Bioinformation by Biomedical Informatics Publishing Group open access

www.bioinformation.net Hypothesis

The application of wavelet-based neural network on DNA microarray data

Jack Lee[1] and Benny Zee[1, *]

[1]Centre for Clinical Trials, School of Public Health; Department of Clinical Oncology, the Chinese University of Hong Kong, Hong Kong SAR; Benny Zee* - Email: bzee@cct.cuhk.edu.hk; *Corresponding author

received May 05, 2008; accepted December 05, 2008; accepted December 10, 2008; ;published December 31, 2008

Abstract:
The advantage of using DNA microarray data when investigating human cancer gene expressions is its ability to generate enormous amount of information from a single assay in order to speed up the scientific evaluation process. The number of variables from the gene expression data coupled with comparably much less number of samples creates new challenges to scientists and statisticians. In particular, the problems include enormous degree of collinearity among genes expressions, likely violation of model assumptions as well as high level of noise with potential outliers. To deal with these problems, we propose a block wavelet shrinkage principal component (BWSPCA) analysis method to optimize the information during the noise reduction process. This paper firstly uses the National Cancer Institute database (NC160) as an illustration and shows a significant improvement in dimension reduction. Secondly we combine BWSPCA with an artificial neural network-based gene minimization strategy to establish a Block Wavelet-based Neural Network model in a robust and accurate cancer classification process (BWNN). Our extensive experiments on six public cancer datasets have shown that the method of BWNN for tumor classification performed well, especially on some difficult instances with large-class (more than two) expression data. This proposed method is extremely useful for data denoising and is competitiveness with respect to other methods such as BagBoost, RandomForest (RanFor), Support Vector Machines (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN).

Keywords: wavelet shrinkage; denoising; ANN; classification of cancer types

Background:

DNA microarray has boosted the application in high-throughput technologies to analyze DNA, RNA or protein from tumors cells for the understanding of living systems. Many novel methods have been developed to improve the analysis to levels that classical methods have been unable to reach. The studies of gene microarray data classification in biosciences provide clear examples of how the cancer analysis in classification faces the challenge to develop computational procedures able to address specific issues, such as modeling multiple, heterogeneous populations and reducing the overwhelming number of variables (genes). In particular, Principal Component Analysis (PCA) disjoint models **[1]**, BPCA model **[2]** and Robust PCA model **[3]** have been developed for the application of gene expression data and provide some promising results. However, these methods either face the problem of outliers' effects, or lack of accuracy for classification of multiple tumor types using a smaller number of variables for future analysis. In this paper, we proposed a novel method of Block Wavelets Shrinkage Principal Component Analysis (BWSPCA) to address the above problems. This method avoids the sensitivity nature towards outlier (i.e. with better robustness property) and reduces dimension with high efficiency and accuracy to identify the most important information from huge amount of data.

In dealing with high dimensional microarray data, Liu and colleagues **[2]** proposed BPCA method to group original data into several 'blocks' of variables. Variables within block are more likely to be related, and likely to have less variability. We performed PCA for each block and select variables from the leading principal components and then combined these selected variables from each block. If the number of these combined variables is still large we will then apply the PCA again and until we obtain a smaller number of variables without losing too much useful information. This method has proven useful and efficient by Liu and colleagues **[2]**. However, BPCA is similar to PCA approach that it is based on the mean and the sample covariance matrix of the data and is very sensitive to outliers. Thus, there is a high possibility that some extreme data points may distort the data structure and hence reducing the accuracy of analysis. For this reason and because of wavelet transform is a tool to able processing both the stationary as well as non-stationary signal and has got multi-resolution capabilities, wavelet analysis has been applied to a large variety of biomedical data. For instance, it has been used effectively in microarray data analysis by Myasnikova and colleagues **[4]**. In particularly, wavelet shrinkage is an efficient technique for denoising and to avoid the problems of outliers.

In many literatures researchers consider the task of diagnosing cancer on the basis of microarray data as *class prediction* and the encompassed methods ranged from modified versions of traditional discriminant analysis, over penalized regression approached, classical nonparametric methods such as the nearest neighbor rules to modern tools of machine learning. See for example Dudoit and Fridlyand (2003) for an overview and references. Since the classification accuracy appears to degrade very rapidly as the number of class increased from gene expression datasets, thus this paper we also proposed an alternated approach based on BWSPCA and artificial Neural Network so call BWNN, which confirms the improvement compared to artificial neural network through a detailed empirical study. Also the BWNN shows very competitive comparing to established classifier including bagboosting, nearest neighbor method and modern tools such as support vector machines and random forest by Marcel Dettling **[6]**. In the next sections, we will introduce the method of BWSPCA and BWNN, and then we apply the proposed method BWSPCA with prediction analysis for microarray in R (PAMR) **[7]** on a NCI60 dataset. At the same time, the approach of BWNN will be applied to the six public cancer datasets to illustrate its efficiency and accuracy in classification with respect to other approaches such as Bagboost, RanFor, SVM, kNN, and ANN. These datasets also are provided by Marcel Dettling **[6]**.

Methodology:

Wavelet shrinkage method

We adapt the techniques *WaveShrink* **[8]** and use wavelets to perform nonparametric regression and consider observed noisy data $Y = (y_1, y_2,\ldots,y_n)$ given by equation (1) (shown in supplementary materials). The estimation of $f = (f_1, f_2,\ldots,f_n)$ with small mean square error is measured by small L_2 risk (minimax property) as equation (2). We post strategies in Figure 1a and the algorithm is shown in supplementary materials.

Block wavelet shrinkage principal component analysis (BWSPCA)

The newly developed method of BWSPCA reduces the dimension and performs denoising by using block (dimension p>> n data size) wavelet shrinkage with robust estimator techniques (deal with 'noising' data) and PCA (deal with vast amount of dimensions). We apply the approach of *screeplot* with a chosen percentage (95% or 98%) of total variation. For the wavelet analysis, a general orthonormal transform *W* (type of discrete wavelet transform (DWT)) is considered to apply on data *Y*. This can often isolated the key features of *Y* into a small number of coefficients. The procedure of BWSPCA is briefly presented in Figure 1b and the algorithm is shown in supplementary materials. This technique can be widely applied in conjunction with other classification methods.

The classification of artificial neural network based on BWSPCA (BWNN)

After we apply the BWSPCA for the denoising and dimension reduction, we use single-hidden-layer neural network, a feed forward neural network *nnet* from *R* library **nnet** for data classification. We also adapt the gene minimization procedure **[9]** as the method of variable selection. That is, each of the input clones is ranked according to its importance to the prediction of ANNs. Increasing number of the top-ranked clones is used to train ANNs, and the resulting classification error is obtained. The minimal number of clones that yield the minimal classification error and the top-ranked clones for each of genes are used to retain the ANNs and predict the test samples without performing a BWSPCA analysis. The work flow diagram for BWNN is presented in Figure 2.

Other classifiers

For the comparison with BWNN, we obtain the mean error rate by applying some of the most popular classification methods for microarray data. These competitors are: 1) BagBoost, using Breiman's bagging as a module in boosting; 2) random forests, following the default implementation in the ***R*** function **randomForest ( )**; 3) support vector machines (SVM), with radial basis kernel using the default cost parameter of c=1 for penalizing observations with negative margins. It is implemented by determined the optimal hyperplan using the function svm from the R library e1071; 4) the k-nearest neighbor (kNN), we set k=1 and this classification method is implemented in the R library class; 5) the artificial neural network (ANN), with PCA for dimension reduction and using an ANN-based gene minimization strategy for classification **[10]**. This method is implemented in the R function nnet ( ). Many of these methods have been used for comparison by Dettling **[6]**.

Discussion:

Implementation of method

We examine BWSPCA based on the following two parts: comparing proposed BWSPCA with other two approaches, namely the classical PCA, BPCA **[2]**, and comparing the classifier BWNN with respect to Bagboot, RanFor, SVM, kNN, and ANN. The first part is to find out which method provides the smallest number of genes to achieve the most accurate classification of cancer types. We used the S-PLUS Wavelets package to perform all Wavelet Shrinkage process (see Figure 1). Others such as semi-partial R^2 plot and selecting variables from each block are done by SAS **[11]**. The final cross-validation part for BWSPCA result is carried out by the package PAMR in R. We repeat similar procedure for PCA and BPCA **[2]** and investigate the performance of all three approaches. The second part we adapt the similar procedure from Dettling **[6]** for other five classification methods. We compare results and investigate the performance of six approaches. The datasets are presented in Table 2 under supplementary materials.

Results for BWSPCA

We present the results for BPCA and BWPCA on NCI60 dataset. There are 11 blocks are obtained and the summary is given in Table1 (under supplementary materials). For BPCA, the number of genes in the 11 blocks ranges from 71 to 168 and yields a total 157 genes. For BWSPCA, the number of genes in the 11 blocks ranges from 3 to 5 and yields a total 39 genes. The range of % variance from BPCA for each block is 95.1% to 96% while the range of % variance from BWPCA for each block is 98.2% to 99.9%. This implies that BWSPCA uses a much smaller number of genes to explain more variances and hence keeping more important information. In fact, for BPCA approach, even we include more principal components to explain above 98% of total variance, the accuracy has not been improved with the increase in the number of genes. It can also be shown that if we repeat the BPCA procedure again to further reduce the dimension we will decrease the accuracy.

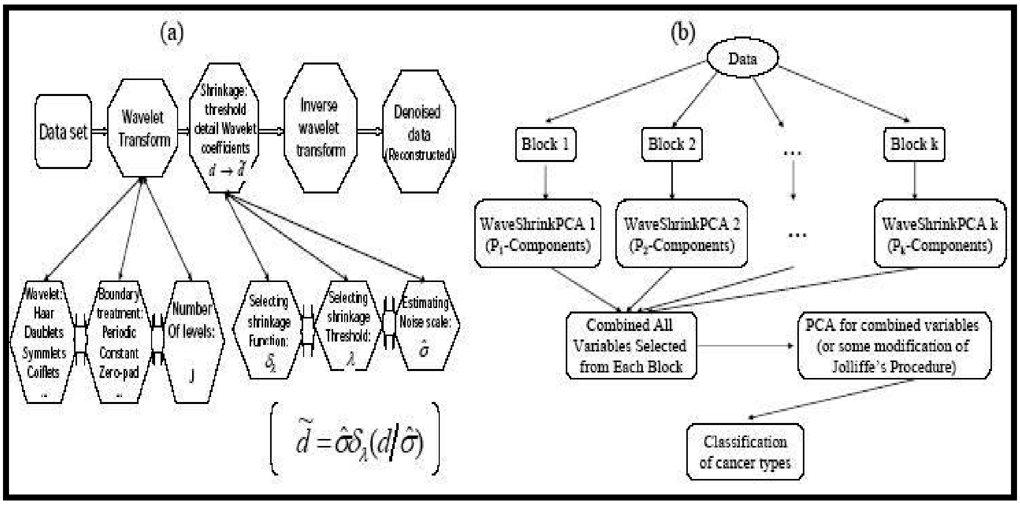
(a)
Data set
Wavelet Transform
Shrinkage: threshold detail Wavelet coefficients
$d \to \tilde{d}$
Inverse wavelet transform
Denoised data (Reconstructed)
Wavelet: Haar Daublets Symmlets Coiflets ...
Boundary treatment: Periodic Constant Zero-pad ...
Number Of levels: J
Selecting shrinkage Function: δ_λ
Selecting shrinkage Threshold: λ
Estimating Noise scale: $\hat{\sigma}$
$\tilde{d} = \hat{\sigma}\delta_\lambda(d/\hat{\sigma})$
(b)
Data
Block 1
Block 2
...
Block k
WaveShrinkPCA 1 (P_1-Components)
WaveShrinkPCA 2 (P_2-Components)
...
WaveShrinkPCA k (P_k-Components)
Combined All Variables Selected from Each Block
PCA for combined variables (or some modification of Jolliffe's Procedure)
Classification of cancer types

Figure 1 (a): Wavelet-shrinkage paradigm **(b)** Schematic diagram of data flow by using BWSPCA method.

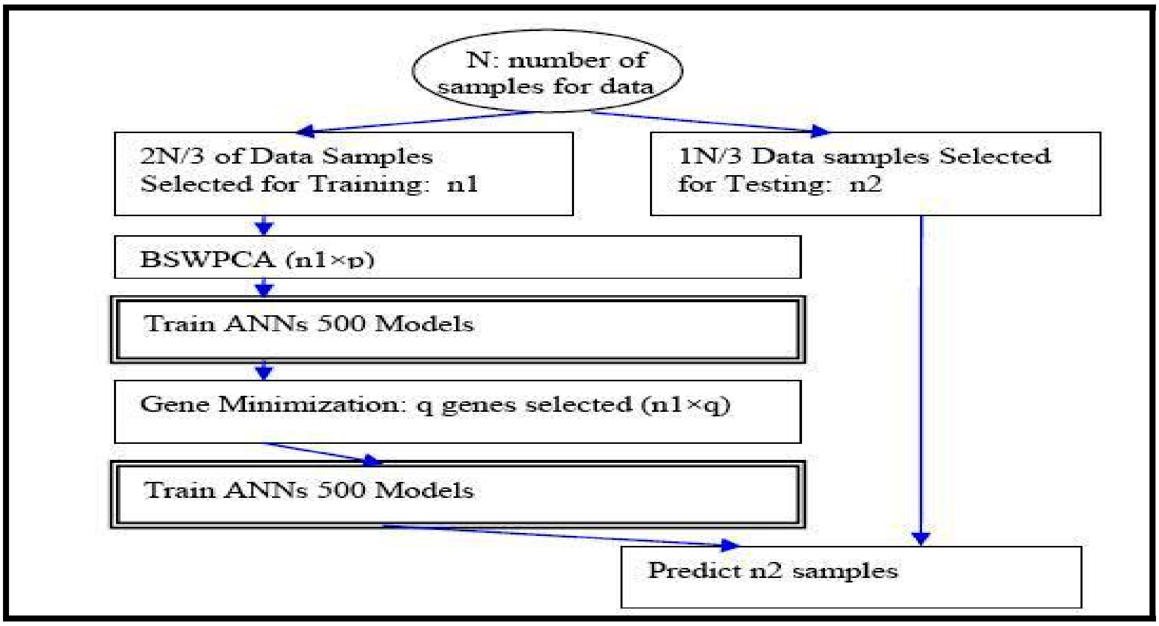
N: number of samples for data
2N/3 of Data Samples Selected for Training: n1
1N/3 Data samples Selected for Testing: n2
BSWPCA (n1×p)
Train ANNs 500 Models
Gene Minimization: q genes selected (n1×q)
Train ANNs 500 Models
Predict n2 samples

Figure 2: Work flow for BWNN analysis for microarray data.

Cross-validation for BWSPCA results

To compare the accuracy of three approaches involving classical PCA, BPCA and BWSPCA, we apply the nearest shrunken centroids (NSC) method **[7]** on NCI60 21-cell lines dataset and successfully identified a set of 9 genes that predict CO, LE and RE, with 100% accuracy. A 6-fold cross-validation is applied and the result is presented in Table 1 under supplementary material. From the approach of BWPCA, we obtain the minimized number of genes without misclassification error is 9 at threshold value 2.41 (Figure 3a). At least 70% chance (cross-validated probabilities) that each class separate clearly for BWSPCA data (at the threshold value 0), i.e., almost perfectly classified each class except a few CO-genes and LE-genes. This indicates we achieve high accuracy for BWSPCA method (Figure 3b).

Discussion:

We present the class prediction performance of BWNN with other five classifiers on six publicly available datasets. Similarly as Marcel Dettling **[6]** did, we reduce the variability by splitting data into the learning and test sets with repeated *n*=50 times and the error estimates are averaged. The results are provided in Table 3 (under supplementary materials). Our approach is one of the best classifier among to these top four competitors: BWNN, BagBoost, SVM and Random Forest. Especially on multi-class cancer type (>=3), we have that the average ranks from BagBoost to BWNN are: 3, 2.33, 3.33, 5, 5.67, and 1.67 respectively. Figure 4 shows a visual illustration for these results. The intervals in this plot provided the variability over 50 random splits and represent ±1.96 standard error from the point estimate (mean).

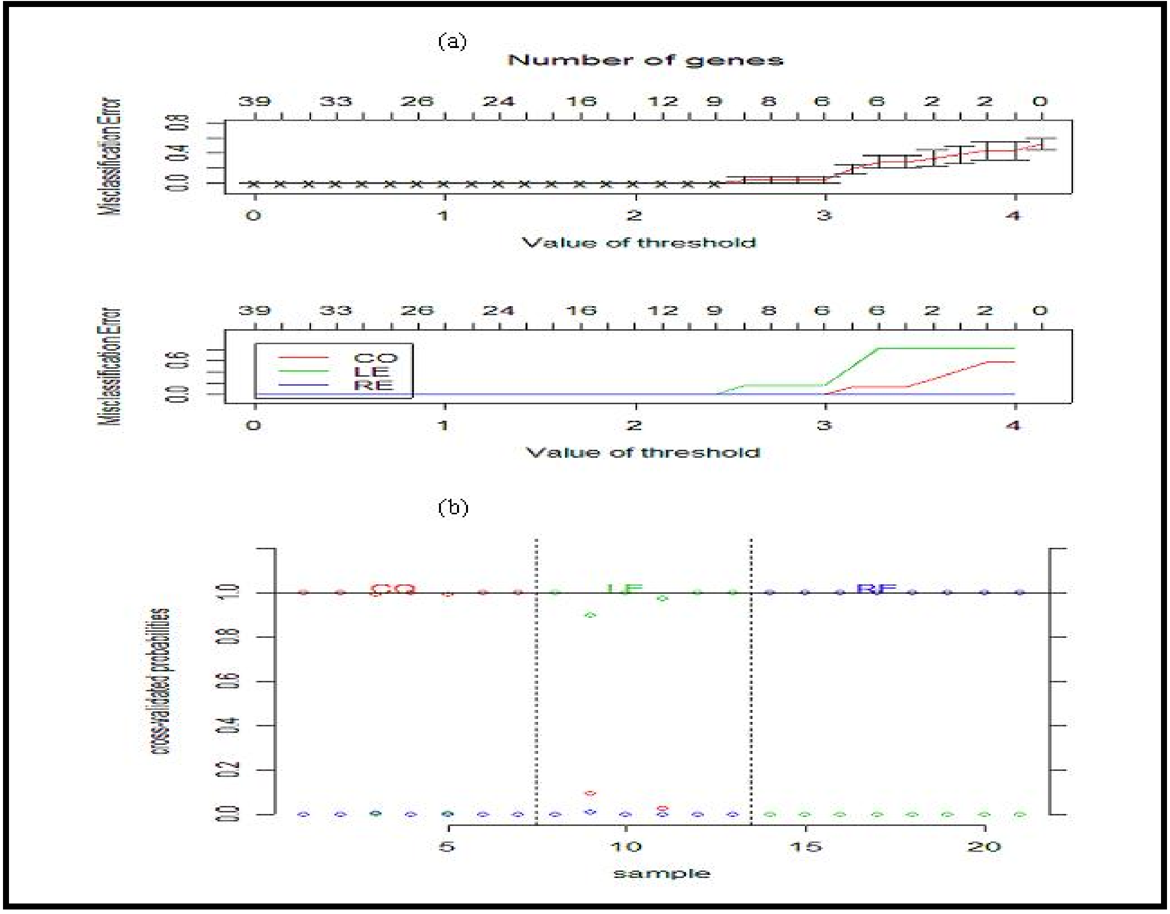

Figure 3: The cross-validated BWSPCA results by applying the method of nearest shrunken centroids in PAMR: **(a)** the cross-validated error curves for a range of threshold values, and **(b)** the plot of cross-validated sample probabilities for the original result of BWSPCA (at threshold value 0).

In summary, we observe that the BWNN is one of the best classifier for tissue classification based on gene expression. For all of these methods, KNN is surprisingly almost good as other classifiers except on the prostate data, which is notably the largest dataset in the analysis. This may be recognized this method is limited to gene expression datasets with small sample size. SVM has a quite good in average performance among all methods. BagBoost and Random forest sounds has the best performance in certain datasets. The method of Random forest performances almost as good as our method on all datasets except the brain data. Finally, the method ANN is comparable perform poor in average among all classifiers. However, overall it is possible to achieve very high accuracy on most of the dataset study. Here even ANN has a >74% on brain data (the worst one). On the Lymphoma, BWNN can achieve almost perfect on test datasets, accuracy is >98%. Moreover, the best performance of classifiers (BWNN) on the Brain dataset is 80.308% which is quite worse with comparable to other datasets. This dataset has noticed with the smallest sample size and largest number of class to be classified. Thus one may conclude that multi-class classification based on gene expression can be effectively solved by increasing the sample size.

We also compare Dettling's result in Table 3 (in Bracket); most of presented misclassification rates are similar to our results. Notice that the random selection of training (testing) samples can cause the results quite different if the repeating time is less. The techniques of variables selection can also improve the accuracy. As Dettling mentioned that the data overlap in the learning sets can cause the error estimates of the 50 splits without independence and therefore can not just interpreted non-overlapping intervals of error bar as a significant advantage. In our study, there is no clear advantage for any one of top four classifiers. Even for Dettling's result, the clear advantage for BagBoosting on the brain tumor data does not happen in our study. We also conjecture that BWNN has a better performance for small sample size datasets (Colon, Lymphoma and Brain). Finally, it is obvious that ANN without BWSPCA technique performing badly and is not suitable for class prediction with gene expression data. For the further ensuring the advantage for our method, probably we need to perform a simulation study in next research.

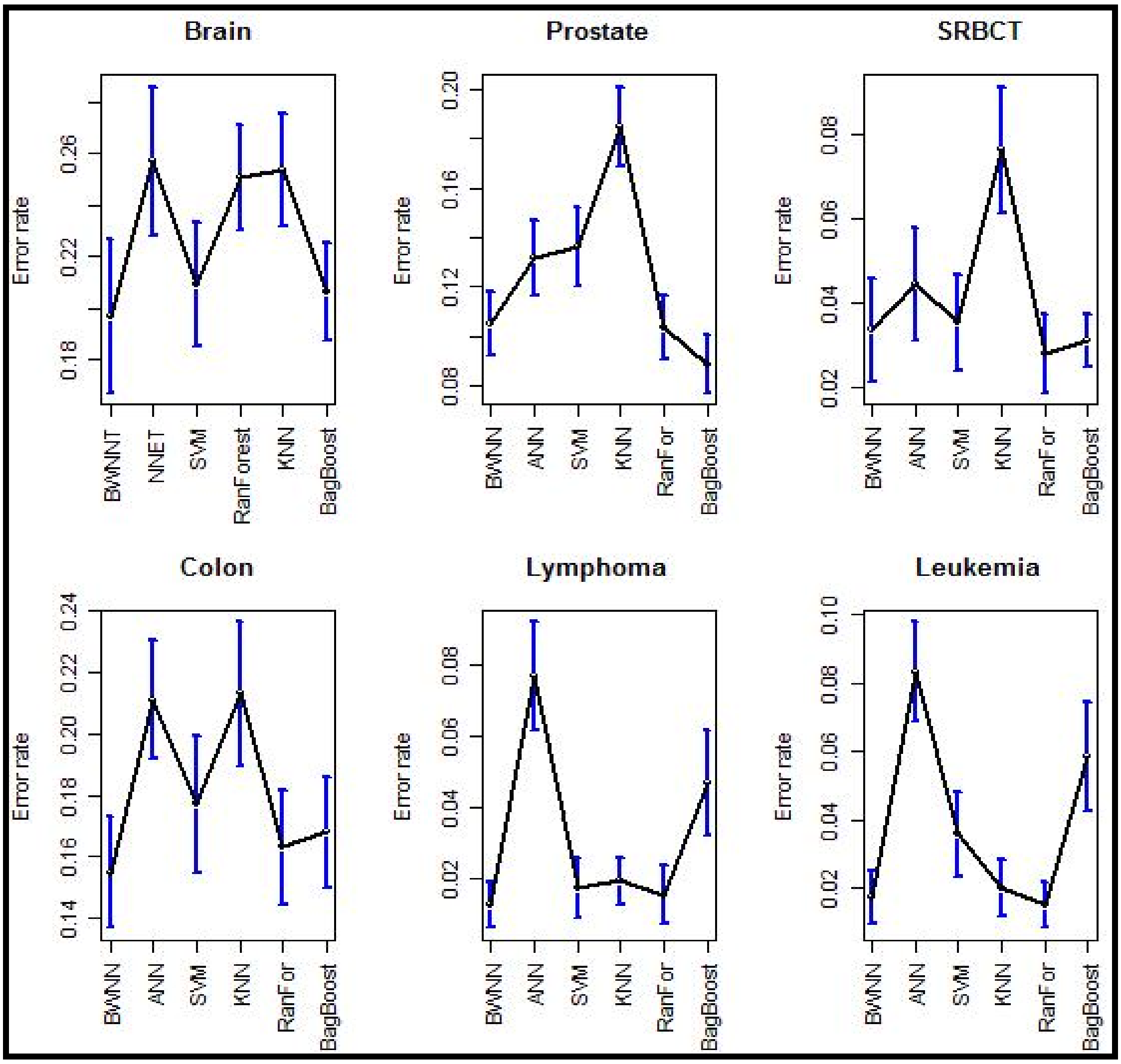

Figure 4: Misclassification rates for 6 classifiers on 6 microarray datasets, based on 50 random splits into training and test sets. The bars represent pseudo-confidence intervals, showing the variability of the point estimates.

Conclusion:

This paper shows that a much smaller number of genes can provide the same insight for the cancer phenotypes as the complete set of genes, which is very useful for huge size of dataset in microarray analysis. We demonstrated that BWSPCA is more effective comparing to classical PCA and BPCA to select the most significant genes. This approach can also be applied in conjunction with other supervised methods such as PAM to identify subsets of genes that concisely classify cancer types. Likewise, we proposed an alternative method BWNN to classify the cancer phenotypes through different type of microarray data. We demonstrated that BWNN is more effective comparing to other commonly used classifiers to select the most significant genes, especially for small sample size. With BWNN, we found a subset from each of six datasets was able to classify the types of cancers with comparable high accuracy in average. Applying the BWNN is a plausible solution to the problem of accuracy degradation as the number of classes increases (e.g. Brain). In addition to the application of gene expression, this method can also be applied to other kinds of genomic array data analysis.

Acknowledgment:

This study is funded by the Research Grant Council of Hong Kong #CUHK4469/03M

References:

[01] S. Bicciato, *et al., Bioinformatics,* 19: 571 (2003) [PMID: 12651714]

[02] Y. Liu, *et al., Statist. Med.,* 21: 3465 (2002)

[03] M. Hubert and S. Engelen, *Bioinformatics,* 20: 1728 (2004) [PMID: 14988110]

[04] E. Myasnikova, *et al., Bioinformatics,* 17: 3 (2001) [PMID: 11222257]

[05] S. Dudoit and J. Fridlyand, *Bioinformatics,* 19: 1090 (2003) [PMID: 12801869]

[06] M. Dettling, *Bioinformatics*, 20: 3583 (2004) [PMID: 15466910]
[07] R. Tibhishirani, *et al., Proc. Natl. Acad. Sci. U.S.A.*, 99: 6567 (2002) [PMID: 12011421]
[08] D. L. Donoho and I. M. Johnstone, *Biometrika*, 81: 425 (1994)
[09] J. Khan, *et al., Nat Med.*, 7: 673 (2001) [PMID: 11385503]
[10] J. S. Wei *et al., Cancer Research,* 64: 6883 (2004) [PMID: 15466177]
[11] R. Khattree and D. N. Naik, *Cary*, N.C.: SAS Institute (2000)
[12] V. G. Cheung, *et al., Nature Genetics Supplement*, 21: 15 (1999) [PMID: 9915495]
[13] D. L. Donoho and I. M. Johnstone, *Journal of American Statistical Association*, 90: 1200 (1995)
[14] J. R. King and D. A. Jackson, *Environmetrics*, 10: 67 (1999)
[15] J. Y. Koo, *et al., Bioinformatics*, 22: 950 (2006) [PMID: 16452113]
[16] T. Li, *et al., Bioinformatics*, 20: 2429 (2004)
[17] S. G. Mallat *et al., IEEE Trans. Pattern Analysis and Machine Intelligence*, 11: 674 (1989)
[18] S. Ramaswamy and T. R. Golub, *Journal of Clinical Oncology*, 20: 1932 (2002)
[19] U. Scherf *et al., Nature Genetics Supplement*, 24: 236 (2000)

Edited by P. Kangueane
Citation: Lee & Zee, Bioinformation 3(5): 223-229 (2008)

Supplementary material

Table1: Comparison of four classification methods for NCI60 21-cell lines dataset

Method	6-fold CV overall error rate	No. of genes
PCA(with NSC)	1/21	33
BPCA(with NSC)	0/21	24
BWSPCA(with NSC)	0/21	9

Table2: Six public datasets

Dataset	References	n	p	#of Class	Response
Leukemia	Golub et al. (1999)	72	3571	2	Subtypes of leukemia
Colon	Alon et al. (1999)	62	2000	2	Tumor/normal tissue
Prostate	Singh et al. (2002)	102	6033	2	Tumor/normal tissue
Lymphoma	Alizadeh et al. (2000)	62	4026	3	Subtypes of lymphoma
SRBCT	Khan et al. (2002)	63	2308	4	Subtypes of lung
Brain A	Pomeroy et al. (2002)	42	5597	5	Different tumor types

Table3: Misclassification rates and average ranking for 6 classifiers on 6 different datasets

Classifiers	Leukemia (%)	Colon (%)	Prostate (%)	Lymphoma (%)	Srbct (%)	Brain (%)	Rank
BagBoost	5.83(4.08)	16.80(16.10)	8.88(7.53)	4.67(1,62)	3.11(1.24)	20.62(23.86)	3.00
RanFor	1.50(2.50)	16.30(15.43)	10.35(7.88)	1.52(1.43)	2.82(2.29)	24.08(34.71)	2.00
SVM	3.58(3.50)	17.70(16.67)	13.65(6.82)	1.71(.95)	3.55(1.81)	20.92(28.14)	3.83
kNN	2.00(3.83)	21.30(16.38)	18.47(10.59)	1.91(1.52)	7.64(1.43)	25.39(29.71)	5.00
ANN	8.33	21.10	13.18	7.71	4.46	25.69	5.33
BWNN	1.75	15.50	10.53	1.24	3.36	19.69	1.83

Algorithm 1 **Denoising of the data Y by Wavelet shrinkage**

Step 1 Perform a suitable wavelet transform of the noisy data.

Step 2 Compute the detailed Wavelet coefficients $\tilde{d}$ depending upon the noise variance (or standard deviation estimate $\hat{\sigma}$) and a suitable shrinkage function δ_λ where λ is the selecting shrinkage threshold.

Step 3 Obtained the coefficients from above after thresholding and is inverted to reconstruct the denoised data.

Algorithm 2 **BWSPCA methodology**

Step 1 Divide the original data Y into *k* blocks (from partitioned variables), say, Y_i for *i=1,…,k.* We determine the number of blocks and select variables from each block by using plots of semi-partial R^2

Step 2 For each block Y_i , we perform wavelet shrinkage method.

Step 3 When we obtain shrunk data in step 2 we apply the PCA on each block.

Step 4 Retain those variables with large coefficients after examine the coefficients of variables *X* in above *p* leading component (see Jackson 1991 and Jolliffe 1986).

Step 5 Collect variables obtained from step 4 and then select a suitable clustering method.

Algorithm 2 BWSPCA classifier

Step 1 For decomposition process, the wavelet transform we chose the biorthogonal wavelet with B-spline at the degree 2 of polynomial. The length of the support for this dual wavelet is 7. The boundary treatment we used zero-pad or periodic (depends on sample size for each block).

Step 2 For Shrinkage process, we select shrinkage function, threshold and noise scale as ‘soft’, ‘adapt’, ‘sure’ and ‘all’ (with MAD estimate function for robustness) correspondingly.

Step 3 For each block, we extract shrunk data and use PCA to obtain *screeplot*. According to the *screeplot*, we selected the leading principal components explaining > 90% of total variance.

Step 4 Use modified (if necessary) Jolliffe’s one variable per leading component method to select variables from each retained principal component.

Step 5 Combine all variables selected from step (4) as a new data set. If there are still too many variables in this new data, we repeat step 3 and 4 again. If no, we continued to the next step.

Step 6 Apply the Artificial Neural Network and obtained the classification of cancer types under the BWSPCA method.

Equations:

Applied nonparametric regression and consider observed noisy data $Y = (y_1, y_2, \ldots, y_n)'$ given by

$$\boxed{y_i = f_i + \varepsilon_i} \qquad \rightarrow \qquad (1)$$

In order to estimate $f = (f_1, \ldots, f_n)'$ with small mean square error, we need to find an estimator $\hat{f}$ with small L_2 risk (minimax property) as

$$\boxed{R\left(\hat{f}, f\right) := \frac{1}{n}\sum_{i=1}^{n} E\left(\hat{f}_i, f_i\right)^2.} \qquad \rightarrow \qquad (2)$$