# Classification of High-throughput Data Using Correlation-shared Gene Clusters

**Pingzhao Hu, Hui Jiang**
Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, Canada

**Abstract** - *Molecular predictor is a new tool for disease diagnosis, which uses gene expression to classify the diagnostic category of a patient. The statistical challenge for constructing such a predictor is that there are thousands of genes to predict disease category, but only a small number of samples are available. We explored a correlation-sharing based method to integrate 'essential' correlation structure among genes into the predictor in order that the cluster structure of genes, which is related to diagnostic classes we look for, can have potential biological interpretation. We evaluated performance of the method with other methods using three real examples. Our results show that the approach has the advantage of computational simplicity and efficiency with lower classification error rates than the compared methods.*

**Keywords:** Correlation-sharing, Principal components, Classification, High-throughput

## 1 Introduction

With the development of microarrays technology, more and more statistical methods have been applied to the disease classification using microarray gene expression data. Microarray data sets often have a large number of features (genes), but only a very limited number of samples are available, which presents unique challenges to feature selection and predictive modeling. In general, these statistical methods can be divided into two categories: one is the supervised classification methods. For example, Golub et al. developed a "weighted voting method" to classify two types of human acute leukemias [1]. Radmacher et al. constructed a 'compound covariate prediction' to predict the BRCA1 and BRCA2 mutation status of breast cancer [2]. Studies have shown that given the same set of selected features, different classification methods often perform quite similarly and simple methods like diagonal linear discriminant analysis (DLDA) and k nearest neighbor (kNN) normally work remarkably well [3]. Thus, finding the most informative features is a crucial task in predictive modeling from microarray data [4-5]. Another is the unsupervised clustering approaches, which are usually used to determine gene clusters that are mostly correlated with clinical outcomes [6]. However, the clustering approach is purely exploratory and methods that can be used to assess the significance of the clustering results are required. It has been widely known that most diseases (such as cancer) are 'caused' or influenced by multiple gene variations more often than only a single gene. Traditional microarray-based disease classification approaches use only individual differentially expressed genes as biomarkers to discriminate classes of cancer and normal samples. However, a large proportion of such genes are irrelevant and functional correlations among those genes are ignored. Since the genes with the best discriminative power are likely to correspond to a limited set of biological functions or pathways, it is rational to focus on these key functional expression patterns for disease prediction. This approach may then provide clues as for the types of biological processes that underlie the expression patterns of sets of genes.

Some attempts have been made to integrate the unsupervised gene clustering and the supervised disease classification approaches into a unified classification process. Li et al. developed cluster-Rasch models, in which a model-based clustering approach was first used to cluster genes and then the discretized gene expression values were input into a Rasch model to estimate a latent factor associated with disease classes for each gene cluster [7]. The estimated latent factors were finally used in a regression analysis for disease classification. They demonstrated that their results were comparable to those previously obtained, but the discretization of continuous gene expression levels usually results in a loss of information. Hastie et al. proposed a tree harvest procedure for finding additive and interaction structure among gene clusters, in their relation to an outcome measure [8]. They found that the advantage of the method could not be demonstrated due to the lack of rich samples. Dettling et al. presented a new algorithm to search for gene clusters in a supervised way. The average expression profile of each

cluster was considered as a predictor for traditional supervised classification methods. However, using simple averages will discard information about the relative prediction strength of different genes in the same gene cluster [9]. Yu also compared different approaches to form gene clusters. The resulting information was used for providing sets of genes as predictors in regression [10].

Recently, gene co-expression networks have become a more and more active research area [11-14]. A gene co-expression network is essentially a graph where nodes in the graph correspond to genes, and edges between genes represent their co-expression relationship. The gene neighbor relations (such as topology) in the networks are usually neglected in traditional cluster analysis [13]. One of the major applications of gene co-expression network has been centered in identifying functional modules in an unsupervised way [11-12], which may be hard to distinguish members of different sample classes. Recent studies have shown that prognostic signatures that could be used to classify the gene expression profiles from individual patients can be identified from network modules in a supervised way [14].

In this paper we explored a clustering-based approach for classification of high-throughput gene expression data. Specifically, we first used a seed based approach to identify correlation-shared gene clusters from gene network. Each of these clusters included a differentially expressed gene between sample classes, which was treated as a seed, and a set of other genes highly co-expressed with the seed gene; then we performed principal component analysis (PCA) to extract meta-gene expression profiles; finally a supervised PCA-based logistic regression (LR) model was built to predict disease outcomes. We call the method as CPCLR. The method returned signature components of tight co-expression with good predictive performance. The performance of this method was compared with other state-of-the-art classification methods. We demonstrated that the approach has the advantage of computational simplicity and efficiency with lower classification error rates than the compared classification methods.

The remainder of this paper is organized as follows: Section 2 gives a detailed description of our classification method and briefly discusses other methods to be compared as well as the evaluation strategy; Section 3 presents the results based on six classification methods and three case studies; Section 4 summarizes our findings in the study.

# 2 Methods

## 2.1 CPCLR algorithm

CPCLR classification algorithm includes three stages: 1) construct correlation-sharing based gene clusters; 2) extract meta-gene expression profiles from the constructed clusters using PCA; 3) classify samples using PCA-based LR model. Here we briefly described each of the three stages:

*Stage 1: construct correlation-sharing based gene clusters.* We modified the correlation-sharing method developed by Tibshirani and Wasserman [15], which was originally proposed to detect differential gene expression. The approach works in the following steps:

*A:* Compute test statistic $T_i (i = 1,2,..., p)$ for each gene *i* using the standard t-statistic or a modified t-statistic, such as significance of microarrays (SAM) [16].

*B:* Select seed genes having larger absolute test statistic values, say top *m* genes.

*C:* Find the cluster membership *s* for each selected seed gene $i^*$. The cluster assignments can be characterized by a many to one mapping. That is, one seeks a particular encoder $C_r(i^*)$ that maximizes

$$i_s^* = \max_{\{0 \le r \le 1\}} ave_{i \in C_r(i^*)} | T_i | \quad (1)$$

where $C_r(i^*) = \{s : abs(corr(x_{i^*}, x_s)) \ge r\}$. The set of genes *s* for each seed gene $i^*$ is an adaptively chosen cluster, which maximizes the average (*ave*) differential expression signal around gene $i^*$. The set of identified genes *s* should have absolute (*abs*) correlation (*corr*) with $i^*$ larger than *r*. The advantage of the correlation-sharing based clustering method is that the membership in different clusters can be overlapped rather than mutually disjoint.

*Stage 2: Principal component analysis of correlation-shared expression profiles:* To do this, for each of the seed-based gene cluster, we performed principal component analysis. Specifically, for a given gene cluster with *C* genes, assume $x^{(j)} = (x_{1j}, x_{2j},..., x_{Cj})^t$ be expression indices of *C* genes in the *j*-th sample and *t* denotes transpose of a vector. Let $\sum$ be covariance matrix of *x* with dimension *C x C*. All positive eigenvalue of $\sum$ are denoted by $\lambda_1 > \lambda_2 > ... > \lambda_C$. The first PC score of the *j*-th sample is given by $x_j^* = e_1^t x^{(j)}$, where $e_1$ is the eigenvector associated with $\lambda_1$. Therefore, we can define the super-gene expression profile for *N* samples in a seed–based gene cluster as $x^* = \{x_1^*, x_2^*,..., x_N^*)^t$. The estimated values for the coefficient $e_1^t$ (eigenvector) of the first PC can be computed using singular value decomposition (SVD) [17]. Briefly, assume *X* be an *NxC* matrix with normalized gene expression values of *C* genes in a given cluster, so we can express the SVD of *X* as $X = ULA^T$,

where $U = \{u_1, u_2, ..., u_d\}$ is a $N$ x $d$ matrix ($d = rank(X)$), $L = diag\{l_1^{1/2}, l_2^{1/2}, ..., l_d^{1/2}\}$ is a $d$ x $d$ diagonal matrix where $l_k$ is $k$-th eigenvalue of $X^t X$, $A = \{e_1, e_2, ..., e_d\}$ is a $C$ x $d$ matrix where $e_k$ is eigenvector of associated with $\lambda_k$ and coefficients for defining PC scores. Magnitude of loadings for the first principal component score can be viewed as an estimate of the amount of contribution from the clustered genes.

*Stage 3: Classification using PCA-based logistic regression model:* Assume $Y$ is a categorical variable indicating the disease status (such as cancer or no cancer). Here we only focus on binary classification and suppose that $Y=1$ denotes the presence and $Y=0$ indicates the absence of the disease. Therefore, we can have following supervised PCA-based logistic regression model:

$$\log(\frac{p_j}{1-p_j}) = \beta_0 + \sum_{i^*}^{m} \beta_{i^*} PC1_{i^* j} + \varepsilon_j \quad (2)$$

where $p_j = \Pr(Y_j = 1 \mid PC1_{i^* j}, i^* = 1,2,...,m)$. $PC1_{i^* j}$ is the first principal component score estimated from the seed gene cluster $i^*$ for sample $j$ and represents the latent variable for the underlying biological process associated with this group of genes. The model was fitted using *GLM* function in stats R package.

## 2.2 Method Comparisons

We compared the prediction performance of CPCLR with other established classification methods, which include, diagonal linear discriminant analysis (DLDA), logistic regression (LR) model, one nearest neighbor method (1NN), support vector machines (SVM) with linear kernel and recursive partitioning and regression trees (Trees). We used the implementation of these methods in different R packages (http://cran.r-project.org/), which are sma for DLDA, stats for LR, class for 1NN, e1071 for SVM and rpart for Trees. Default parameters were used. In the comparison, we selected seed genes using t-test and SAM and evaluated the performance of DLDA, LR, 1NN, SVM and Trees using different number of top seed genes and that of CPCLR using the gene clusters built on the selected seed genes.

## 2.3 Cross-validation

We performed ten-fold cross-validation to evaluate the performance of these classification methods. The basic principle is that we split all samples in a study into 10 subsets of (approximately) equal size, set aside one of the subsets from training and carried out seed gene selection, gene cluster construction, extracted super-gene expression profiles and classifier fitting using the remaining 9 subsets. We then predicted the class label of the samples in the omitted subset based on the constructed classification rule. We repeated this process 10 times so that each sample is predicted exactly once. We determined the classification error rate as the proportion of the number of incorrectly predicted samples to the total number of samples in a given study. This 10-fold cross-validation procedure was repeated 10 times and the averaged error rate was reported.

# 3 Experimental Results

## 3.1 Real datasets

We applied the CPCLR algorithm and the established classification methods mentioned in Section 2.2 to three microarray data sets. The detailed description of these data sets is shown in Table 1. We got the preprocessed Colon cancer microarray expression data from http://genomics-pubs.princeton.edu/oncology/. For prostate cancer and lung cancer microarray data, we downloaded the raw data from gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/) and preprocessed them using robust multi-array average (RMA) algorithm [18].

Table 1: Descriptive characteristics of data sets used for classification

| Disease | Groups | No. Samples | No. Genes | Studies |
|---|---|---|---|---|
| Colon Cancer | Tumor/ Normal | 40 / 22 | 2000 | [6] |
| Prostate Cancer | Tumor/ Normal | 50 / 38 | 12635 | [19] |
| Lung Cancer | Tumor/ Normal | 60 / 69 | 22215 | [20] |

Tables 2, 3 and 4 listed the prediction performance of different classification methods applied to colon cancer, prostate cancer and lung cancer microarray gene expression data using different number of top seed genes. As we can see, for the colon and lung cancer data sets, CPCLR algorithm has better or comparable classification performance than other well-established classification methods based on different number of top seed genes or significantly differentially expressed genes (Tables 2, 4 and 5). However, for the prostate cancer data, the best performance was observed by using SVM predictors (Table 3). In order to save the time to search for genes which were correlated with a given seed gene and maximized their averaged test statistic value (formula 1), we tested 10

cutoffs of correlation *r* from 0.5 to 0.95 with interval 0.05. We observed that the averaged correlation of genes in the constructed gene cluster is usually between 0.65 and 0.85 with the number of genes in the clusters from 2 to 60, suggesting the genes in the constructed gene clusters are highly co-expressed.

Table 2: Error rates (%) of six classification methods applied to colon cancer data set

| No. Genes | DLDA | 1NN | Tree | SVM | LR | CPCLR |
|---|---|---|---|---|---|---|
| 5 | 11.3 | 21.0 | 22.6 | 11.3 | 11.3 | 9.7 |
| 10 | 17.7 | 16.1 | 29.0 | 12.9 | 14.5 | 9.7 |
| 15 | 12.9 | 12.9 | 24.2 | 14.5 | 12.9 | 11.3 |
| 20 | 12.9 | 16.1 | 25.8 | 12.9 | 14.5 | 11.3 |
| 30 | 12.9 | 16.1 | 19.4 | 14.5 | 19.4 | 12.9 |

Table 3: Error rates (%) of six classification methods applied to prostate cancer data set

| No. Genes | DLDA | 1NN | Tree | SVM | LR | CPCLR |
|---|---|---|---|---|---|---|
| 5 | 23.9 | 26.1 | 22.7 | 21.6 | 22.7 | 21.6 |
| 10 | 19.3 | 28.4 | 31.8 | 17.0 | 26.1 | 19.3 |
| 15 | 22.7 | 26.1 | 29.5 | 26.1 | 26.1 | 23.9 |
| 20 | 22.7 | 25.0 | 27.3 | 19.3 | 21.6 | 20.5 |
| 30 | 21.6 | 23.9 | 29.5 | 21.6 | 22.7 | 21.6 |

Table 4: Error rates (%) of six classification methods applied to lung cancer data set

| No. Genes | DLDA | 1NN | Tree | SVM | LR | CPCLR |
|---|---|---|---|---|---|---|
| 5 | 17.0 | 18.6 | 20.1 | 16.2 | 19.3 | 17.0 |
| 10 | 14.7 | 18.6 | 19.3 | 17.0 | 20.1 | 14.7 |
| 15 | 16.2 | 20.1 | 17.8 | 13.2 | 17.8 | 15.5 |
| 20 | 16.2 | 17.0 | 19.3 | 17.8 | 19.3 | 15.5 |
| 30 | 12.5 | 13.2 | 19.3 | 14.7 | 20.1 | 12.5 |

We also used SAM [16] to select top seed genes and evaluated the prediction performance following the same procedure as described above. Similar prediction results were also observed as shown in Table 5 for lung cancer data. Overall, the CPCLR method has lower error rate than other being compared classification methods.

In all cases, we found that the simple method, DLDA, works well. Its performance is comparable with the advanced method, such as SVM. We also observed that the performance of the predictors with more genes is not necessary better than that of the predictors with fewer genes. For example, the best performance was obtained with only 5 genes for CPCLR predictors in colon cancer data set (Table 2), 10 genes for SVM predictors in prostate cancer data set (Table 3). For lung cancer data set, the best performance was observed using 30 genes for DLDA and CPCLR predictors (Table 4).

Table 5: Error rates of six classification methods applied to lung cancer data set (seed genes selected by SAM)

| No. Genes | DLDA | 1NN | Tree | SVM | LR | CPCLR |
|---|---|---|---|---|---|---|
| 5 | 17.0 | 19.3 | 22.5 | 16.2 | 18.6 | 17.8 |
| 10 | 17.0 | 20.9 | 19.3 | 17.8 | 17.8 | 15.5 |
| 15 | 14.7 | 20.1 | 22.5 | 14.6 | 20.1 | 13.2 |
| 20 | 16.2 | 18.6 | 17.8 | 18.6 | 17.0 | 15.5 |
| 30 | 17.8 | 13.2 | 19.3 | 10.1 | 14.7 | 10.1 |

## 4 Discussions and Conclusions

In this study we investigated a correlation-sharing based method for classification of high-throughput gene expression data. The core idea of the method is to identify 'essential' correlation structure among genes and extract representative features from the correlated gene clusters in a supervised classification procedure. The method takes into account the fact that genes act in networks and the gene clusters identified from the networks act as the features in constructing a classifier. The rationale is that we usually expect tightly co-expressed genes to have a meaningful biological explanation. For example, if gene A and gene B has high correlation, it sometimes hints that the two genes belong to the same pathway or are co-expressed. Instead of using individual genes as predictors in our classification models, we constructed meta-gene expression profiles representing information from each co-expressed gene cluster as predictors to classify disease outcomes. The advantage of this method over other methods has been demonstrated by three real data sets. Our results show that this algorithm is working well for improving class prediction.

## 5 References


[1] 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, E.S. Lander. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring"; *Science*, vol. 286, pp. 531-536, 1999.

[2] M.D.Radmacher, L.M. McShane, R. Simon. "A paradigm for class prediction using gene expression profiles"; *J Comput Biol*, vol. 9, pp. 505-512, 2002.

[3] S. Dudoit, J. Fridlyand, T.P. Speed. "Comparison of discrimination methods for the classification of tumors

using gene expression data"; *Journal of the American Statistical Association,* vol. 97, pp. 77-87, 2002.

[4] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys."Robust biomarker identi¯cation for cancer diagnosis with ensemble feature selection methods"; *Bioinformatics*, vol. 26, pp. 392-398, 2010.

[5] T. Elizabeth, O. Leonardo, B. Pilar, A. Laura. "Multiclass classification of microarray data samples with a reduced number of genes"; *BMC Bioinformatics*, vol. 12, 59, 2011.

[6] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays"; *Proc Natl Acad Sci U S A*, vol. 96, pp. 6745–6750, 1999.

[7] H. Li, F. Hong. "Cluster-Rasch models for microarray gene expression data"; *Genome Biol.*, vol. 2, pp. 0031.1 -0031.13, 2001.

[8] T. Hastie, R. Tibshirani, D. Botstein, P. Brown. "Supervised harvesting of expression trees"; *Genome Biol.*, vol. 2, pp. 0003.1 -0003.12, 2001.

[9] D. dettling, P. Bühlmann. "Supervised Clustering of Genes"; *Genome Biol.*, vol. 3, pp. 0069.1-0069.15.

[10] X. Yu. "Regression methods for microarray data"; Ph.D. thesis, Stanford University, 2005.

[11] L. Elo, H. Jarvenpaa, M. Oresic, R. Lahesmaa, T. Aittokallio. "Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process"; *Bioinformatics,* vol. 23, pp. 2096-103, 2007.

[12] A. Presson, E. Sobel, J. Papp, C. Suarez, T. Whistler, M. Rajeevan, S. Vernon, S. Horvath. "Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome"; *BMC Syst Biol*. vol. 2, 95, 2008.

[13] S. Horvath, J. Dong. "Geometric interpretation of gene coexpression network analysis"; *PLoS Comput Biol.* vol.4, e1000117, 2008.

[14] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, J. L. Wrana. "Dynamic modularity in protein interaction networks predicts breast cancer outcome"; *Nat Biotechnol.* vol. 27, 199-204, 2009.

[15] R. Tibshirani, L. Wasserman. "Correlation-sharing for detection of differential gene expression"; *arXiv*, math. ST, math/0608061.

[16] V. Tusher, R. Tibshirani, G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response"; *Proc Natl Acad Sci USA*, vol. 98, pp. 5116-5121, 2001.

[17] I.T. Jolliffe. *Principal component analysis*. Springer, New York, 2002.

[18] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T. P. Speed. "Summaries of Affymetrix GeneChip probe level data"; *Nucleic Acids Research*, vol. 31, pp. E15, 2003.

[19] R. O. Stuart, W. Wachsman, C.C. Berry, J. Wang-Rodriguez , L. Wasserman, I. Klacansky, D. Masys, K. Arden, S. Goodison, M. McClelland, Y. Wang, A. Sawyers, I. Kalcheva, D. Tarin, D. Mercola. "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer"; *Proc Natl Acad Sci USA*, vol. 101, pp. 615-620, 2004.

[20] A. Spira, J.E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M.E. Lenburg, J.S. Brody. "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer"; *Nat Med.,* vol.13, pp. 361-366, 2007.