivariate normal distribution $N(\mathbf{0},\Psi)$, $\Psi = \mathbf{diag}\left(\sigma_1^2,\ldots,\sigma_p^2\right)$. The resulting covariance matrix Σ is

$$\Sigma = \mathbf{W}\mathbf{W}^T + \Psi.$$

Estimation of Σ reduces to estimating $\mathbf{W}$ and Ψ, totaling $p(q+1)$ parameters. Usually, we have q much less than p. The factor model is learned via the EM algorithm (McLachlan and Krishnan, 2008), initialized by probabilistic principal component analysis (Tipping and Bishop, 1999).

In our experiments, we set $q = 5$, which typically accounted for 60% of the energy. We also tried $q = 3$ and 7 and observed that the relative performance remained unchanged, although the absolute performance of all methods does change with q.

Five phenotypic classes were used in our simulation: breast cancer, lymphoma, bladder cancer, leukemia and CNS. 100 synthetic data sets were generated randomly according to the learned class models from the real data of these five cancer types. The dimension for each sample is 15,927. For each data set, the training sample size was the same as used in the real data experiments, with 11, 22, 11, 30, and 20 samples in the five respective classes; and the testing set consisted of 3,000 samples, 600 per class.

3.5 Evaluation of Performance

For a given gene-selection method and for each data set (indexed by $i = 1,\ldots,100$), the classifier F_i is learned. We then evaluate F_i on the i-th testing set, and measure the error rate ε_i. Since the testing set has quite large sample size, we would expect ε_i to be close to the true classification error rate for F_i. Over 100 simulation data sets, we then calculated both the average classification error $\overline{\varepsilon}$ and the standard deviation σ.

Furthermore, let $\varepsilon_{i,PUG}$ denote the error rate associated with PUGs on testing set i, and similarly, let $\varepsilon_{i,SNR}$, $\varepsilon_{i,t\text{-}stat}$, $\varepsilon_{i,BW}$ and $\varepsilon_{i,SVMRFE}$ denote the error rates associated with the four peer gene selection methods. The error rate difference between two methods, for example, PUG and SNR, is defined by

$$D_i(PUG,SNR) = \varepsilon_{i,PUG} - \varepsilon_{i,SNR}.$$

For each synthetic data set, we define the “winner” as the one with the least testing error rate. For each method, the mean and standard deviation of the error rate and the frequency of winning are examined for performance evaluation. In addition, the histogram of error rate differences between PUG and peer methods are provided.

3.6 Experimental Results on the Simulation Data Sets

We tested all gene selection methods using the common OVRSVM classifier. All the experiments were done using the same procedure as on the real data sets, except with LOOCV error estimation replaced by the error estimation using large size independent testing data. Figure 4, analogous to Figure 3 while on the realistic synthetic data whose model was estimated from GCM data set (simulation data under design II), shows the comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE). Tables 5 and 6 show the average error, standard deviation, and frequency of winning, estimated based on the 100 simulation data sets. PUG has the smallest average error over all competing methods. PUG also is the most stable method (with the smallest standard deviation). Tables 7 and 8 provide the comparison results of the five competing methods on the first ten data sets.

Figures 5 and 6 show histograms of the error difference between PUG and other methods, where a negative value of the difference indicates better performance by PUG. The red bar shows the position where the two methods are equal. We can see that the vast majority of differences are negative. Actually, as indicated in Tables 5 and 6, there is no positive difference in the subfigures of Figure 5 and at most one positive difference in the subfigures of Figure 6.

	PUG	SNR	t-stat	BW	SVMRFE
mean	**0.0724**	0.1129	0.1135	0.1165	0.1203
std deviation	**0.0052**	0.0180	0.0188	0.0177	0.0224
frequency of ‘winner’	**100**	0	0	0	0

Table 5: The mean and standard deviation of classification error and the frequency of winner based on 100 simulation data sets with design I.

	PUG	SNR	t-stat	BW	SVMRFE
mean	**0.0712**	0.1311	0.1316	0.2649	0.0910
std deviation	**0.0201**	0.0447	0.0449	0.0302	0.0244
frequency of ‘winner’	**99**	0	0	0	1

Table 6: The mean and standard deviation of classification error and the frequency of winner based on 100 simulation data sets with design II.

3.7 Comparison Between PUGs and PDGs

In this experiment, we selected PDGs according to the definition given in (6) and evaluated gene selection based on PUGs, PDGs, and based on their union, as given in (7). Again, all gene selection methods were coupled with the OVRSVM classifier. Table 9 shows classification performance for

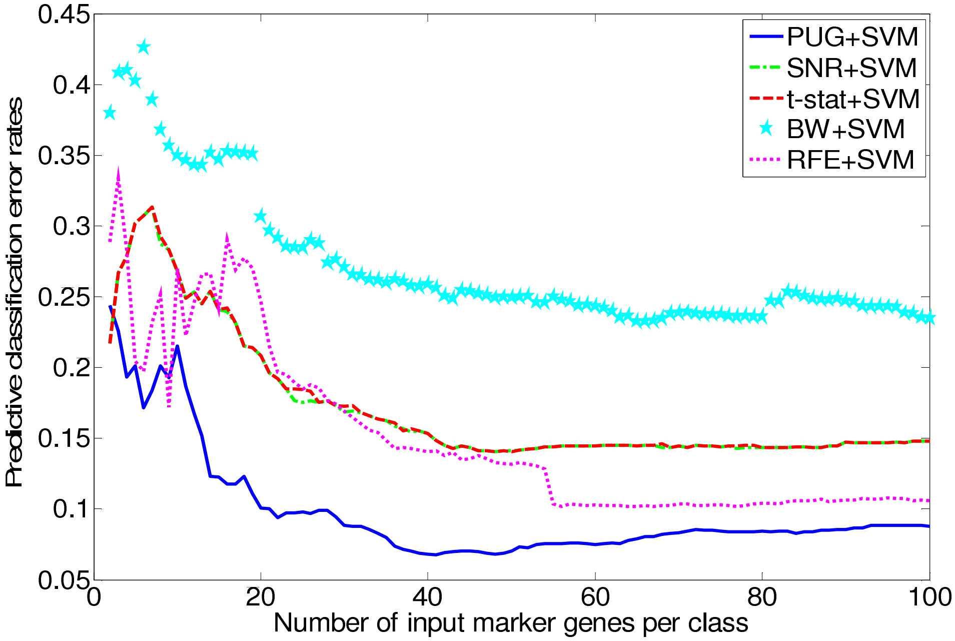

Figure 4: Comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) on one simulation data set under design II. The curves of classification error rates were generated by using OVRSVM committee classifiers with varying size of the input gene subset.

	sim_1	sim_2	sim_3	sim_4	sim_5	sim_6	sim_7	sim_8	sim_9	sim_10
PUG	**0.0864**	**0.0773**	**0.0697**	**0.0681**	**0.0740**	**0.0761**	**0.0740**	**0.0721**	**0.0666**	**0.0758**
SNR	0.1078	0.1092	0.1028	0.1279	0.1331	0.1004	0.1011	0.1253	0.0817	0.0838
t-stat	0.1109	0.1089	0.1022	0.1251	0.1333	0.0991	0.1016	0.1268	0.0823	0.0832
BW	0.1127	0.0995	0.1049	0.1271	0.1309	0.1107	0.1044	0.1291	0.0903	0.0845
SVMRFE	0.1030	0.1009	0.0967	0.1219	0.1248	0.1016	0.1107	0.1191	0.1198	0.0933

Table 7: Comparison of the classification error for the first ten simulation data sets with design I.

PUGs, PDGs and PUGs+PDGs. Clearly, PDGs have less discriminatory power than PUGs, and the inclusion of PDGs (generally) worsens classification accuracy, compared with just using PUGs.

	sim_1	sim_2	sim_3	sim_4	sim_5	sim_6	sim_7	sim_8	sim_9	sim_10
PUG	**0.0694**	**0.0610**	**0.0748**	**0.0675**	0.0536	**0.0474**	**0.0726**	**0.0818**	**0.0560**	**0.0700**
SNR	0.1559	0.0659	0.1142	0.1211	0.0508	0.1937	0.1568	0.1464	0.0797	0.0711
t-stat	0.1559	0.0659	0.1142	0.1210	0.0508	0.1939	0.1568	0.1464	0.0797	0.0712
BW	0.2373	0.2698	0.2510	0.2650	0.3123	0.2464	0.3070	0.2236	0.2800	0.3055
SVMRFE	0.0906	0.0739	0.0864	0.0852	**0.0426**	0.0776	0.0863	0.0973	0.0655	0.0730

Table 8: Comparison of the classification error for the first ten simulation data sets with design II.

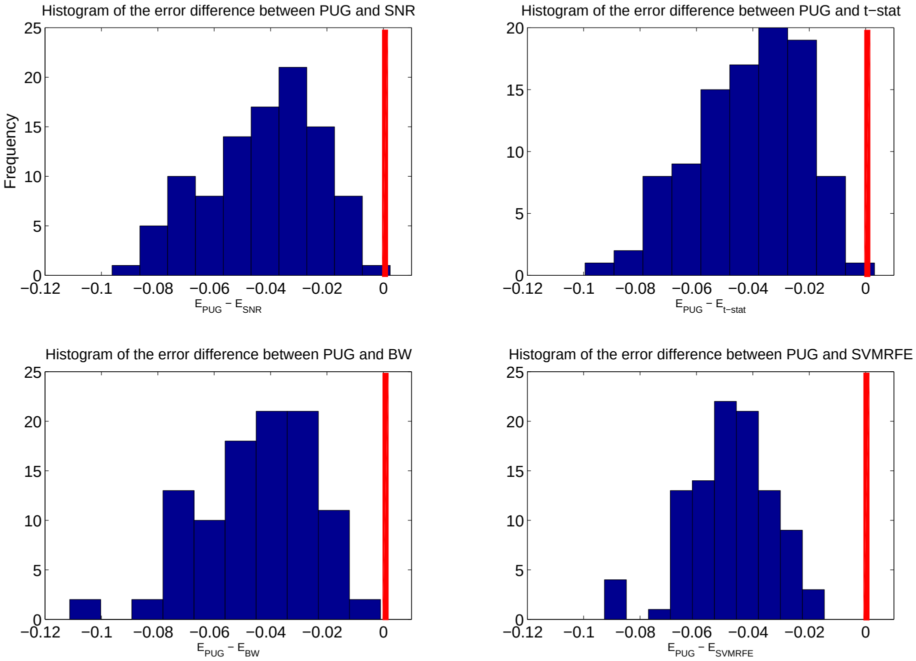

Figure 5: Histogram of the error difference between PUG and other methods with design I.

Error Rate	GCM	NCI60	UMich	CNS	MD	NAS
PUG	**11.05%**	**27.33%**	**1.08%**	**7.14%**	**19.67%**	**13.16%**
PDG	17.58%	30.33%	1.98%	9.52%	26.28%	25.79%
PUG+PDG	14.53%	30.67%	1.13%	**7.14%**	23.14%	15.79%

Table 9: Classification comparison of PUG and PDG on the six benchmark data sets.

There are several potential reasons that may jointly explain the non-contributing or even negative role of the included PDGs. First, the number of PDGs are much less than that of PUGs, that is, PUGs represent the significant majority of informative genes when PUGs and PDGs are jointly considered, as shown in Table 10 (Top PUG+PDGs were selected with 10 genes per class and we counted how many PUGs are included in the total). Second, PDGs are less reliable than PUGs due to the noise characteristics of gene expression data, that is, low gene expressions contain relatively large additive noise after log-transformation (Huber et al., 2002; Rocke and Durbin, 2001). This is further exacerbated by the follow-up one-versus-rest classifier because there are many more samples in the 'rest' group than in the 'one' group. This practically increases the relative noise/variability associated with PDGs in the 'one' group. In addition, PUGs are consistent with the practice of molecular pathology and thus may have broader clinical utility, for example, most currently available disease gene markers are highly expressed (Shedden et al., 2003).

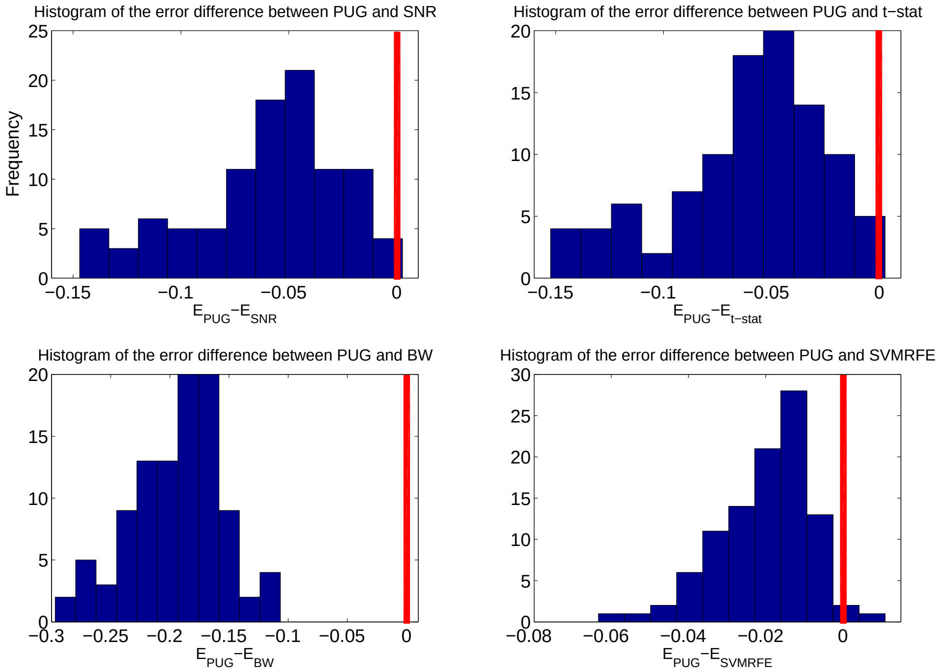

Figure 6: Histogram of the error difference between PUG and other methods with design II.

	GCM	NCI60	UMich	CNS	MD	NAS
No. of PUG	113	76	56	33	76	65
No. of PUG+PDG	140	90	60	50	130	70
% of PUG	80.71%	84.44%	93.33%	66.00%	58.46%	92.86%

Table 10: Classification comparison of PUG and PDG on the six benchmark data sets.

3.8 Marker Gene Validation by Biological Knowledge

We have applied existing biological knowledge to validate biological plausibility of the selected PUG markers for two data sets, GCM and NAS. The full list of genes most highly associated with each of the 14 tumor types in the GCM data set are detailed in the Supplementary Tables 8 and 9.

3.8.1 BIOLOGICAL INTERPRETATION FOR GCM DATA SET

Prolactin-induced protein, which is regulated by prolactin activation of its receptors, ranks highest among the PUGs associated with breast cancer. Postmenopausal breast cancer risk is strongly associated with elevated prolactin levels (PubMed IDs 15375001, 12373602, 10203283). Interestingly, prolactin release proportionally increases with increasing central fat in obese women (PubMed ID 15356045) and women with this pattern of obesity have an increased risk of breast cancer mortality (PubMed ID 14607804). Other genes of interest that rank among the top 10 breast cancer PUGs include CRABP2, which transports retinoic acid to the nucleus. Retinoids are important regulators of breast cell function and show activity as potential breast cancer chemopreventive agents (PubMed IDs 11250995, 12186376). Mammglobin is primarily expressed in normal breast epithelium and breast cancers (PubMed ID 12793902). Carbonic anhydrase XII is expressed in breast cancers and is generally considered a marker of a good prognosis (PubMed ID 12671706). The selective expression and/or function of these genes in breast cancers are consistent with their selection as PUGs in the classification scheme.

The top 10 PUGs associated with prostate cancer include several genes strongly associated with the prostate including prostate specific antigen (PSA) and its alternatively spliced form 2, and prostatic secretory protein 57. The role of PSA gene KLK3 and KLK1 as a biomarker of prostate cancer is well established (PubMed ID 19213567). Increased NPY expression is associated with high-grade prostatic intraepithelial neoplasia and poor prognosis in prostate cancers (PubMed ID 10561252). ACPP is another prostate specific protein biomarker (PubMed ID 8244395). The strong representation of genes that show clear selectivity for expression within the prostate illustrates the potential of the PUGs as bio-markers linked to the biology of the underlying tissues.

Several of the selected PUG markers for uterine cancer fit very well with our current biological understanding of this disease. It is well-established that estrogen receptor alpha (ESR1) is expressed or amplified in human uterine cancer (PubMed IDs 18720455, 17911007, 15251938), while the Hox7 gene (MSX1) contributes to uterine function in cow and mouse models, especially at the onset of pregnancy (PubMed IDs 7908629, 14976223, 19007558). Mammaglobin 2 (SCGB2A1) is highly expressed in a specific type of well-differentiated uterine cancer (endometrial cancers) (PubMed ID 18021217), and PAM expression in the rat uterus is known to be regulated by estrogen (PubMed IDs 9618561, 9441675). Other PUGs provide novel insights into uterine cancer that are deserving of further study. Our PUG selection ranks HE4 higher than the well-established CA125 marker, which may suggest HE4 as a promising alternative for the clinical management of endometrial cancer. One recent study (PubMed ID 18495222) shows that, at 95% specificity, the sensitivity of differentiating between controls and all stages of uterine cancer is 44.9% using HE4 versus 25.2% using CA125 ($p = 0.0001$).

Osteopontin (OPN) is an integrin-binding protein that is involved in tumorigenesis and metastasis. OPN levels in the plasma of patients with ovarian cancer are much higher compared with plasma from healthy individuals (PubMed ID 11926891). OPN can increase the survival of ovarian cancer cells under stress conditions in vitro and can promote the late progression of ovarian cancer in vivo, and the survival-promoting functions of OPN are mediated through Akt activation (PubMed ID 19016748). Matrix metalloproteinase 2 (MMP2) is an enzyme degrading collagen type IV and other components of the basement membrane. MMP-2 is expressed by metastatic ovarian cancer cells and functionally regulates their attachment to peritoneal surfaces (PubMed ID 18340378). MMP2 facilitates the transmigration of human ovarian carcinoma cells across endothelial extracellular matrix (PubMed ID 15609323). Glutathione peroxidase 3 (GPX3) is one of several isoforms of peroxidases that reduce hydroperoxides to the corresponding alcohols by means of glutathione (GSH) (PubMed ID 17081103). GPX3 has been shown to be highly expressed in ovarian clear cell adenocarcinoma. Moreover, GPX3 has been associated with low cisplatin sensitivity (PubMed ID 19020706).

3.8.2 Biological Interpretation for NAS Data Set

Several top-ranking gene products identified by our computational method have been well established as tumor-type specific markers and many of them have been used in clinical diagnosis. For example, mucin 16, also known as CA125, is a FDA-approved serum marker to monitor disease progression and recurrence in ovarian cancer patients (PubMed ID 19042984). Likewise, kallikrein family members including KLK6 and KLK8 are known to be ovarian cancer associated markers which can be detected in body fluids in ovarian cancer patients (PubMed ID 17303231). TITF1 (also known as TTF1) has been reported as a relatively specific marker in lung adenocarcinoma (PubMed ID 17982442) and it has been used to assist differential diagnosis of lung cancer from other types of carcinoma. Fatty acid synthase (FASN) is a well-known gene that is often upregulated in breast cancer (PubMed ID 17631500) and the enzyme is amenable for drug targeting using FASN inhibitors, suggesting that it can be used as a therapeutic target in breast cancer. The above findings indicate the robustness of our computational method in identifying tumor-type specific markers and in classifying different types of neoplastic diseases. Such information could be useful in translational studies (PubMed ID 12874019). Metastatic carcinoma of unknown origin is a relatively common presentation in cancer patients and an accurate diagnosis of the tumor type in the metastatic diseases is important to direct appropriate treatment and predict clinical outcome. The distinctive patterns of gene expression characteristic to various types of cancer may help pathologists and clinicians to better manage their patients.

3.9 Gene Comparisons Between Methods

It may be informative to provide some initial analysis on how the selected genes compare between methods; however, without definitive ground truth on cancer markers, the utility of this information is somewhat limited and should, thus, be treated as anecdotal, rather than conclusive. Specifically, we have now done some assessment of how differentially these gene selection methods rank some known cancer marker genes. The overlap rate is defined as the number of genes commonly selected by two methods over the maximum size of the two selected gene sets. Let G_1 and G_2 denote the gene sets selected by gene selection methods 1 and 2, respectively, and $|G|$ denote the cardinality (the size) of set G. The overlap rate between G_1 and G_2 is

$$R = \frac{|G_1 \cap G_2|}{\max(|G_1|, |G_2|)}.$$

Table 11 shows the overlap rate between methods on the top 100 genes per class. We can see that the overlap rates between methods are generally low except for the pair of SNR and t-stat. BW genes are quite different from the genes selected by all other methods and have only about 15% overlap rate with PUG and SVMRFE. The relatively high overlap rate between SNR and t-stat may be expected since they use quite similar summary statistics in their selection criteria.

We have also examined a total of 16 genes with known associations with 4 tumor types. These 16 genes are well-known markers supported by current biological knowledge. The rank of biomedical importance of these genes produced by each method is summarized in Table 12. When a gene is not listed in the top 100 genes by a wrapper method like SVMRFE, we simply assign the rank as '$>$100'. Generally but not uniformly across cancer types, these validated marker genes are highly ranked in the PUGs list as compared to other methods, and thus will be surely selected by PUG criterion.

4. Discussion

In this paper, we address several critical yet subtle issues in multicategory molecular classification applied to real-world biological and/or clinical applications. We propose a novel gene selection methodology matched to the multicategory classification approach (potentially with an unbalanced

Overlapping Rate	PUG	SNR	t-stat	BW	SVMRFE
PUG	1	0.4117	0.3053	0.1450	0.4057
SNR	0.4117	1	0.7439	0.3307	0.3841
t-stat	0.3053	0.7439	1	0.2907	0.3941
BW	0.1450	0.3307	0.2907	1	0.1307
SVMRFE	0.4057	0.3841	0.3941	0.1307	1

Table 11: The overlapping rate between methods on the top 100 genes per class.

Breast Cancer Relevant Genes						Prostate Cancer Relevant Genes					
Gene Symbol	Rank					Gene Symbol	Rank				
	PUG	SNR	t-stat	BW	SVMRFE		PUG	SNR	t-stat	BW	SVMRFE
PIP	1	5745	6146	473	>100	KLK3	4	5	11	61	15
CRABP2	4	5965	6244	498	>100	KLK1	5	3	9	76	16
SCGB2A2	6	6693	6773	458	14	NPY	7	18	22	344	30
CA12	9	6586	6647	518	>100	ACPP	3	4	8	71	12
Uterine Cancer Relevant Genes						Ovarian Cancer Relevant Genes					
Gene Symbol	Rank					Gene Symbol	Rank				
	PUG	SNR	t-stat	BW	SVMRFE		PUG	SNR	t-stat	BW	SVMRFE
ESR1	1	2	16	130	5	OPN	15	334	517	371	63
Hox7	2	4	52	307	12	MMP2	42	2626	3045	481	>100
SCGB2A1	8	3	19	190	4	GPX3	7	411	812	446	>100
PAM	10	83	281	365	71						
HE4	3	1	3	99	5						

Table 12: Detailed comparison between methods on several validated marker genes.

mixture distribution) that is not a straightforward pooled extension of binary (two-class) differential analysis. We emphasize the statistical reproducibility and biological plausibility of the selected gene markers under small sample size, supported by their detailed biological interpretations. We tested our method on six benchmark and in-house real microarray gene expression data sets and compared its performance with that of several existing methods. We imposed a rigorous performance assessment where each and all components of the scheme including gene selection are subjected to cross-validation, for example, held-out/unseen samples in a testing fold are not used for any phase of classifier training.

Tested on six benchmark real microarray data sets, the proposed PUG-OVRSVM method outperforms several widely-adopted gene selection and classification methods with lower error rates, fewer marker genes, and higher performance sustainability. Moreover, while for some data sets, the absolute gain in classification accuracy percentage of PUG-OVRSVM is not dramatically large, it must be recognized that the performance may be approaching the minimum Bayes error rate, in which case PUG-OVRSVM is achieving nearly all the improvement that is theoretically attainable. Furthermore, the improved performance is achieved by correct classifications on some of the most difficult cases, which is considered significant for clinical diagnosis (Ramaswamy et al., 2001). Lastly, although improvements will be data set-dependent, our multi-data set tests have shown that PUG-OVRSVM is the consistent winner as compared to several peer methods.

We note that we have opted for simplicity as opposed to theoretical optimality in designing our method. Our primary goal was to demonstrate that a small number of reproducible phenotypic-dependent genes are sufficient to achieve improved multicategory classification, that is, small sample sizes and a large number of classes need not preclude a high level of performance. Our studies suggest that using genes' marginal associations with the phenotypic categories as we do here has the potential to stabilize the learning process, leading to a substantial reduction in performance variability with small sample size; whereas, the current generation of complex gene selection techniques may not be stable or powerful enough to reliably exploit gene interactions and/or variations unless the sample size is sufficiently large. We have not explored the full flexibility that this method readily allows, with different numbers of subPUGs used by different classifiers. Presumably, equal or better performance could be achieved with fewer genes if more markers were selected for the most difficult classifications, involving the nearest phenotypes. However, such flexibility could actually degrade performance in practice since it introduces extra design choices and, thus, extra sources of variation in classification performance. We may also extend our method to account for variation in fold-changes, with the uncertainty estimated on bootstrap samples judiciously applied to eliminate those PUGs with high variations.

Notably, multicategory classification is intrinsically a nonlinear classification problem, and this method (using one-versus-everyone fold-change based PUG selection, linear kernel SVMs, and the MAP decision rule) is most practically suitable to discriminating unimodal classes. Future work will be required to extend PUG-OVRSVM for multimodal class distributions. An elegant yet simple strategy is to introduce unimodal pseudo-classes for the multi-modal classes via a pre-clustering step, with the final class decision readily made without the need of any decision combiner. Specifically, for each (pseudo-class, super pseudo-class) pair (where, for a pseudo-class originating from class k, the paired super pseudo-class is the union of all pseudo-classes that do not belong to class k), a separating hyperplane is constructed. Accordingly, in selecting subPUGs for each pseudo-class, the pseudo-classes originating from the same class will not be considered.

Acknowledgments

This work was supported in part by the US National Institutes of Health under Grants CA109872, CA149147, CA139246, EB000830, NS29525, and under Contract No. HHSN261200800001E.

References

M. Bakay, Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, V. Sartorelli, J. Seo, E. Pegoraro, C. Angelini, B. Shneiderman, D. Escolar, Y. W. Chen, S. T. Winokur, L. M. Pachman, C. Fan, R. Mandler, Y. Nevo, E. Gordon, Y. Zhu, Y. Dong, Y. Wang, and E. P. Hoffman. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. *Brain*, 129(Pt 4):996–1013, 2006.

U. M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–80, 2004.

Z. Cai, R. Goebel, M. R. Salavatipour, and G. Lin. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. *BMC Bioinformatics*, 8:206, 2007.

R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*, 8(1):37–49, 2008.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, New York, 2nd edition, 2001.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457): 77–87, 2002.

B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:S105–10, 2002.

G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104–11, 2005.

H. Gish. A probabilistic approach to the understanding and training of neural network classifiers. In *IEEE Intl. Conf. Acoust., Speech, Signal Process.*, pages 1361–1364, 1990.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439):531–7, 1999.

I. Guyon, J. Weston, S. Barnhill, and V. Vladimir. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

B. Hanczar and E. R. Dougherty. On the comparison of classifiers for microarray data. *Current Bioinformatics*, 5(1):29–39, 2010.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, New York, 2001.

W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

C. Lai, M. J. Reinders, L. J. van't Veer, and L. F. Wessels. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, 7: 235, 2006.

F. Li and Y. Yang. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21(19):3741–7, 2005.

T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–37, 2004.

H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60, 2002.

J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling. Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21(11):2691–7, 2005.

M. Loog, R. P. W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Trans Pattern Anal Machine Intell*, 23(7):762–766, 2001.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, Hoboken, N.J., 2nd edition, 2008.

C. H. Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44, 2003.

S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–42, 2002.

S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26):15149–54, 2001.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2002.

D. M. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *J Comput Biol*, 8(6):557–69, 2001.

K. A. Shedden, J. M. Taylor, T. J. Giordano, R. Kuick, D. E. Misek, G. Rennert, D. R. Schwartz, S. B. Gruber, C. Logsdon, D. Simeone, S. L. Kardia, J. K. Greenson, K. R. Cho, D. G. Beer, E. R. Fearon, and S. Hanash. Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *Am J Pathol*, 163 (5):1985–95, 2003.

L. Shi, W. D. Jones, R. V. Jensen, S. C. Harris, R. G. Perkins, F. M. Goodsaid, L. Guo, L. J. Croner, C. Boysen, H. Fang, F. Qian, S. Amur, W. Bao, C. C. Barbacioru, V. Bertholet, X. M. Cao, T. M. Chu, P. J. Collins, X. H. Fan, F. W. Frueh, J. C. Fuscoe, X. Guo, J. Han, D. Herman, H. Hong, E. S. Kawasaki, Q. Z. Li, Y. Luo, Y. Ma, N. Mei, R. L. Peterson, R. K. Puri, R. Shippy, Z. Su, Y. A. Sun, H. Sun, B. Thorn, Y. Turpaz, C. Wang, S. J. Wang, J. A. Warrington, J. C. Willey, J. Wu, Q. Xie, L. Zhang, L. Zhang, S. Zhong, R. D. Wolfinger, and W. Tong. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics*, 9 Suppl 9:S10, 2008.

A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–43, 2005.

J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, J. P. Mesirov, E. S. Lander, and T. R. Golub. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, 98(19):10787–92, 2001.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, 2002.

M. E. Tipping and C.M. Bishop. Probabilistic principle component analysis. *Journal of the Royal Statistical Society. Series B*, 61(3):611–622, 1999.

V. N. Vapnik. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998.

Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke. Iterative normalization of cNDA microarray data. *IEEE Trans Info. Tech. Biomed*, 6(1):29–37, 2002.

Y. Wang, J. Zhang, J. Khan, R. Clarke, and Z. Gu. Partially-independent component analysis for tissue heterogeneity correction in microarray gene expression analysis. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 24–32, 2003.

Y. Wang, D. J. Miller, and R. Clarke. Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br J Cancer*, 98(6):1023–8, 2008.

Z. Wang, Y. Wang, J. Xuan, Y. Dong, M. Bakay, Y. Feng, R. Clarke, and E. P. Hoffman. Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data. *Bioinformatics*, 22(6):755–61, 2006.

J. Xuan, Y. Wang, Y. Dong, Y. Feng, B. Wang, J. Khan, M. Bakay, Z. Wang, L. Pachman, S. Winokur, Y. W. Chen, R. Clarke, and E. Hoffman. Gene selection for multiclass prediction by weighted fisher criterion. *EURASIP J Bioinform Syst Biol*, page 64628, 2007.

C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17 Suppl 1:S316–22, 2001.

J. Zhang, L. Wei, X. Feng, Z. Ma, and Y. Wang. Pattern expression non-negative matrix factorization: Algorithm and application to blind source separation. *Computational Intelligence and Neuroscience*, page Artical ID 168769, 2008.

Y. Zhao, M. C. Li, and R. Simon. An adaptive method for cDNA microarray normalization. *BMC Bioinformatics*, 6:28, 2005.

X. Zhou and D. P. Tuck. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, 23(9):1106–14, 2007.