2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Efficacy of the Extended Principal Orthogonal Decomposition Method on DNA Microarray Data in Cancer Detection

Carlyn Lee [1] and Charles H. Lee [2]

[1] Department of Computer Science, California State University, Fullerton

[2] Department of Mathematics, California State University, Fullerton

Abstract. Recent advances in microarray technology offer the ability to study the expression of thousands of genes simultaneously. The DNA data stored on these microarray chips can provide crucial information for early clinical cancer diagnosis. The Principal Orthogonal Decomposition (POD) method has been widely used as an effective feature detection method. In this paper, we present an enhancement to the standard approach of using the POD technique as a disease detection tool. In the standard method, cancer diagnosis of an arbitrary sample is based on its correlation value with the cancerous or normal signature extracted using the POD method on DNA microarray data. In this paper, we extend the POD method by feeding the extracted principal features into Machine Learning algorithms to detect cancer. Particularly, Linear Support Vector Machine, Feed Forward Back Propagation Networks, and Self-Organizing Maps have been used on liver cancer, colon cancer, and leukemia data. Sensitivity, specificity, and accuracy have been used as a mean to evaluate predictive abilities of the proposed extended POD methods. Our results indicate overall the proposed methods provide slight improvements over the standard POD method.

Keywords: DNA Microarray, Principal Orthogonal Decomposition, Machine Learning, Artificial Neural Networks, Support Vector Machine, Self-Organizing Map, Cancer Detection

1. Introduction

Expressions of thousands of individual genes can be stored in a DNA microarray, which allows one to see genes that are 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 standard Principal Orthogonal Decomposition (POD) method had been used to effectively detect liver and bladder cancers [1]-[2]. In this paper, we proposed to extend the standard Principal Orthogonal Decomposition (POD) method to include Machine Learning (ML) algorithms for cancer detection. Namely, we use the POD technique to extract the principal features, both cancerous and normal. We then feed them to ML algorithms such as the Support Vector Machine (SVM), Feed Forward Back Propagation Networks (FFBPN) and Self-Organizing Map (SOM) to train the classifiers for detection of different types of cancers.

2. Predictive Classifiers using the Extended POD Methods

Given a cancer training set $\{T_i^C\}_{i=1}^{N_C}$ and a normal training set $\{T_j^N\}_{j=1}^{N_N}$, we apply the POD technique to extract the primary dominant features Φ^C and Φ^N, respectively. We use $\{T_k\}_{k=1}^{N_C+N_N}$ to include both training sets, whose elements can be represented as,

$$X^{(k)} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ y^{(k)} \end{bmatrix}, \text{ where } x_1^{(k)} = \left\langle T_k, \Phi^C \right\rangle, x_2^{(k)} = \left\langle T_k, \Phi^N \right\rangle, \text{ and } y^{(k)} = \begin{cases} 0 & T_k \in \left\{T_j^N\right\}_{j=1}^{N_N} \\ 1 & T_k \in \left\{T_i^C\right\}_{i=1}^{N_C} \end{cases}. \quad (1)$$

The following Machine Learning algorithms are used to construct classifiers F based on the values of $\{X^{(k)}\}_{k=1}^{N_C+N_N}$ of the training sets so that $F(X^{(k)})=y^{(k)}$ for as many k as possible. We denote by $\{S_m^C\}_{m=1}^{M_C}$, $\{S_n^N\}_{n=1}^{M_N}$, $\{S_l\}_{l=1}^{M_C+M_N}$ the test sets of cancer, normal, and both, respectively. For each member of the testing set, we define the corresponding metric

$$X^{(l)}=\begin{bmatrix} x_1^{(l)} \\ x_2^{(l)} \\ y^{(l)} \end{bmatrix}, \text{ where } x_1^{(l)}=\left\langle S_l,\Phi^C\right\rangle, x_2^{(l)}=\left\langle S_l,\Phi^N\right\rangle, \text{ and } y^{(l)}=\begin{cases} 0 & S_l\in\left\{S_n^N\right\}_{n=1}^{M_N} \\ 1 & S_l\in\left\{S_m^C\right\}_{m=1}^{M_C} \end{cases}. \tag{2}$$

2.1. Linear Support Vector Machine

Since we perform our projections onto the dominant cancer and normal POD features, the hyper-plane is two dimensional and SVM draws a contour between the cancerous and normal classes [3]. For simplicity, we assume that the training data is linearly separable and utilize a linear SVM. The SVM algorithm constructs the line $y=mx+b$ that maximizes the margin between the positive and negative groups. In this case, the classifier is given by

$$F_{SVM}(X^{(l)})=\begin{cases} 0 & x_2^{(l)}>mx_1^{(l)}+b \\ 1 & \text{otherwise} \end{cases}. \tag{3}$$

2.2. Feed Forward Back Propagation

For the Feed Forward Back Propagation Network (FFBPN), we assume a simple, single layer perceptron with two inputs and one output (see [4] for more details). The FFBPN is constructed using the MATLAB command "newff", where the weights (w_1,w_2) and the bias parameter θ are found based on the training sets. The network architecture is activated by a hyperbolic tangent sigmoid function,

$$F_{FFBPN}(X^{(l)})=\begin{cases} 0 & G(w_1x_1+w_2x_2+\theta)<\tau_{cutoff} \\ 1 & \text{otherwise} \end{cases} \text{ where } G(s)=\frac{e^s-e^{-s}}{e^s+e^{-s}}. \tag{4}$$

Note that the cut-off value τ_{cutoff} is determined from the Receiver-Operating-Characteristic (ROC) curve described in Section 2.5.

2.3. Self-Organizing Map

SOM starts out with an initial two-dimensional map and, when introduced to the training set, it updates the map iteratively to fit the distribution of the clusters in the training set. When a testing set is fed into the map, the map classifies it according to its nearest cluster of the training set. We implement the SOM scheme using all four neighborhood functions (Bubble, Gaussian, Cut-Gaussian, and Epanechicov) sequentially to exhaust all possible maps. Both the batch and the sequential training algorithms are also explored in this study. SOMs are implemented using the somtoolbox (see [5] for further details).

2.4. Performance Measures

Sensitivity, specificity, and accuracy are used to determine the performance of classifiers in this study. Sensitivity measures the ability to correctly identify those with the disease, whereas specificity measures the ability to identify those without the disease. Accuracy shows the ratio of true predictions (true positives and true negatives) out of all predictions. For all test set predictions, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are determined. Sensitivity, specificity, and accuracy are evaluated to determine the quality of the network:

$$\text{Sensitivity}=\frac{TP}{TP+FN}, \quad \text{Specificity}=\frac{TN}{TN+FP}, \quad \text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}. \tag{5}$$

2.5. Cut-off Thresholds

Note that our predictive non-binary classifiers, such as those using the standard POD and the FFBPN, produce output that varies from 0 to 1. To produce a meaningful cut-off threshold τ , we exhaustively search a value between 0 and 1 with small increments that attains the highest fitness. In our case the fitness value, based on the interest of the ROC curve, is defined as

$$f_{fitness}(\tau) = \text{Sensitivity}(\tau) \times \text{Specificity}(\tau)\,. \tag{6}$$

3. Data Sets

For liver cancer detection, we examined the DNA microarray data from reference [6]. The data, containing both normal and cancerous tissues, are obtained from the Stanford Microarray Database at *genome-www5.stanford.edu.* Only genes with expressions in over 80% of the samples are included. Missing data for a particular gene are imputed with the average of the values for that gene from the other samples. The liver cancer data set contained data from 76 normal tissue samples and 105 primary liver cancer samples, where data for 5520 genes are extracted.

For colon cancer detection, we examined DNA microarray data from reference [7]. Colon cancer data consisted of 40 cancerous samples and 22 normal samples. Samples are taken from epithelial cells of colon cancer patients. The original data contained 6000 gene expression levels. Only 2000 gene expression levels are used based on the confidence in the measured expression levels.

For leukemia detection, we examined DNA microarray data from reference [8]. Leukemia data consisted of 48 samples of Acute Myeloid Leukemia (AML) and 25 samples of Acute Lymphoblastic Leukemia (ALL). The measurements are taken from 63 bone marrow samples and 10 peripheral blood samples. Data for 7129 gene expression levels are extracted.

Mean values for each gene are subtracted off before selecting the most prominent genes for performing the orthogonal decomposition for all data. We define the Signal-to-Noise ratio for each gene g as

$$SNR(g) = \left| \frac{\underset{1 \le i \le N_C}{mean}(T_i^C(g))}{\underset{1 \le i \le N_C}{std}\,(T_i^C(g))} \right| + \left| \frac{\underset{1 \le j \le N_C}{mean}(T_j^N(g))}{\underset{1 \le j \le N_C}{std}\,(T_j^N(g))} \right|. \tag{7}$$

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

4. Methods

The top 10 prominent genes with the highest SNR scores are used for analysis. Samples are randomly partitioned into training and testing sets. Training sets consist of 90% of cancerous samples and 90% of normal samples. The remaining samples are used for testing. The projections onto the POD cancer and normal features are normalized from 0 to 1 and cut-off thresholds are selected to obtain maximum fitness (6).

Predictions for the testing set are made using the highest fitness value defined in equation (6). Training, validation and testing processes are repeated multiple times with randomly selected partitions. Averaged sensitivity, specificity and accuracy from these predictions are recorded.

5. Results

Results from machine learning techniques demonstrate slightly improved predictions when compared to the standard POD method. A study [9] using SVM and SOM without POD on colon and leukemia cancers has been compared to our proposed methods. Our methods produce better results (+90% versus +70%); however, there are different assumptions such as hold-out percentages (10% vs 50%) for training and testing.

5.1. Liver Cancer Data

POD feature extraction for one trial is plotted in Figure 1. The horizontal axis is the case number and the vertical axis represents its projection value. Cancerous and normal samples are numbered 1-105 and 106-181, respectively. The cut-off points are drawn as a horizontal line along the plots for each projection in Figure 1.

A large percentage of projections from cancerous tissue samples exceeded this cut-off. Similarly, a large percentage of projections from normal tissue samples are less than this cut-off. In this case, the standard POD method performs rather well as a predictive classifier. The ROC curve for training data using POD cancerous feature (blue), POD normal feature (green), and FFBPN (red) are shown in Figure 2. Points with the largest fitness values are circled and the corresponding cut-off thresholds are used for predicting the test set. In addition, we find from Figure 2 that while the FFBPN obtains a smaller false positive rate than the POD normal feature, it obtains a higher true positive rate than the POD cancer feature. The SOM method, displayed in Figure 3, indicates distinct cancerous and normal clusters. Labeled neurons show that only a small percentage of the map neurons have predictive capabilities prior to pruning. The SVM hyper-plane, shown in Figure 4, is constructed using the training set, denoted with red and green. Test data is denoted in magenta and cyan. Average accuracies for five random trials and all classifiers are shown in Table 1.

5.2. Colon Data

Prediction results from test data using our proposed methods are recorded in Table 2. Accuracy for our proposed extended POD methods exceed the recognition rate for SVM and SOM methods described in [9]. This suggests that the use of the POD as a pre-processing to the ensemble classifiers [9] may improve their overall accuracy.

5.3. Leukemia Data

In this data set, there are no normal samples and the classes are AML and ALL. Here we treat the ALL samples as if they are normal samples. Results from AML and ALL predictions are shown in Table 3. Accuracy of POD predictions improved only slightly using machine learning techniques. Accuracy for this data using SVM and SOM pre-processed with POD exceed recognition rate for feature selection methods proposed in [9]. Furthermore, using POD feature reduction to pre-process this data obtains slightly better results to a majority vote ensemble classifier [9] (97.8% versus 97.1%).

6. Conclusion

The resulting average sensitivity, specificity, and accuracy across all three data sets suggest that the proposed method is reliable for only particular data. However, the proposed method achieves over 90% accuracy with various classifiers and for a variety of data. Such high recognition rates for predictions pre-processed with POD suggest that introducing POD for use in ensemble classifiers may improve accuracy for general cancer detection [9].

7. Acknowledgements

Computation and graphics were generated with Matlab2010. SOM Toolbox was used for constructing all self organizing maps. The authors would like to thank Professor Spiros Courellis of the department of Computer Science at California State University, Fullerton for his sharing his knowledge of machine learning and inspiring this study.

8. 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] S. Abe, *Support Vector Machines for Pattern Classification,* Springer*, 2005*

[4] C. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006

[5] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas (2000), SOM Toolbox for Matlab 5, report, Helsinki Univ. of Technol., Helsinki, Finland.

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

[7] 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.

[8] 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.

[9] S. Cho and H. Won.Machine learning in DNA microarray analysis for cancer classification. *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics.* APBC '03. 2003.

9. Tables and Figures

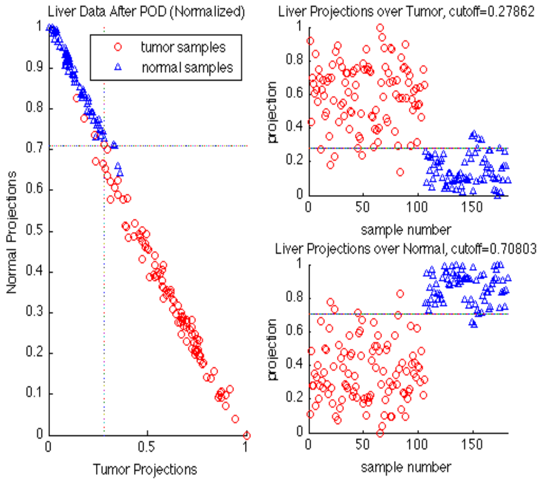

Fig. 1: Normalized data plotted with cutoffs.

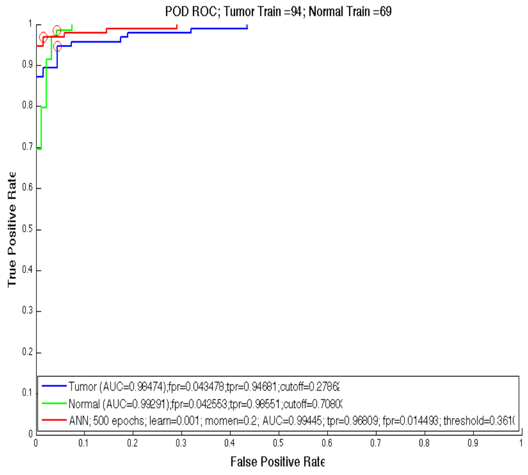

Fig. 2: ROC curve for normalized POD predictions and with FFBPN predictions. Points with max fitness are circled.

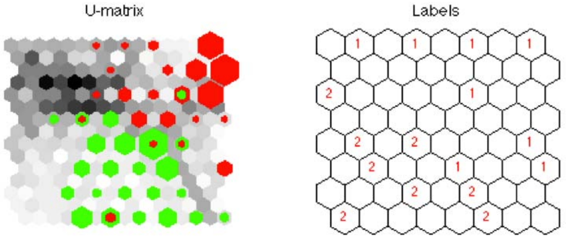

Fig. 3: SOM for Liver Cancer Data before pruning.

Left: Umatrix of Liver data clustering where red=“tumor” samples, and green=“normal” samples. Right: Neuron Labels 1=”tumor”; 2=”normal”

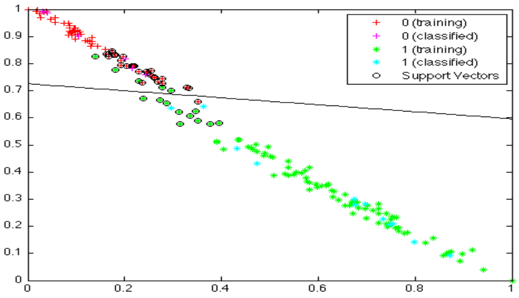

Fig 4: Hyper-plane separating tumorous and normal samples.

Table 1: Average accuracy measurements for liver cancer test set predictions over 5 trials

Measure	Method				
	POD cancer	***POD normal***	***FFBPN***	***SOM***	***SVM***
Sensitivity	0.9518	0.9620	0.9618	0.9676	0.9520
Specificity	0.9475	0.9654	0.9661	0.9800	0.9789
Accuracy	0.9502	0.9636	0.9637	0.9726	0.9636

Table 2: Average accuracy measurements for colon test set predictions over 100 random trials

Measure	Method				
	POD cancer	***POD normal***	***FFBPN***	***SOM***	***SVM***
Sensitivity	0.8050	0.8638	0.8613	0.8287	0.9293
Specificity	0.8541	0.8447	0.8640	0.7559	0.7587
Accuracy	0.8218	0.8567	0.8621	0.8029	0.8687

Table 3: Average accuracy measurements for leukemia test set predictions over 100 random trials

Measure	Method				
	POD cancer	***POD normal***	***FFBPN***	***SOM***	***SVM***
Sensitivity	0.7905	0.9695	0.9570	0.9740	0.9870
Specificity	1.00	0.9700	0.9950	0.9833	0.8583
Accuracy	0.8628	0.9713	0.9710	0.9779	0.9453