Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

Effect of Feature Extraction and Feature Selection on Expression Data from Epithelial Ovarian Cancer

Y. Turkeli[1], A. Ercil[1], O. U. Sezerman[1]
[1]Laboratory of Computational Biology, Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

Abstract— **Classifying the gene expression levels of normal and cancerous cells and identifying the genes most contributing to this distinction propose an alternative means of diagnosis. We have investigated the effect of feature extraction and feature selection on clustering of the expression data on two different data sets for ovarian cancer. One data set consisted of 2176 transcripts from 30 samples, nine from normal ovarian epithelial cells and 21 from cancerous ones. The other data set had 7129 transcripts coming from 27 tumor and four normal ovarian tissues. Hierarchical clustering algorithms employing complete-link, average-link and Ward's method were implemented for comparative evaluation. Principal component analysis was applied for feature extraction and resulted in 100% segregation. Feature selection was performed to identify the most distinguishing genes using CART® software. Selected features were able to cluster the data with 100% success. The results suggest that adoption of feature extraction and selection enhances the quality of clustering of gene expression data for ovarian cancer. Identification of distinguishing genes is a more complex problem that requires incorporating pathway knowledge with statistical and machine learning methods.**

Keywords— **feature extraction, feature selection, ovarian epithelial cells, principal component analysis, tree structured classifiers**

I. Introduction

Advances in DNA microarrays provide the basis for experiments that quantify the relative gene expression levels under different circumstances. This technology has been adopted in investigation of cancer since the gene expression profiles of the cancerous cells can be compared against the healthy cells, which constitute a natural control group [1]. Experiments have been performed on several types of cancer such as leukemia, breast and ovarian cancer, yielding large gene expression data [2]. Analysis of this proliferating data requires application of computational and statistical techniques for deriving inferences and gaining insights into fundamental cancer biology at the molecular level [3].

Feature extraction is a dimension reduction technique in which a transformation is applied to the vector of all input data followed by the selection of the best subset of transformed features. Principal component analysis (PCA) is employed as a feature extraction method. PCA reduces dimensionality by trying to achieve the optimal sum-of-squared reconstruction error [4, 5]. It has previously been applied to gene-expression analysis [6, 7].

Identification of genes that are most effective in successful classification corresponds to feature selection addressed commonly in pattern recognition [4]. Not only does feature selection ameliorate the quality of classification by eliminating the redundancy in data, it also reveals the underlying molecular mechanisms and nominates the discriminating genes as candidate markers for diagnosis, suggesting which genes might be examined further. Feature selection is performed using the CART software [8, 9], which directly reports some of the variables as being important in derivation of the classification tree.

We have performed agglomerative hierarchical clustering of expression profiles from normal and malignant epithelial ovarian cells [10, 11]. Ovarian cancer is ranked fifth in the number of reported deaths from cancer for women in the United States. Even though the five-year rate of survival for early detection is around 90%, this rate declines to 25% if the diagnosis is made when the disease has already disseminated. More severely, ovarian cancer is mostly diagnosed at a distant stage due to lack of symptoms and molecular determinants [12]. Therefore, characterization of the key molecules involved in ovarian carcinogenesis is crucial in order to decrease the mortality rate [13]. Although the mutations in the BRCA 1 and 2 genes and alterations of *p53*, *c-erb-B2* and *c-myc* are considered to be implicated in ovarian cancer, these changes have not been diagnostic [10, 11].

Agglomerative hierarchical clustering algorithms are applied to both the complete data and the data reduced by feature extraction and feature selection. Feature extraction allowed 100% discrimination between healthy and cancerous cells while feature selection pointed out several genes as candidate diagnostic markers.

II. Methodology

A. Data Sets

One of the data sets consisted of 30 samples with the expression levels of 2176 transcripts [9]. 21 of the samples were epithelial cells with Cedar Sinai Ovarian Cancer (CSOC) and nine were normal cells, Human Ovarian surface Epithelia (HOSE). These transcripts are selected by [10] as being differentially expressed with a statistical significance of p-value<0.05. They were already normalized to a target intensity of 2500 and the values below the background intensity of 300 were increased to 300. No further processing was done on the data set.

0-7803-7789-3/03/$17.00 ©2003 IEEE

The second data set contained 31 samples, 27 of which were driven from tumor tissue and 4 from normal ovarian tissue [11]. All of the 7129 transcripts were used.

B. Hierarchical Clustering

For both of the data sets, the transcripts (genes) are considered as the features and the data points are the tissue samples which are clustered. A perfect clustering is obtained when the data set is divided into two branches at the top most level and all healthy samples are placed in one branch whereas all the tumorous samples are placed in the other. Any healthy sample placed into the tumorous branch is considered as a misclassification and vice versa.

We have implemented agglomerative hierarchical clustering. Initially, each observation was a cluster with a single element. For the next step, the two nearest clusters were found and merged. This decreased the number of clusters by one. The algorithm stopped when a single cluster containing all the observations was obtained. The only part that was different in each type of hierarchical clustering was the computation of distance between the clusters. The merging cost function was found by complete link, average link and Ward's method [4, 14].

The programs for agglomerative hierarchical clustering were coded using the Matlab™ software package (The MathWorks, Inc., Natick, MA).

C. Feature Extraction using PCA

The covariance matrix of the standardized data was calculated and the eigenvalues were found. The eigenvalues are sorted and plotted as shown in Fig. 1a. In order to see how much variance can be reconstructed, we divide the cumulative sum of eigenvalues to sum of all eigenvalues each time after we add a new eigenvalue to the set. The eigenvalues are added in sorted order. Proportions of variance attained after adding the eigenvalues one by one in the sorted order is seen in Fig. 1b.

Singular value decomposition (SVD) was used to identify the principle components (PCs) corresponding to the largest eigenvalues. The data set was projected using the PCs to a new data set. Experiments were carried out by selecting the top n features of the transformed data points for several different values of n. This way, a reduced data set was obtained, with each sample represented by a feature vector of length n. The new reduced data set was fed to the clustering programs and the number of correct and incorrect classifications were counted. This procedure is repeated for both of the data sets.

D. Feature Selection using CART

CART® is a binary tree structured classifier which divides the data set into two descendent subsets, iteratively [8]. Like most of the decision tree building algorithms, the most important part of the problem is determining the splits and when to stop splitting. Splitting is performed by selecting a feature level. For example, a parent node would be split when the expression level of gene x_i is less than a certain value ($x_i < a$). Splits are selected such that the data in the child nodes are purer than the parent node, where purity function increases as the node contains samples mostly from one class. A node is recursively split until the decrease in impurity is below a threshold. The terminal nodes, the leaf nodes, are labeled as the class of the dominating data points in them. For each feature (or variable) a rank is determined by summing the decrease in impurity produced in all the rest of the nodes if splitting was performed on that feature [9].

The complete data sets were fed to CART. The classification performance was observed. Moreover, the variables (features) reported as important were recorded as the distinguishing features. These features are then selected and the resulting reduced data sets were fed to the agglomerative clustering algorithms and the clustering performance was observed.

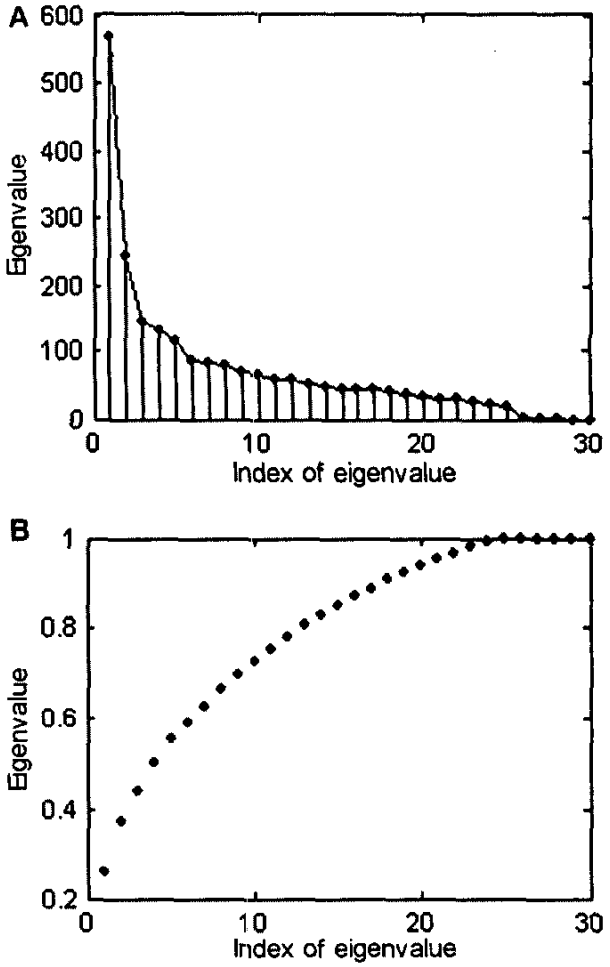

Fig. 1. Plot of (a) eigenvalues (b) proportion of variance attained by the inclusion of i^{th} eigenvalue.

III. Results

A. Clustering the Complete Data

Initially, all of the data was fed to the clustering programs. Two of the agglomerative clustering methods resulted in four misclassifications. The samples C815, C918, C889 and C858 were classified with the normal cells. Ward's method placed only the samples C815 and C918 in the healthy branch.

Hierarchical clustering performed by [10] using Cluster and Treeview [3] also generated four misclassifications. However, their results showed three COSC samples (C918, C889 and C858) in a branch within the HOSE group. In addition, [10] placed a HOSE sample (H263) into the HOSE branch.

Clustering the data set 2 using all 7129 genes as features grouped the four normal samples together. However, they were not placed in a separate branch of their own, rather they were a subgroup under a group of tumor samples.

B. Clustering after PCA

We have applied PCA to both of the data sets prior to clustering. For the first data set, the eigenvalues of the covariance matrix after the 25th were negligible and those after 29th were zero (Fig. 1). Experiments were repeated using the top n PCs selected from the ordered list of PCs for different values of n. Table 1 shows the number of misclassifications for all the different methods for sets of PC's along with the values of the variance accounted for. The results confirm that PCA improves classification performance. For both of the data sets, only two PCs were sufficient to segregate the normal and malignant tissue samples with no misclassifications.

C. Selecting Genes Using CART

Both of the data sets were fed to the CART software [9] which used decision tree structures for classification. CART was able to classify the learning sets with 100% success.

TABLE 1
CLUSTERING AFTER PCA USING DIFFERENT NUMBER OF PCS

Number of PCs	Data Set 1			
	% Variance	Average-Link misc.	Complete-Link misc.	Ward's method misc.
1	26 %	1	1	1
2	37 %	0	0	0
3	44 %	0	0	0
Number of PCs	**Data Set 2**			
	% Variance	Average-Link misc.	Complete-Link misc.	Ward's method misc.
1	23 %	11	10	10
2	34 %	0	0	0
3	41 %	0	0	0

Also, CART ranked several features as important variables for determining the classification. The number of features picked varied depending on the parameters we set. Different numbers of these selected transcripts were fed into our agglomerative clustering programs. With the six genes listed in Table 2, we were able to distinguish all of the CSOC samples from the HOSE samples as shown in Fig. 2.

One of the selected six genes is the proto-oncogene c-myc, whose alteration is known to play a role in development of epithelial ovarian carcinoma [10]. As Fig. 2 presents, CSOC samples are characterized by an increase in the level of *c-myc*. Likewise, AL050051 has elevated levels in CSOCs. The expression level of the rest of the four genes are decreased in CSOC samples.

CART analysis of the data set of [11] also distinguished tumor tissues from the normal tissues successfully. However, a different set of variables were denoted as being most discriminative, with the smallest set presented in Table 2. Among these genes, LISCH7 was also designated as important by [11].

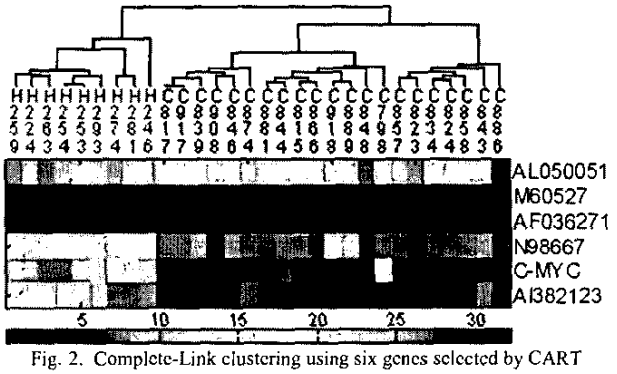

Fig. 2. Complete-Link clustering using six genes selected by CART

TABLE 2
MINIMUM SET OF GENES PICKED BY CART

Probe Set	Data Set 1	
	GeneBank accession	Description
39551_at	N98667	yy66d05.r1 Homo sapiens cDNA
40900_at	AI382123	tc30a09.x1 Homo sapiens cDNA,
37580_at	AF036271	Homo sapiens EEN-B2-L3 mRNA
1936_s_at	Alt. Splice 3, Orf 114	Proto-Oncogene C-Myc
38687_at	AL050051	Homo sapiens mRNA
886_at	M60527	Human deoxycytidine kinase mRNA
Probe Set	**Data Set 2**	
AB000450_at	AB000450	Human mRNA for VRK2
AD000684_cds1_at	AD000684	LISCH7 gene
AFFX-HUMGAPDH/M33197_3_at	M33197	GAPDH mRNA
D00763_at	D00763	Human mRNA for proteasome subunit HC9
AF006084_at	AF006084	Homo sapiens Arp2/3 protein complex subunit p41-Arc (ARC41) mRNA,

IV. Discussion

In this study, feature extraction is performed by applying PCA directly. Eigenvectors of the covariance matrix of a data set identify a linear projection that produces uncorrelated features. It is known that projection of the data on to the eigenvector with the largest eigenvalue does not necessarily preserve the distinction among classes and thus does not always lead to an improved clustering [7]. Yet, perfect clustering observed in the tests we performed using only two PCs suggests that PCA is appropriate for our data set as a feature reduction technique prior to clustering. It also indicates that there is an evident redundancy which produces noise and causes deterioration in the results when all the features are employed.

Since the PCs are some linear combination of the original features, genes in our case, we cannot directly conclude which genes are redundant and which are most effective. As a remedy, we have performed feature selection. CART generated a set of genes that were important in the determination of the tree structure it used for classification. The list of selected genes contained both the genes mentioned as differentially expressed in ovarian cancer in other studies and those not reported previously. Strikingly, the proto-oncogene *c-myc* was one of the distinguishing genes. Among the rest, AI382123, and LISCH7 gene were also detected as preferentially expressed previously [10, 11].

We have investigated why the selected genes did not agree for different data sets. The preprocessed data set of [10] and [11] had 4989 overlapping transcripts. Although both of the data sets were obtained from oligonucleotide arrays and both [10] and [11] used Student's *t*-test, their measurements were not parallel for the transcripts. The genes they have designated as distinguishing were not common as well. Differences might be due to experimental errors, such as *in vitro* processing of the samples. Also, taking the measurements at different stages of the illness or obtaining them from different cell lines may account for the inconsistency in the data we processed.

Samples with randomly selected sets of genes were also fed to the clustering programs. The results generally deteriorated, suggesting that the genes picked by CART are indeed effective. Yet, these genes are mostly distinct from those picked by [10] and [11]. Moreover, selecting several of the genes from literature also produced acceptable clustering results even though no single gene was sufficient by itself. This indicates that discriminating genes are functional as a group and that there may be more than one such group, possibly accounting for different pathways.

V. Conclusion

Comparison of the results of feature extraction and selection to the clustering of the complete data showed that preprocessing indeed resulted in a better segregation.

Genes function in a combined manner and this behavior should also be captured by the algorithm identifying the distinguishing genes. Both PCA and CART consider this combinatory effect. Comparison of data sets indicates that there is redundancy in both of the data sets and PCA is successful for dimension reduction. Finding the distinguishing genes is a more complex process involving different pathways that is highly data dependent and requires larger data sets obtained from consistent experimental conditions.

Acknowledgment

We thank Umut Naci for his comments and discussion. We also thank Matei *et al.* and Welsh *et al.* for providing the data set on their web sites.

References

[1] K. A. Cole, D. B. Krizman, and M. R. Emmert-Buck, "The genetics of cancer- a 3D model," *Nature Genet. Supp.*, vol. 21, pp. 38-41, Jan. 1999.

[2] E. T. Liu, "Classification of cancers by expression profiling," *Curr. Opin. Genet. Dev.*, vol. 13, pp. 97–103, Feb. 2003.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proc. Natl. Acad. Sci. USA.*, vol. 95, pp. 14863–14868, Dec. 1998.

[4] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification.* New York, NY: Wiley-Interscience, 2001.

[5] I. T. Jolliffe, *Principal Component Analysis.* New York, NY: Springer, 1980.

[6] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal component analysis to summarize microarray experiments: application to sporulation time series," in *Proc. Pacific. Symp. Biocompuingt.*, vol. 5, 2000, pp. 452–463.

[7] K. Y. Yeung, and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bionformatics*, vol. 17, no. 9, pp. 763-774, Sep. 2001.

[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Pacific Grove: Wadsworth, 1984.

[9] D. Steinberg, and P. Colla, *CART: Tree-Structured Non-Parametric Data Analysis.* San Diego, CA: Salford Systems, 1995.

[10] D. Matei, T. G. Graeber, R. L. Baldwin, B. Y. Karlan, J. Rao, and D. D. Chang, "Gene expression in epithelial ovarian carcinoma," *Oncogene*, vol. 21, pp. 6289-6298, Sep. 2002.

[11] J. B. Welsh *et al.*, "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," in *Proc. Natl. Acad. Sci. USA,* vol. 98, no. 3, pp. 1176-1181, Jan. 2001.

[12] R. T. Greenlee, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics," *CA Cancer J. Clin.*, vol. 50, no.1, pp. 7-33, Jan-Feb. 2000.

[13] M. Schummer *et al.*, "Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas," *Gene,* vol. 238, no. 2, pp. 375–385, Oct. 1999.

[14] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Am. Stat. Assoc.*, vol. 58, pp. 236–244, 1963.