2013 13th International Conference on Computational Science and Its Applications

Cancer Screening Using Multi-Modal Differential Principal Orthogonal Decomposition

Lee, Carlyn-Ann B.
Department of Mathematics
California State University, Fullerton
Fullerton, USA
cblee@csu.fullerton.edu

Lee, Charles H.
Department of Mathematics
California State University, Fullerton
Fullerton, USA
CharlesHLee@fullerton.edu

***Abstract*— Signatures of a cancer may be encrypted in DNA microarrays, and once found, can be used for diagnoses. The standard Principal Orthogonal Decomposition (POD) method has been used to effectively detect liver and bladder cancers. Supporting work demonstrated feasibility of detecting leukemia and colon cancer via extending the standard POD to use principal features extracted from cancer and healthy sets as input to Support Vector Machine (SVM). In this study, we improved screening performances with inclusion of multiple dominant extracted modes from both cancer and healthy samples. We also investigate the efficacy of combining gene expressions with their derivative information to improve the accuracy of disease detection from previous work. We report sensitivity, specificity, and accuracy from classifications using extended POD with SVM trained with weighted projections onto multiple modes extracted from cancer and normal gene expressions and their derivatives. This is equivalent to mining not only the resembling features, but also the behavioral features. By using multiple modes, classification and prediction can be more distinctively definitive. We found that, in many cases, our new approach using multi-modal POD tends to improve cancer-screening accuracy.**

Keywords*— *Principal orthogonal decomposition; data mining; support vector machine; cancer screening

I. Introduction

Expressions of thousands of individual genes can be stored in a DNA microarray, which allows one to see genes induced or repressed in an experiment. Signatures of a cancer may be encrypted in DNA microarrays, and once found, can be used for diagnoses. The Principal Orthogonal Decomposition (POD) method had been used to effectively detect liver and bladder cancers [1]-[2]. Additional work demonstrated feasibility of detecting leukemia and colon cancer via extending the standard POD to use principal features extracted from cancer and healthy sets as input to Support Vector Machine (SVM) [3],[4]. In these previous studies, the diagnoses using POD was a quantitative assessment of gene expression levels. In this study we look at relative behavior among genes and use derivative information to detect specific gene-gene interactions that might typically occur in different cancers. This type of assessment provides for a more qualitative approach to cancer screening.

We propose to include multiple dominant extracted modes from both cancer and healthy samples as input to SVM. We also investigate the efficacy of combining gene expressions with their derivative information. This is equivalent to mining the resembling features and behavioral features of cancer gene signatures. We report results from varying the number of modes extracted from cancer projections and normal projections, as well as their derivatives.

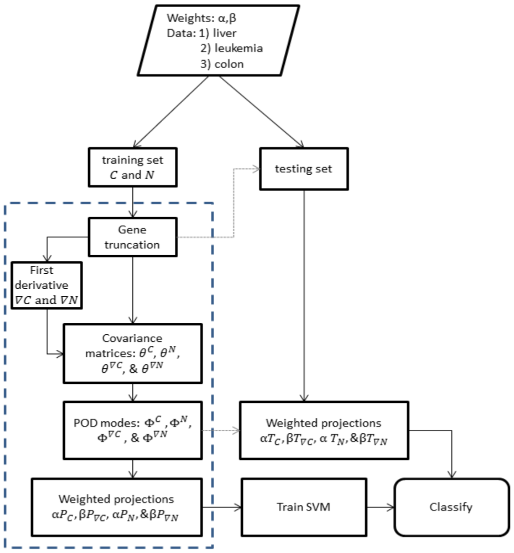
Weights: α,β
Data: 1) liver
2) leukemia
3) colon
training set
C and N
testing set
Gene truncation
First derivative
∇C and ∇N
Covariance matrices: θ^C, θ^N, $\theta^{\nabla C}$, & $\theta^{\nabla N}$
POD modes: Φ^C , Φ^N, $\Phi^{\nabla C}$, & $\Phi^{\nabla N}$
Weighted projections
αT_C, $\beta T_{\nabla C}$, $\alpha\ T_N$, &$\beta T_{\nabla N}$
Weighted projections
αP_C, $\beta P_{\nabla C}$, αP_N, &$\beta P_{\nabla N}$
Train SVM
Classify

Fig. 1. Cancer screening process with derivative information and weighted projections extracted from the primary dominant mode only. For this study, the results are averaged over 100 randomly selected training data.

II. Methodology

A. Data Sets

Liver cancer, leukemia, and colon cancer data are collected and preprocessed as described in reference [3]. The liver cancer data set contained 76 normal tissue samples and 105 primary liver cancer samples from reference [8]. Data for 5520 genes were extracted from tissue samples. Leukemia data consisted of 48 samples of Acute Myeloid Leukemia (AML) and 25 samples of Acute Lymphoblastic Leukemia (ALL) from reference [10]. The measurements were taken from 63 bone marrow samples and 10 peripheral blood samples for 7129 genes. Colon cancer data consisted of 40 cancerous samples and 22 normal samples from reference [9]. Gene expression levels for 2000 genes were extracted from epithelial cells of patients.

978-0-7695-5045-9 2013

DOI 10.1109/ICCSA.2013.16

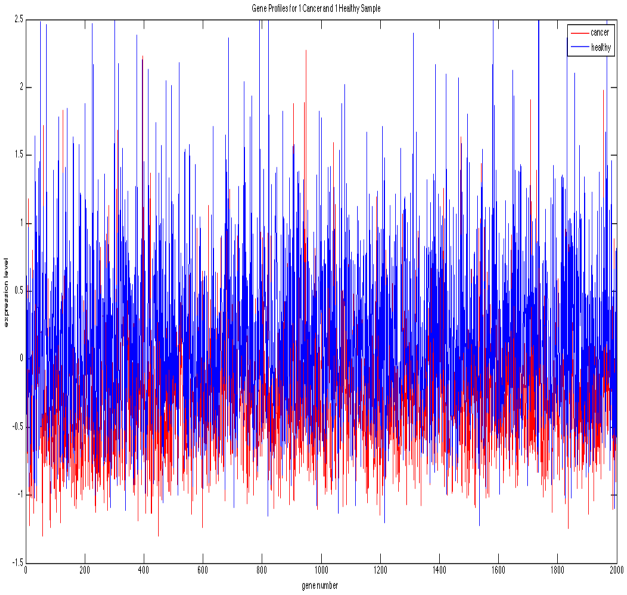

Fig. 2. One arbitrary colon cancer sample profile (blue) plotted along with one healthy sample (red). The expression levels are indicated along the y-axis for each gene on the x-axis. Cancer and normal signatures are shown in gene expression profiles. Derivative behavior between genes contributes a degree of freedom.

Mean values for each gene are subtracted off before selecting the most prominent genes for performing the orthogonal decomposition for all data. A priori information is required for this method. Given a cancer training set $\{C_k\}_{k=1}^{n_{Tr}^C}$ with n_{Tr}^C samples, and a healthy training set $\{N_l\}_{l=1}^{n_{Tr}^N}$ with n_{Tr}^N, we define the differential signal-to-noise ratio as:

$$SNR(g) = \left| \frac{\underset{1\le l\le n_{Tr}^N}{\text{mean}} \; N_l(g)}{\underset{1\le l\le n_{Tr}^N}{\text{std}} \; N_l(g)} - \frac{\underset{1\le k\le n_{Tr}^C}{\text{mean}} \; C_k(g)}{\underset{1\le k\le n_{Tr}^C}{\text{std}} \; C_k(g)} \right|. \tag{1}$$

We sort the differential SNR values for the genes in descending order and select only the genes with the highest differential SNR score for our analyses.

B. Principal Orthogonal Decomposition

In studies [1]-[4], the standard POD was computed using a L^2 inner product:

$$\langle S_i, S_j \rangle = \sum_{k=1}^{n_g} S_i(k) S_j(k), \tag{2}$$

for sample S_i and S_j with n_g genes. We can express the POD modes as a linear combination of a weighting factor $\vec{a}$ and all n samples, $\{S\}_{j=1}^{n}$:

$$\Phi_m^{POD} = \sum_{i=1}^{n} a_m \, S_i, \tag{3}$$

where $\vec{a}_m$ is the mth dominant eigenvector of the covariance matrix, θ, with elements $\theta_{ij}=\langle S_i S_j \rangle$. The Eigen theory approach was used in references [5] and [6] to combine signals from antenna arraying models in deep space communications. The idea behind [5] and [6] is to find optimal weight coefficients from noise statistics to maximize signal to noise ratio in received observables. A forward linear combining of channel outputs and weighting coefficients are used to extract the resulting signal.

Similarly, in this study, we use gene profile statistics to find weighting coefficients corresponding to the eigenvalue with the maximum signal to noise ratio. We construct the covariance matrices for cancer gene expressions, derivatives of the cancer gene expressions, healthy gene expressions, and derivatives of the healthy gene expressions. We extract modes from cancer and normal training sets, $\{C_k\}_{k=1}^{n_{Tr}^C}$ and $\{N_l\}_{l=1}^{n_{Tr}^N}$, and their first derivatives, $\{\nabla C_k\}_{k=1}^{n_{Tr}^C}$ and $\{\nabla N_l\}_{l=1}^{n_{Tr}^N}$. The elements of covariance matrix, θ, for each of the sets are defined by the inner products:

$$\begin{aligned} \theta_{ij}^{C} &= \langle C_i, C_j \rangle, \\ \theta_{ij}^{\nabla C} &= \langle \nabla C_i, \nabla C_j \rangle, \\ \theta_{ij}^{N} &= \langle N_i, N_j \rangle, \text{ and} \\ \theta_{ij}^{\nabla N} &= \langle \nabla N_i, \nabla N_i \rangle. \end{aligned} \tag{4}$$

From the covariance matrices we extract the most dominant POD modes Φ_m^C, $\Phi_m^{\nabla C}$, Φ_m^N, and $\Phi_m^{\nabla N}$ respectively, where $1 \le m \le n_{Tr}^{NC} = \min\{n_{Tr}^C, n_{Tr}^N\}$. Note that the training samples, $Tr = \{C_k\}_{k=1}^{n_{Tr}^C} \cup \{N_l\}_{l=1}^{n_{Tr}^N}$, are used to construct the dominant POD modes: $\{\Phi_m^C, \Phi_m^{\nabla C}, \Phi_m^N, \Phi_m^{\nabla N}\}_{m=1}^{n_{Tr}^{NC}}$.

We choose at random n_{Tr}^N healthy samples and n_{Tr}^C cancer samples to compose training set, Tr, such that the training sample size is $n_{Tr} = n_{Tr}^N + n_{Tr}^C$. The remaining samples are used to compose the testing set, Te, which consists of n_{Te}^C cancer testing samples and n_{Te}^N normal samples. The testing set, Te, has sample size $n_{Te} = n_{Te}^C + n_{Te}^N$. We project the samples from testing set, Te, onto the dominant POD modes extracted from training set, Tr.

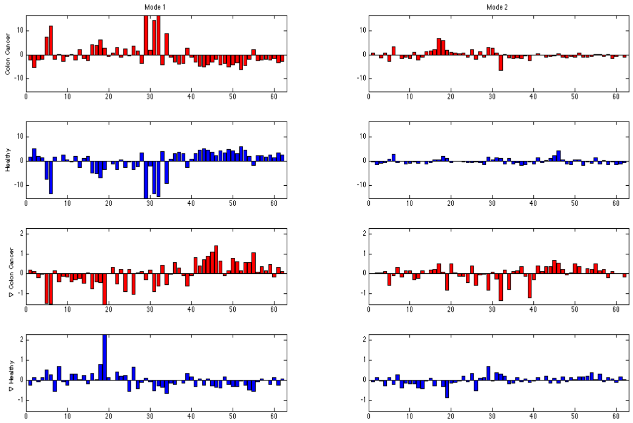

Fig. 3. The primary dominant mode from only the cancer and healthy gene expressions (first column, rows 1 and 2 respectively) show the best indicators for classification, however additional projections can contribute to improved results.

Figure 3 shows projections of all samples on to the first and second modes extracted from training with colon cancer samples (red) and healthy samples (blue) and derivatives (notated with "∇"). Cancer projections are labeled along the horizontal axis from 1-40 and healthy projections along 41-62. The projections for each sample are typically correlated with the training information. The primary dominant mode from only cancer and normal gene expressions show the best indicators for classification, however additional projections can contribute to improved results.

C. *Weighted Projections*

Given any arbitrary sample, S, we apply Gram Schmidt orthogonalization to compute its projection onto the dominant mth POD mode(s). This has been done in reference [2], however in this approach we evaluate the projections onto behavioral information as well as the gene expression levels. The projections are defined by the inner products:

$$\begin{aligned} P_m^C(S) &= \langle S, \Phi_m^C \rangle, \\ P_m^{\nabla C}(S) &= \langle \nabla S, \Phi_m^{\nabla C} \rangle, \\ P_m^N(S) &= \langle S, \Phi_m^N \rangle, \text{ and} \\ P_m^{\nabla N}(S) &= \langle \nabla S, \Phi_m^{\nabla N} \rangle. \end{aligned} \tag{5}$$

Each of these projections has capability to screen for cancers. Figure 4 summarizes the cases where cancer signatures are detected from the projection of arbitrary sample, S, onto a given mode, m. The left column illustrates cases for the projection of sample, S, to infer a healthy sample, whereas the right column demonstrates cases where a cancer signature is detected in the sample.

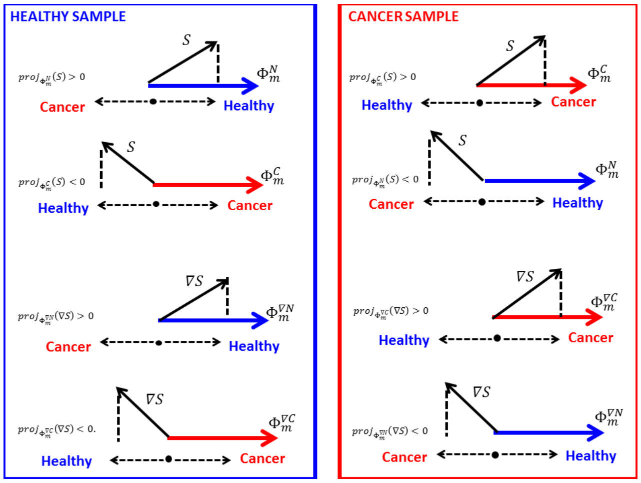

Fig. 4. Projections of an arbitrary sample, S, and its derivative, ∇S, are classified as healthy or cancer depending on their projection over extracted modes. The left column shows cases for a sample that is classified as healthy and the right column shows cases for detection of a cancer signature.

A collection of all the projections discussed in eq. (5), $\{\mathrm{P}_m^C(Tr), \mathrm{P}_m^{\nabla C}(Tr), \mathrm{P}_m^N(Tr), \mathrm{P}_m^{\nabla N}(Tr)\}_{m=1}^{n_{Tr}^{NC}}$, are used to extract the representative features for training set, Tr. Figure 5 shows the combination of all projections from qualitative and quantitative information used in this study. Rows highlighted in purple are the projections extracted from quantitative data, while rows highlighted in green are the projections extracted from qualitative information.

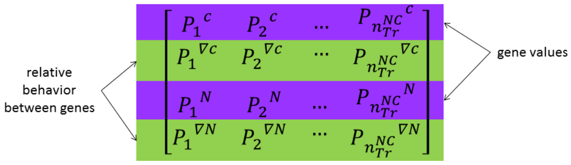

Fig. 5. Projections for all data are evaluated for four modes: projections from cancer modes (row 1), projections from derivative cancer modes (row 2), projections from healthy modes (row 3), and projections from derivative healthy modes (row 4). These projections form the bases for the feature set.

Individual projections from eq. (5) can screen for cancer but with inherent errors. In this study, we combine the each of the individual projections to increase the degrees of freedom in representations of each sample. Effectively, we use a combination of multiple projections to compensate for inherent errors present in screening with one projection only. The previous studies, [3] and [4], have used this approach but combine only quantitative data extracted from gene expression levels. This study extracts signals from qualitative information found in ∇S and combines them with quantitative data in S.

Furthermore we apply weights α and β to vary the emphasis each projection has on the characterization of the sample. In this study, we wish to determine how qualitative information contributes to improved results. By varying a weight, β, we can see how gene behavior affects overall performance. The weight, α, is applied to the quantitative data.

For training set, Tr, the resulting features extracted with multimodal differential POD, for each mode used, is a matrix of size $4 \times n_{Tr}$:

$$\vec{x}_m^{proj}(Tr) = \begin{bmatrix} \alpha \cdot P_m^C(Tr) \\ \beta \cdot P_m^{\nabla C}(Tr) \\ \alpha \cdot P_m^N(Tr) \\ \beta \cdot P_m^{\nabla N}(Tr) \end{bmatrix}_{4 \times n_{Tr}} . \tag{6}$$

Note that α and β are weights applied to the projections from gene expression values and derivatives of the gene expressions respectively. The underlying assumption is that certain behavioral patterns of a gene's profile can characterize cancer or healthy signatures. The resulting feature vector, $\vec{x}_m^{proj}(Tr)$, for arbitrary training sample, Tr, uses both behavioral patterns with gene expression values to contribute a degree of freedom for stronger characterization of the signatures. Weighting $P_m^{\nabla C}(Tr)$ and $P_m^{\nabla N}(Tr)$ by β increases or decreases the contribution of relative gene behavior to characterization of expression profiles in the feature extraction.

Using the dominant, M, modes the representative matrix is constructed with $\vec{x}_m^{proj}(Tr)$ for all modes $m = 1 \dots M$. The final representation of the training data, Tr, is a matrix with dimensions $(4M) \times n_{Tr}$:

$$\vec{X}_M^{proj}(Tr) = \begin{bmatrix} \vec{x}_1^{proj}(Tr) \\ \vdots \\ \vec{x}_M^{proj}(Tr) \end{bmatrix}_{(4M)\times n_{Tr}} . \tag{7}$$

Inclusion of projections over multiple modes contributes additional degrees of freedom, which typically results in stronger characteristic feature vectors, and can be used to screen for cancer signatures with higher accuracy.

D. POD With Weighted Modes

Using this method, projections are defined by the inner products of arbitrary sample , S , and sum of weighted modes, m, from quantitative and qualitative information:

$$\tilde{P}_m^C(S) = \langle S, \alpha \cdot \Phi_m^C + \beta \cdot \Phi_m^{\nabla C} \rangle \quad \text{and}$$
$$\tilde{P}_m^N(S) = \langle S, \alpha \cdot \Phi_m^N + \beta \cdot \Phi_m^{\nabla N} \rangle. \tag{8}$$

This approach applies weight, α, to the modes extracted from gene expression levels, Φ_m^C and Φ_m^N , and weight, β, to modes from derivative behavior, $\Phi_m^{\nabla C}$ and $\Phi_m^{N\nabla}$. With this approach we construct definitive bases of cancer and healthy samples through a combination of both gene expression profiles and the behavioral patterns of the gene expressions.

The objective for this approach is to construct stronger and more representative feature vectors by projecting data onto bases that best define the cancer or healthy signatures. A weight, $\beta > 0$, includes behavioral information in the construction of bases vectors over which we project arbitrary data sample, S. The resulting features from applying this method, for each mode used, results in a matrix with dimensions $2 \times n_{Tr}$:

$$\vec{x}_m^{mode}(Tr) = \begin{bmatrix} \tilde{P}_m^C(Tr) \\ \tilde{P}_m^N(Tr) \end{bmatrix}_{2\times n_{Tr}} . \tag{9}$$

We lose a degree of freedom compared to the method described in section C, however with this approach we are able to extract features that are more representative of inherent characteristics of cancer and healthy samples. We can include additional degrees of freedom by using multiple modes. The feature extraction using the dominant, M, modes is described by the matrix with dimensions $(2M) \times n_{Tr}$:

$$\vec{X}_M^{mode}(Tr) = \begin{bmatrix} \vec{x}_1^{mode}(Tr) \\ \vdots \\ \vec{x}_M^{mode}(Tr) \end{bmatrix}_{(2M)\times n_{Tr}} . \tag{10}$$

E. Classification

In this study we use a linear SVM that constructs a hyper plane described by:

$$\langle \vec{w}, \vec{x} \rangle - b = 0. \tag{11}$$

As shown in Figure 6, the weighting vector, $\vec{w}$, maximizes the distance between the hyper plane and support vectors from two classes of data. Namely, we construct a hyper plane to separate the cancer and normal training samples extracted with multimodal differential POD. To determine $\vec{w}$ and b, SVM takes as input the resembling matrix, X , and its class information, γ_{Tr}:

$$\begin{bmatrix} b \\ \vec{w} \end{bmatrix} = SVM(X, \gamma_{Tr}). \tag{12}$$

Note that the characterizing matrix, X, is extracted from raw training set, Tr, using the multimodal differential POD and its class information, γ_{Tr}:

$$X = \begin{cases} \vec{X}_M^{proj}(Tr) & \begin{matrix} weighted \\ projections \end{matrix} \\ \vec{X}_M^{mode}(Tr) & \begin{matrix} weighted \\ modes \end{matrix} \end{cases}, \gamma_{Tr} = \begin{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n_{Tr}^C \times 1} \\ \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n_{Tr}^N \times 1} \end{bmatrix} . \tag{13}$$

In column vector, γ_{Tr}, an entry of "1" corresponds to a cancer sample and "0" corresponds to a healthy sample.

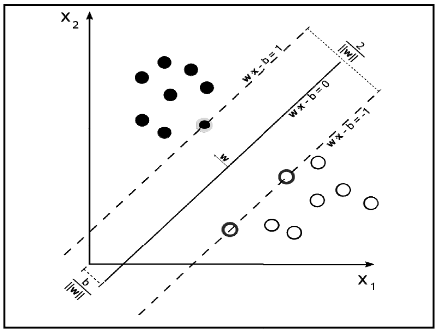

Fig. 6. The separating line with maximum distances from two classes is constructed using SVM algorithm and is used for classification [7].

Matrix, X, is extracted with one of the transformations described in (7) and (10), $X = \vec{X}_M^{proj}(Tr)$ or $X = \vec{X}_M^{mode}(Tr)$, and have dimensions $4M \times n_{Tr}$ or $2M \times n_{Tr}$ respectively. Figure 7 shows that observing multiple dimensions can improve separation between two-class data. Similarly, we increase the number of modes, M, to add additional degrees of freedom which can possibly improve cancer detection.

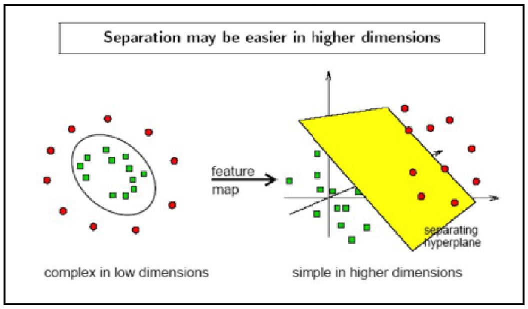
Separation may be easier in higher dimensions
feature map
separating hyperplane
complex in low dimensions
simple in higher dimensions

Fig. 7. Separation between two classes may be more distinct in higher dimensions [7]. Similarly, we increase the number of modes, M, used for feature extraction to gain additional degrees of freedom.

We use b and $\vec{w}$ from the hyper plane described by (11) to construct a SVM classification rule. The SVM classification rule is used to screen for cancer signatures in input matrix, Y, which represents the raw data extracted from testing set, Te:

$$\gamma(Y) := \{sgn(\langle \vec{w}, Y \rangle - b) > 0\}. \tag{14}$$

Note that the characterizing matrix, Y, is extracted from raw testing data, Te, using multimodal differential POD. We require the definitions of matrix, Y, and its targets, γ_{Te}^{target}, to screen for cancer signatures with the SVM classification and to evaluate accuracy of the screening:

$$Y = \begin{cases} \vec{X}_M^{proj}(Te) & \begin{matrix} weighted \\ projections \end{matrix} \\ \vec{X}_M^{mode}(Te) & \begin{matrix} weighted \\ modes \end{matrix} \end{cases}, \gamma_{Te}^{target} = \begin{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n_{Te}^C \times 1} \\ \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n_{Te}^N \times 1} \end{bmatrix}. \tag{15}$$

We apply the transformation to obtain X from Tr in definition (13) and apply it to obtain Y , which has the same number of columns as X .

Output, $\gamma(Y)$, from equation (14) returns a column vector of size $n_{Te} \times 1$ with elements, "0" and "1," corresponding to healthy and cancer SVM classifier diagnoses respectively. We compare output, $\gamma(Y)$, with the testing set targets, γ_{Te}^{target}, to evaluate sensitivity, specificity, and accuracy based on the following definitions:

$$sensitivity = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ negatives} \tag{16}$$

$$specificity = \frac{\#\ of\ true\ negatives}{\#\ of\ true\ negatives + \#\ of\ false\ positives}$$

$$accuracy = \frac{\#\ of\ true\ negatives + \#\ of\ true\ positives}{\#\ of\ all\ samples}$$

III. Results

We observe results of varying weights in (6) and (8), using $\beta = \{0.01, 0.1, 1, 10, 100\}$, and $\alpha = 1$ held constant. In doing so we increase or decrease presence of derivative information in the extracted features. A large β emphasizes the contribution of qualitative information gleaned from relative gene behaviors, whereas a small β maintains strong emphasis on quantitative data present in gene expression levels. As a baseline reference results are recorded for $\alpha = 1$ and $\beta = 0$, which corresponds to classification using the method for extended POD described in references [3] and [4]. Results for weighted projections with derivative information are compared to scores from the baseline reference to show performance improvements from previous methods. Results are also evaluated for projections using only behavioral information, where $\alpha = 0$ and $\beta = 1$, to obtain a purely qualitative assessment.

For this study we apply only the linear SVM kernel, however the multimodal differential POD does not guarantee that the extracted features of two-class data are exactly linearly separable. Further investigation of other SVM kernels, hyperbolic, polynomial, or Gaussian radial basis functions, may improve results.

Average sensitivity, specificity, and accuracy are recorded for 100 random trials using the methods described in section C and D, for all observed pairs, (α, β), and for $M = 1, 2, and\ 3$. A combined geometric mean score, GM, is evaluated from average sensitivity, specificity, and accuracy from eq. (16) for 100 trials:

$$GM = \sqrt[3]{sensitivity \times specificity\ \times accuracy}\,.$$

The combined performance scores are evaluated to compare generalized performances across data sets for pairs α and β and use of additional, M, modes:

$$Combined\ Perf. = \sqrt[3]{GM_{\ (liver)} \times GM_{(leuk)}\ \times GM_{(colon)}}\,.$$

Average sensitivity, specificity, and accuracy for 100 trials are recorded in tables 1 and 2 for weighted projections and weighted modes respectively. For each trial we use 80% of the healthy and cancer samples at random to compose, Tr. Gene truncation is performed to 40 genes with the highest SNR using eq. (1) from samples of Tr. The remaining 20% of samples are used to compose Te. Screening is performed for liver cancer, colon cancer, and leukemia, using each weighted pair (α, β) , and $M = 1, 2, and\ 3$.

A. Weighted Projections

Combined performances for weighted projections are summarized in Figure 8. Compared to the baseline reference using only one mode, slight performance gain is observed when the reduced feature vector contains weighted projections from derivative information. Using only one mode, the results show approximately %1 with inclusion of derivative information with any weight, β > 0. Varying weight, β, only scales the feature vector in the derivative dimensions without affecting the ability to classify data.

Low performance is observed when only the derivative is used (=0 and =1) and infers that derivative information is not as strong of an indicator as gene expressions alone. For two and three modes, performance decreases after introducing weighted projections from derivative information.

B. Weighted Modes

Combined performances for weighted modes are summarized in Figure 9. The performance score for the case =0 and =1 infer that the derivative information can contribute significantly to performance. We can see the effects of derivative information by observing the performance of one mode when >0. Compared with the baseline reference using one mode, a gain is observed in the combined performance scores when derivative information is introduced in the feature extraction. Performance steadily increases as the observed values of β increases to $\beta = 10$. Using only the primary dominant modes, the improvement is approximately 2% compared to the baseline reference. The optimal score for one mode is comparable to the performance score for the baseline reference for three modes by ~0.4%.

For multiple modes, a slight gain can be obtained by adding weighted derivative information. Using two modes, there is no improvement in combined performance score for the β values observed in this study. However, for three modes, there is a slight gain of approximately 0.3% for combined performance score when the weight, $\beta = 0.1$. We infer that it may be possible to optimize the performance, using a minimal number of modes, by selecting the appropriate weight, β.

IV. Conclusion

In this study, we show that cancer screening can be improved using qualitative gene profile behavior in addition to the quantitative values of gene expressions. It was shown in reference [3] and [4] that accuracy typically increases using multiple projections extracted from raw gene expressions with extended POD. These performance accuracies from methods [3] and [4] are summarized in the baseline references in this study. As shown in the results, including derivative information can further improve performance accuracy from previous methods.

The results demonstrate that performance gain obtained through use of additional modes can also be achieved by an optimal minimum number of modes. This may be approached by including derivative information to the modes and selecting appropriate weights. We believe it may also be possible to optimize performance accuracy using only the primary dominant features to obtain comparable results from baseline performance observed with multiple modes by selecting the appropriate weights. The solution to this optimization problem can be beneficial for classifying large datasets since the computational complexity of the extended POD increases as additional modes are used.

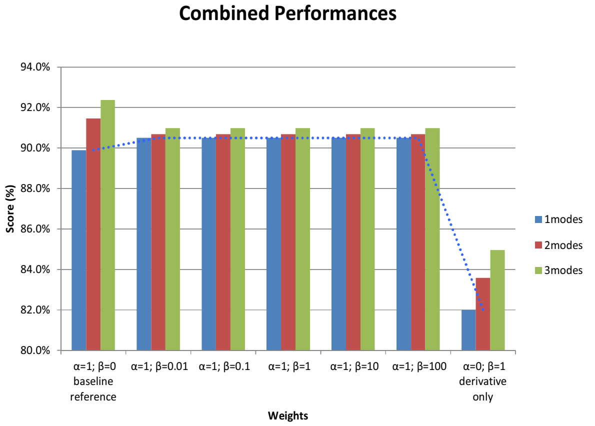

Fig. 8. Combined performance scores for weighted projections. Derivative information contributes to ~1% gain in performance, however it is a weak screening method when used alone (=0 and =1).

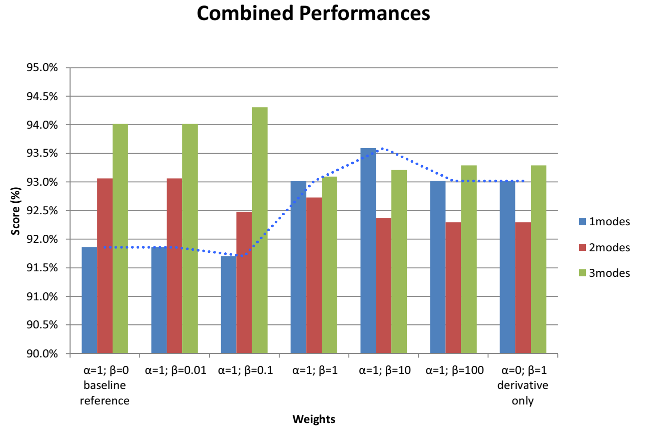

Fig. 9. Combined performance for weighted modes improve with qualitative information. There may be an optimal weight, , that maximizes performance.

Acknowledgment

Carlyn-Ann Lee received partial support for travel from the CSU Program for Education and Research in Biotechnology (CSUPERB)

References

[1] D. Peterson and C.H. Lee, A DNA-based Pattern Recognition Technique for Cancer Detection, Proceedings of the 26th Annual Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 2 (2004) pp. 2956-2959.

[2] C. H. Lee and N. Abbasi, Feature Extraction Techniques on DNA Microarray Data for Cancer Detection, 2007 World Congress on Bioengineering Proceedings, Bangkok, Thailand, July 2007.

[3] C. B. Lee and C. H. Lee, Extended Principal Orthogonal Decomposition Method for Cancer Screening, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 2, No. 2, March 2012.

[4] C. B. Lee, C. H. Lee, Efficacy of the extended principal orthogonal decomposition method on DNA microarray data in cancer detection, 2012 4th International Conference on Bioinformatics and Biomedical Technology, Singapore, Feb 2012.

[5] K.-M. Cheung, Eigen Theory for Optimal Signal Combining: A Unified Approach, The Telecommunications and Data Acquisition Progress Report 42-126, April-June 1996, Jet Propulsion Laboratory, Pasadena, California, pp. 1-9, August 15, 1996.

[6] C.H. Lee, V. Vilnrotter, E. Satorius, Z. Ye, D. Fort, and K.-M. Cheung, Large Array Signal Processing for Deep-Space Applications, JPL Interplanetary Network Progress Report 42-150, Jet Propulsion Laboratory, Pasadena, CA, pp. 1-28, August 15, 2002.

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] X. Chen, et. al. Gene expression patterns in human liver cancers. Molecular Biology of the Cell. 2002, 13: 1929-1939.

[9] U. Alon, et al., Broad patterns of gene expression revealed by clustering analysis of cancer and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. 96: 6745–6750.

[10] T. R. Golub, D. K. Slonim, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999, 286: 531-537.

TABLE I. RESULTS FOR 100 TRIALS (WEIGHTED PROJECTIONS)

Modes		α=1; β=0 (reference)			α=1; β=0.01			α=1; β=0.1			α=1; β=1			α=1; β=10			α=1; β=100			α=0; β=1 (derivative)		
		liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon
1	sensitivity	95.6%	97.0%	77.9%	96.0%	97.4%	80.8%	96.0%	97.4%	80.8%	96.0%	97.4%	80.8%	96.0%	97.4%	80.8%	96.0%	97.4%	80.8%	88.7%	82.0%	68.9%
	specificity	98.4%	97.8%	76.1%	97.6%	94.8%	78.4%	97.6%	94.8%	78.4%	97.6%	94.8%	78.4%	97.6%	94.8%	78.4%	97.6%	94.8%	78.4%	92.2%	85.2%	78.7%
	accuracy	96.8%	97.3%	77.2%	96.7%	96.5%	79.8%	96.7%	96.5%	79.8%	96.7%	96.5%	79.8%	96.7%	96.5%	79.8%	96.7%	96.5%	79.8%	90.2%	83.1%	72.3%
	combined mean	96.9%	97.3%	77.0%	96.7%	96.2%	79.6%	96.7%	96.2%	79.6%	96.7%	96.2%	79.6%	96.7%	96.2%	79.6%	96.7%	96.2%	79.6%	90.4%	83.4%	73.2%
	combined perf.	89.9%			90.5%			90.5%			90.5%			90.5%			90.5%			82.0%		
2	sensitivity	96.3%	96.4%	82.0%	96.0%	97.3%	80.4%	96.0%	97.3%	80.4%	96.0%	97.3%	80.4%	96.0%	97.3%	80.4%	96.0%	97.3%	80.4%	91.3%	84.3%	70.6%
	specificity	98.8%	95.6%	81.4%	97.9%	94.4%	79.9%	97.9%	94.4%	79.9%	97.9%	94.4%	79.9%	97.9%	94.4%	79.9%	97.9%	94.4%	79.9%	94.5%	86.0%	78.3%
	accuracy	97.4%	96.2%	81.7%	96.8%	96.3%	80.2%	96.8%	96.3%	80.2%	96.8%	96.3%	80.2%	96.8%	96.3%	80.2%	96.8%	96.3%	80.2%	92.6%	84.9%	73.3%
	combined mean	97.5%	96.1%	81.7%	96.9%	96.0%	80.1%	96.9%	96.0%	80.1%	96.9%	96.0%	80.1%	96.9%	96.0%	80.1%	96.9%	96.0%	80.1%	92.8%	85.1%	74.0%
	combined perf.	91.5%			90.7%			90.7%			90.7%			90.7%			90.7%			83.6%		
3	sensitivity	96.3%	96.8%	85.0%	96.0%	97.6%	81.0%	96.0%	97.6%	81.0%	96.0%	97.6%	81.0%	96.0%	97.6%	81.0%	96.0%	97.6%	81.0%	92.4%	87.8%	70.9%
	specificity	98.6%	94.8%	83.8%	97.7%	94.6%	80.8%	97.7%	94.6%	80.8%	97.7%	94.6%	80.8%	97.7%	94.6%	80.8%	97.7%	94.6%	80.8%	95.2%	87.6%	79.4%
	accuracy	97.2%	96.2%	84.5%	96.7%	96.6%	80.8%	96.7%	96.6%	80.8%	96.7%	96.6%	80.8%	96.7%	96.6%	80.8%	96.7%	96.6%	80.8%	93.6%	87.7%	73.8%
	combined mean	97.4%	95.9%	84.4%	96.8%	96.2%	80.9%	96.8%	96.2%	80.9%	96.8%	96.2%	80.9%	96.8%	96.2%	80.9%	96.8%	96.2%	80.9%	93.7%	87.7%	74.6%

Sensitivity, specificity, accuracy recorded for 100 random trials using weighted projections. Combined performances are highlighted for comparison. Compared with the baseline reference, there is a slight gain (~1%) in combined performance for one mode when a any weight, β>0, is used. Decreased performance for weights, $\alpha = 0$ and $\beta = 1$ infer that derivative information is not as strong of an indicator as gene expressions.

TABLE II. RESULTS FOR 100 TRIALS (WEIGHTED MODES)

Modes		α=1; β=0 (reference)			α=1; β=0.01			α=1; β=0.1			α=1; β=1			α=1; β=10			α=1; β=100			α=0; β=1 (derivative)		
		liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon	liver	leuk	colon
1	sensitivity	95.7%	100.0%	82.5%	95.7%	100.0%	82.5%	95.7%	100.0%	82.5%	95.7%	100.0%	83.8%	95.7%	100.0%	86.3%	95.7%	100.0%	86.3%	95.7%	100.0%	86.3%
	specificity	99.3%	96.0%	79.0%	99.3%	96.0%	79.0%	99.3%	98.0%	76.5%	99.3%	100.0%	81.0%	99.3%	100.0%	81.5%	99.3%	96.0%	81.5%	99.3%	96.0%	81.5%
	accuracy	97.3%	98.6%	81.6%	97.3%	98.6%	81.6%	97.3%	99.3%	80.8%	97.3%	100.0%	83.1%	97.3%	100.0%	84.7%	97.3%	98.6%	84.7%	97.3%	98.6%	84.7%
	combined mean	97.4%	98.2%	81.0%	97.4%	98.2%	81.0%	97.4%	99.1%	79.9%	97.4%	100.0%	82.6%	97.4%	100.0%	84.1%	97.4%	98.2%	84.1%	97.4%	98.2%	84.1%
	combined perf.	91.9%			91.9%			91.7%			93.0%			93.6%			93.0%			93.0%		
2	sensitivity	96.2%	99.0%	86.3%	96.2%	99.0%	86.3%	96.2%	99.0%	83.8%	96.2%	98.0%	83.8%	96.2%	99.0%	86.3%	95.7%	99.0%	86.3%	95.7%	99.0%	86.3%
	specificity	99.3%	94.0%	84.0%	99.3%	94.0%	84.0%	98.7%	94.0%	84.0%	99.3%	94.0%	86.0%	100.0%	92.0%	81.0%	100.0%	92.0%	81.0%	100.0%	92.0%	81.0%
	accuracy	97.5%	97.3%	85.6%	97.5%	97.3%	85.6%	97.3%	97.3%	84.1%	97.5%	96.6%	84.8%	97.8%	96.6%	84.7%	97.5%	96.6%	84.7%	97.5%	96.6%	84.7%
	combined mean	97.7%	96.7%	85.3%	97.7%	96.7%	85.3%	97.4%	96.7%	84.0%	97.7%	96.2%	84.8%	98.0%	95.8%	83.9%	97.7%	95.8%	83.9%	97.7%	95.8%	83.9%
	combined perf.	93.1%			93.1%			92.5%			92.7%			92.4%			92.3%			92.3%		
3	sensitivity	96.2%	99.0%	88.8%	96.2%	99.0%	88.8%	96.2%	99.0%	88.8%	96.2%	99.0%	86.3%	96.2%	100.0%	87.5%	96.7%	100.0%	87.5%	96.7%	100.0%	87.5%
	specificity	98.0%	94.0%	88.0%	98.0%	94.0%	88.0%	98.0%	96.0%	88.0%	100.0%	94.0%	83.5%	100.0%	92.0%	83.5%	100.0%	92.0%	83.5%	100.0%	92.0%	83.5%
	accuracy	97.0%	97.3%	88.7%	97.0%	97.3%	88.7%	97.0%	98.0%	88.7%	97.8%	97.3%	85.6%	97.8%	97.2%	86.3%	98.1%	97.2%	86.3%	98.1%	97.2%	86.3%
	combined mean	97.1%	96.7%	88.5%	97.1%	96.7%	88.5%	97.1%	97.6%	88.5%	98.0%	96.7%	85.1%	98.0%	96.4%	85.8%	98.2%	96.4%	85.8%	98.2%	96.4%	85.8%
	combined perf.	94.0%			94.0%			94.3%			93.1%			93.2%			93.3%			93.3%		

Sensitivity, specificity, accuracy observed for 100 random trials with weighted modes. Combined performances are highlighted for comparison. Compared with the baseline reference, there is a slight gain (~2%) in combined performance for one mode when a large weight, β, is used. The performance score for β=10 is comparable to the performance score when additional modes are used.