BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 14 2005, pages 3114–3121
doi:10.1093/bioinformatics/bti483

Gene expression

Ensemble dependence model for classification and prediction of cancer and normal gene expression data

Peng Qiu[1,*], Z. Jane Wang[2] and K. J. Ray Liu[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA and
[2]Department of Electrical and Computer Engineering, University of British Columbia, Canada

Received on October 29, 2004; revised on April 5, 2005; accepted on May 1, 2005
Advance Access publication May 6, 2005

ABSTRACT

Motivation: DNA microarray technologies make it possible to simultaneously monitor thousands of genes' expression levels. A topic of great interest is to study the different expression profiles between microarray samples from cancer patients and normal subjects, by classifying them at gene expression levels. Currently, various clustering methods have been proposed in the literature to classify cancer and normal samples based on microarray data, and they are predominantly data-driven approaches. In this paper, we propose an alternative approach, a model-driven approach, which can reveal the relationship between the global gene expression profile and the subject's health status, and thus is promising in predicting the early development of cancer.

Results: In this work, we propose an ensemble dependence model, aimed at exploring the group dependence relationship of gene clusters. Under the framework of hypothesis-testing, we employ genes' dependence relationship as a feature to model and classify cancer and normal samples. The proposed classification scheme is applied to several real cancer datasets, including cDNA, Affymetrix microarray and proteomic data. It is noted that the proposed method yields very promising performance. We further investigate the eigenvalue pattern of the proposed method, and we discover different patterns between cancer and normal samples. Moreover, the transition between cancer and normal patterns suggests that the eigenvalue pattern of the proposed models may have potential to predict the early stage of cancer development. In addition, we examine the effects of possible model mismatch on the proposed scheme.

Availability: see Supplemental website at http://dsplab.eng.umd.edu/~genomics/edm

Contact: qiupeng@umd.edu

INTRODUCTION

With the rapid development of microarray expression technologies in the past few years, it is possible to monitor the expression levels of thousands of genes simultaneously (Lockhart and Winzeler, 2000; Young, 2000). The large amount of data generated by expression microarrays have stimulated the development of many computational methods to study different biological processes at the gene expression level. Among these, understanding the difference between cancer and normal cells is of particular interest. This includes the difficult task of distinguishing cancerous subtypes, such as benign, invasive, neoplastic and metastatic. Cancer is the fourth most common disease and the second leading cause of death in the United States. Therefore, detection of cancer is a research topic with significant importance. Recently, gene array techniques have been shown to provide insight into cancer study (Chang *et al*., 2003; Van't Veer *et al*., 2002), and molecular profiling, based on gene expression array technology, is expected to offer the promise of precise cancer detection and classification. We plan to address this challenge in this paper.

Current methods for the classification of microarray gene expression data can usually be divided into two categories. One is based on the clustering of samples, which can be used to distinguish cancer and normal samples and to distinguish subtypes of cancers. Some example schemes include hierarchical clustering (Eisen *et al*., 1998), local maximum clustering (Wu *et al*., 2004), self-organizing map (SOM) (Kohonen, 1997) and K-means clustering and its variations (Tavazoie *et al*., 1999). These clustering methods are mainly data-driven approaches. Usually, they do not require many prior assumptions, i.e. an underlying model. However, determining the number of clusters is a challenging problem in itself, and there is a lack of widely accepted measures to evaluate the clustering performance.

The other category is based mainly on a machine-learning approach. Motivated by the success of machine-learning algorithms in image and speech processing, many researchers have applied them to microarray data analysis, for example, K-nearest neighbors (KNN) (Duda *et al*., 2001), support vector machine (SVM) (Furey *et al*., 2000) and neural network analysis (O'Neill and Song, 2003). Machine-learning methods generally yield better results than the traditional clustering methods. However, in these machine-learning methods, the features used for classification are preselected genes identified by statistical tests on training datasets. Although selected genes form a feature vector and are processed jointly, they are still treated in quite a separate fashion. Genes' group behaviors and interactions are not considered. In this work, we propose to take genes' group behaviors and interactions into account by developing an ensemble dependence model (EDM).

In this paper, we propose an EDM-based classification approach. Because of the limited size of current data, it is not feasible to examine the regulation relationship between all genes. Also, the microarray gene expression data is noisy. However, if genes are clustered properly, the noise level in the resulting cluster expression will be reduced, and we will be able to reveal the ensemble dependence dynamics of gene clusters. This paper is organized as follows: we start by introducing the EDM. In the 'Model-based classification'

*To whom all correspondence should be addressed.

section, the major components of the proposed classification method are discussed, including feature selection, clustering of genes and hypothesis-testing. The proposed scheme is then applied to several publicly available datasets, with results reported in the 'Results' section. Finally, in the 'Model-based prediction and performance analysis' section, we explain why the proposed method works well. We show the two different patterns in the eigen domain between cancer and normal cases, and suggest that the eigen pattern can be used for predicting the transition from the healthy state to the disease state. In addition, we discuss the effects of model mismatch on the proposed scheme.

ENSEMBLE DEPENDENCE MODEL

Because of the limited size of current data, it is not feasible to examine the regulation relationship between all genes. In the proposed EDM, genes are clustered into several clusters. We predict, given appropriate and well-sorted clustering results, that genes' group behavior and ensemble dynamics can be revealed. In what follows, several clustering methods are compared, and we will discuss the appropriate way to cluster genes. In this section, we assume we can cluster genes appropriately and focus on the proposed EDM.

After clustering, each cluster contains specific genes that have a well-defined mathematical relationship to one another. To average out experiment noise and enhance genes' common expression within each cluster, the average gene expression profile is used to represent each cluster. Without any prior knowledge, we assume that each cluster is, to some extent, dependent on all the other clusters. A linear dependence relationship is studied here, as shown in Figure 1, where each arrow represents an inter-cluster dependence relationship. There is a weight a_{ij} associated with each arrow, which indicates to what extent cluster i depends on cluster j. The so-called self-regulation is assumed to be zero, i.e. $a_{ii} = 0$, $i = 1, 2, 3, 4$. Because the cluster average is used to represent each cluster, the intra-cluster dependence relationship within each cluster is averaged out. Later, it is clear that, from a mathematic point of view, allowing non-zero a_{ii} terms will make the model-learning process trivial and unreasonable, since the results will simply be $a_{ii} = 1$ for any i, and $a_{ij} = 0$ for any $i \neq j$.

The dependence relationship shown in Figure 1 can be expressed as the following linear equation:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ a_{21} & 0 & a_{23} & a_{24} \\ a_{31} & a_{32} & 0 & a_{34} \\ a_{41} & a_{42} & a_{43} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix}, \qquad (1)$$

or equivalently defined as

$$\mathbf{X} = \mathbf{A}\mathbf{X} + \mathbf{N}, \qquad (2)$$

where matrix $\mathbf{A}$ is what we call the dependence matrix and x_i, $i = 1, 2, 3, 4$, are the expression profiles for each gene cluster. There is a noise-like term $\mathbf{N}$, which could be contributed by the model mismatch (i.e. those clusters' expression profiles may not be completely linearly dependent) and measurement uncertainty from microarray experiments. For simplicity, the noise-like term is modeled as a Gaussian random vector. Later, we will show that the dependence matrix and statistics of the noise term could be used to distinguish cancer and normal samples.

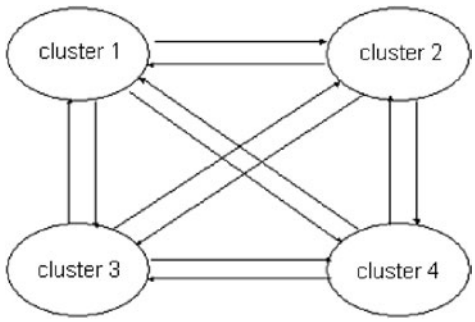

Fig. 1. Ensemble dependence model.

Equation (1) may appear similar to the space–time model of a discrete linear time invariant system in control theory. However, it is quite different. In the state–space model of a discrete linear time invariant system, matrix $\mathbf{A}$ describes how the system state will evolve from the current time step to the next time step. In our case, there is no time concept in the dependence model. The $\mathbf{X}$ vectors on both sides are actually the same. Therefore, the elements of the dependence matrix $\mathbf{A}$ do not imply any time evolution, but only indicate to what extent one gene cluster is dependent on another cluster.

MODEL-BASED CLASSIFICATION

Since not all genes' expression profiles are informative in understanding the difference between cancer and normal cases, feature selection is needed to exclude irrelevant genes. And, as required in the EDM, gene clustering is performed to group together genes with similar expressions. After feature selection and clustering, selected genes are divided into several groups. Then, the proposed EDM is used to describe the dynamics of gene clusters—one model for the cancer case and another for the normal case. With these two dependence models, a hypothesis-testing-based method is applied to classify cancer and normal data. The main flow of the proposed classification method is shown in Figure 2. It includes four main components: feature selection, gene clustering, EDM and hypothesis-testing. We will discuss these components in turn.

Feature selection

In this study, we employ two feature selection methods. The t-Test feature selection criterion is quite popular in microarray analysis. In the t-test, each gene is given a score which evaluates the similarity between its expression profiles in cancer and normal samples. All genes are ranked according to their t-test scores. A P-value is chosen, and genes with scores lower than the P-value are believed to behave most differently between cancer and normal samples.

We also apply another feature selection criterion used in Golub *et al.* (1999) and Slonim *et al.* (2000). Equation (3) is used to calculate a score for each gene:

$$F(x_j) = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-} \right|, \qquad (3)$$

where, μ_j^+ and σ_j^+ are the mean and standard deviation of gene j's expression level in cancer samples, and μ_j^- and σ_j^- are the mean and standard deviation of gene j's expression level in normal samples. Similarly, genes are ranked according to $F(x_j)$ scores. Compared

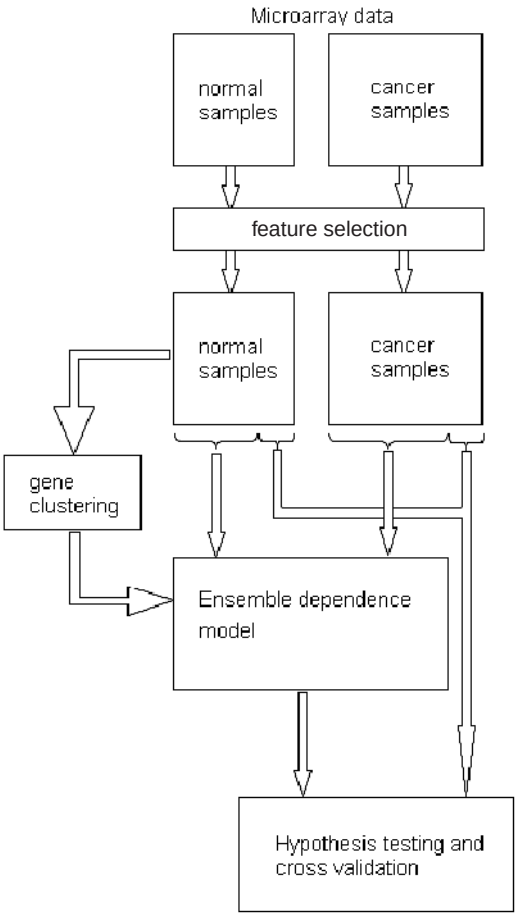

Fig. 2. Classification procedure.

with the t-test approach, for this criterion, genes with the highest scores are believed to behave most differently between cancer and normal samples.

Clustering of genes

As mentioned above, a proper way of gene clustering is required by the EDM. Three standard clustering algorithms are compared: K-means (Tavazoie *et al*., 1999), SOM (Kohonen, 1997) and Gaussian mixture model (GMM) (Steinhoff *et al*., 2003). In these clustering algorithms, the number of clusters can be predefined, as we do in the proposed dependence model. However, K-means clustering is an unstructured method, and it depends more on algorithm initials. SOM is a soft clustering method, but it blurs the difference between adjacent clusters, which is what we want to examine. Therefore, GMM is chosen to cluster genes, since it is a soft clustering method, it can capture cluster difference and it is much more stable than K-means clustering.

No matter which clustering method is chosen, a measure of similarity should be defined. In this study, Euclidean distance of genes' expression profiles is chosen to measure the similarity because genes with similar expression profiles are likely to share similar functionality (Eisen *et al*., 1998). One may argue that Euclidean distance may not cluster genes correctly in terms of their functionalities. Genes in different clusters may share similar functions or be functionally closely related. For example, suppose that two genes, gene a and gene b, are directly down-regulated by each other. When expression of gene a increases, the expression of gene b decreases. In terms of the Euclidean distance of their expression profiles, gene a and gene b could be far away from each other and would thus be likely to fall into different clusters. In this case, mutual information or Euclidean distance of the expressions' derivatives as similarity criteria would be more appropriate. However, in the proposed method, the average gene expression profile over all genes within one cluster is used to represent each cluster. Even if gene a and gene b are in the same cluster, the example above will be averaged out. This is why we choose the Euclidean distance of genes' expression as the similarity criterion. Although functionally related genes may fall into different clusters, at least genes with similar behaviors will be grouped together, and thus will represent ensemble mean behaviors more clearly.

Before clustering, the number of clusters needs to be decided. The optimal number of clusters is difficult to determine, because it may depend on different diseases and different sets of genes under investigation. To determine this parameter, we examine different choices, apply the proposed classification method and suggest the best one by comparing the overall classification performance. In this study, the number of clusters is chosen to be four, an in the 'Results' section. In two of the investigated datasets, the number of normal samples is only $\sim$6, which means we cannot afford to analyze many clusters with the limited current data size. Although the appropriate number of clusters is hard to determine, in general the more clusters, the more the dependence relationship is examined, and the more the difference between cancer and normal samples can be revealed.

Hypothesis-testing

In binary hypothesis-testing problems (Poor, 1994), there are two possible hypotheses, H_0 and H_1, associated with two probability distribution functions, f_0 and f_1, on the observation space. In this study, H_0 and H_1 represent the normal case and the cancer case, respectively. Under each hypothesis, the observation Y, gene expression, follows a certain probability distribution, written as

$$\begin{aligned} H_0 &: Y \sim f_0, \\ H_1 &: Y \sim f_1, \end{aligned} \tag{4}$$

where f_0 and f_1 are the distribution of the gene expression in cancer and normal samples, respectively. A decision rule δ is a partition of the observation space Γ into Γ_1 and $\Gamma_0 = \Gamma_1^c$, where Γ_1^c is the complement set of Γ_1. In this study, the maximum likelihood (ML) approach is used to form the decision rule, that is, to compare the conditional probability of observation Y, given underlying hypothesis H_0 or H_1,

$$\Gamma_1 = \{Y \in \Gamma | f_1(Y) > f_0(Y)\}. \tag{5}$$

Model learning and classification

Given the gene-clustering result, cluster expression profiles can be easily obtained by taking the cluster average. Then, the dependence matrix **A** can be estimated row by row, based on the least squares (LS) criterion. For example, for the first row of matrix **A**,

$$x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + n_1, \tag{6}$$

using the LS criteria, coefficients $a_{1i}, i = 2, 3, 4$, which minimize noise term n_1 are estimated. The statistics of the noise-like term n_1 is estimated at the same time.

The classification procedure is illustrated in Figure 2. For each dataset, after feature selection and gene-clustering, a portion of the cancer samples are used to estimate a model for the cancer case, represented by the dependence matrix ($\mathbf{A_c}$) and the distribution of the noise term ($\mathbf{N_c}$); a portion of the normal samples are used to estimate a model for the normal case, represented by the dependence matrix ($\mathbf{A_n}$) and the distribution of the noise term ($\mathbf{N_n}$). These two models form a hypothesis-testing problem:

$$\begin{aligned} H_1 &: \mathbf{X} = \mathbf{A_c}\mathbf{X} + \mathbf{N_c}. \\ H_0 &: \mathbf{X} = \mathbf{A_n}\mathbf{X} + \mathbf{N_n}. \end{aligned} \tag{7}$$

For each incoming unknown sample X (samples not used in model learning), the ML decision rule is applied to predict whether it is cancer or normal. That is, we check whether the incoming sample fits the cancer model better or fits the normal model better, by comparing the following two log-likelihoods

$$\Pr(\mathbf{X}|H_1) = -0.5\log((2\pi)^k|\mathbf{V_c}|) - 0.5(\mathbf{X}-\mathbf{M_c})^{\mathrm{T}}\mathbf{V_c}^{-1}(\mathbf{X}-\mathbf{M_c}), \tag{8}$$

$$\Pr(\mathbf{X}|H_0) = -0.5\log((2\pi)^k|\mathbf{V_n}|) - 0.5(\mathbf{X}-\mathbf{M_n})^{\mathrm{T}}\mathbf{V_n}^{-1}(\mathbf{X}-\mathbf{M_n}), \tag{9}$$

where k is the number of clusters, and $\mathbf{V_c}$, $\mathbf{M_c}$ and $\mathbf{V_n}$, $\mathbf{M_n}$ are the first- and second-order statistics of the Gaussian noise-like terms in cancer and normal cases, respectively.

DATASETS

Since in general cDNA microarray gene expression data follows a standard format and preprocessing operations (e.g. normalization), five publicly available cDNA datasets are investigated in detail first. Each of them contains both cancer samples and normal samples. They are a gastric cancer dataset (Chen *et al*., 2003) containing 90 cancer samples and 22 normal samples; a liver cancer dataset (Chen *et al*., 2002) containing 82 cancer samples and 74 normal samples; a prostate cancer dataset (Dhanasekaran *et al*., 2001) containing 4 stages of samples [normal adjacent prostate (NAP), benign prostatic hyperlasia (BPH), localized prostate cancer (PCA) and metastatic cancer (MET)] which can be roughly regarded as 15 normal samples (7 NAP and 8 BPH) and 25 cancer samples (14 PCA and 11 MET); a cervical cancer dataset (Wong *et al*., 2003), containing 25 cancer samples and 8 normal samples; and a lung cancer dataset (Garber *et al*., 2001), containing 37 cancer samples and 6 normal samples.

To be complete, we also investigate three Affymetrix datasets and one proteomic dataset.

RESULTS

For each dataset, we use Golub's approach for feature selection, employ the GMM to group selected genes into four clusters and apply the proposed classification scheme to perform leave-one-out cross-validation (Antoniadis *et al*., 2003). The results are shown in Table 1. From Table 1, we can see that the proposed scheme yields high classification accuracy for the first three datasets. For the last two datasets, because there are only 6–8 normal samples, a lack of training data results in relatively poor classification performance for normal samples. However, the proposed model can still make the correct classification for cancer samples.

In the reference papers mentioned in the 'Datasets' section, a hierarchical clustering method is applied to group samples. Since hierarchical clustering does not give precise classification results, it is hard to compare the proposed method with it. To examine the proposed scheme, we compare it with the widely applied linear SVM approach. The SVM algorithm is a supervised machine-learning algorithm. It is a powerful tool in classification and pattern recognition commonly used in the areas of face detection (Jonsson *et al*., 2002), speaker/speech recognition (Dong and Zhaohui, 2001) and handwriting recognition (Choisy and Belaid, 2001). It has also been applied to the problem of microarray data classification (Furey *et al*., 2000; Rifkin *et al*., 2003), where SVM is shown to provide excellent classification performance.

Table 1. Correct classification rate of ensemble dependence model for different datasets, with t-test feature selection and number of clusters being four (%)

	Correct classification for cancer samples	Correct classification for normal samples	Overall classification rate
Gastric cancer	100	100	100
Liver cancer	97.5	100	98.72
Rostate cancer	100	93.3	97.5
Ervical cancer	100	75	93.9
Lung cancer	100	66.7	95.35

In Table 2, for linear SVM and EDM, different feature selection approaches and different choices of clusters are examined using the gastric cancer dataset. From Table 2, we can see that the choice of feature selection does not affect the classification performance significantly. We believe that using a purely mathematical criterion to select genes is not enough, and that a more meaningful gene selection method which can incorporate biological knowledge is desirable. In the proposed method, different choices of the number of clusters yield slightly different results. Although it is hard to conclude which choice is the best, in general, with sufficient samples, the more clusters, the more the dependence relationship is examined, and thus the better the classification performance that can be achieved. Since the number of samples is limited, we cannot afford to analyze many clusters. As illustrated in Table 2, the performance of the five-cluster case is worse than that of the four-cluster case. The number of clusters is heuristically chosen to be four. We also investigated four other datasets and observed similar results (see Supplemental website).

From Table 2, for we also notice that the linear SVM and the proposed algorithm perform comparably, both providing very high classification accuracy. An interesting observation during the result-checking procedure is that the classification errors in nearly all leave-one-out validation experiments happen with the same two samples, which may be because of sample mislabeling. We will explore this issue further in our future related work.

Although the SVM and EDM provide similar classification performance, it is worth mentioning that the proposed approach has its advantages. The linear SVM is a hard test approach since a hyperplane in feature space is generated to classify test samples. In the proposed EDM, two likelihoods are evaluated to determine the class index. The proposed scheme is a soft test approach, where not only the class index is determined, but also the confidence level of each classification operation can be obtained.

It is worth mentioning that all the five datasets reported above are from cDNA microarray experiments. Besides cDNA, there are commercial oligobased expression arrays, such as Agilent's 60mer platform and Affymetrix's 25mer genechip® system. Since different

Table 2. Classification performance comparison on gastric cancer dataset (%)

	Golub's approach 100 genes	Golub's approach 500 genes	*t*-test 3319 genes	All features 6688 genes
Linear SVM	98.8/95.4	98.8/100	98.8/100	98.8/100
EDM 2 clusters	98.8/95.4	98.8/95.4	98.8/100	98.8/100
EDM 3 clusters	98.8/100	98.8/95.4	100/100	98.8/100
EDM 4 clusters	98.8/100	98.8/100	100/100	98.8/100
EDM 5 clusters	98.8/90.9	98.8/100	100/100	98.8/100

'EDM # clusters' means ensemble dependence model with choice of # clusters. In each cell, '#/#' means 'correct classification rate for cancer samples/correct classification rate for normal samples'.

normalization schemes are utilized for these two systems (loess for two-channel normalization with Agilent and MAS5, RMA, GC-RMA and dChip with Affymetrix), for a specific classifier the resulting classification performance may be sensitive to different normalization preprocessing. For instance, it may be sensitive to different levels of normalization from mild (MAS5) to robust (GC-RMA). Here we apply a simple normalization approach: expression data is normalized by the mean intensity of each experiment, as suggested by Alon *et al*. (1999). To illustrate the classification performance, we investigated three oligonucleotide microarray datasets: the colon cancer, lung cancer and prostate cancer datasets. We notice that the overall classification performance ranges from 85 to 98% for different types of cancer. Also, we notice that the performance of the proposed EDM approach is comparable to that of the SVM scheme. To further investigate the generality of the proposed EDM, we also applied our model to a proteomics dataset for ovarian cancer, where we obtained a classification performance of 97.63% (see Supplemental website for detailed information). This example indicates that the proposed EDM might be generally applicable to both gene and protein expression data. We will further examine this issue in future work.

MODEL-BASED PREDICTION AND PERFORMANCE ANALYSIS

Prediction in the eigen domain

The proposed EDM yields excellent classification performance. Now we want to explore intuitively why it works well. Below is a typical example of the estimated cancer dependence matrix $\mathbf{A_c}$ and the normal dependence matrix $\mathbf{A_n}$:

$$\mathbf{A_c} = \begin{bmatrix} 0 & 0.3676 & 0.1098 & -0.0398 \\ 1.6274 & 0 & -0.5400 & 0.0067 \\ 0.2103 & -0.2336 & 0 & 0.3922 \\ -0.1537 & 0.0058 & 0.7912 & 0 \end{bmatrix}. \quad (10)$$

$$\mathbf{A_n} = \begin{bmatrix} 0 & 0.4502 & 0.5154 & -0.4208 \\ 1.8188 & 0 & -1.0142 & 0.5021 \\ 0.6592 & -0.3210 & 0 & 0.7028 \\ -0.7767 & 0.2294 & 1.0145 & 0 \end{bmatrix}. \quad (11)$$

Looking at these two matrices directly does not reveal a clear difference. However, when exploring the eigenvalue domain, we observe that there are clearly two different patterns. In Figure 3, 200 different subsets of the gastric cancer dataset are used to estimate cancer and normal dependence matrices. Their eigenvalues are calculated and plotted. It is clear that, in general, the eigenvalues for the normal dependence matrix have larger absolute values than those of the cancer case. The difference is most distinct at the smallest eigenvalue. We believe that the different patterns in the eigenvalue domain could play an important role in predicting whether an unknown sample is normal or cancer.

Recall that, after gene-clustering, the dependence matrix is obtained from cluster expression profiles. What is the relationship between cluster expression profiles and the eigenvalue pattern of the dependence matrix? What kind of cluster expression profiles will result in the two different patterns observed in Figure 3? To answer these questions, an ideal case is defined where there is no noise-like term in Equation (1), meaning that the four cluster expression profiles are completely linearly dependent and that their rank is three. In other words, each cluster expression profile could be exactly written as the linear combination of the other clusters' expression profiles. Thus, the noise-like term is zero. More specifically, if the four clusters' expression profiles satisfy

$$x_1 = \alpha_1 x_2 + \alpha_2 x_3 + \alpha_3 x_4, \quad (12)$$

then the noise-like term is zero. In this case, the dependence matrix will have the special structure

$$\mathbf{A_{ideal}} = \begin{bmatrix} 0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \frac{1}{\alpha_1} & 0 & -\frac{\alpha_2}{\alpha_1} & -\frac{\alpha_3}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & 0 & -\frac{\alpha_3}{\alpha_2} \\ \frac{1}{\alpha_3} & -\frac{\alpha_1}{\alpha_3} & -\frac{\alpha_2}{\alpha_3} & 0 \end{bmatrix}. \quad (13)$$

We can show that the eigenvalues of the above matrix are 1, 1, 1, -3, no matter what the values of $\alpha_i, i = 1, 2, 3$. We define the above case in Equation (13) as the ideal case.

Based on the ideal case model, we gradually introduce larger and larger random variation to make the four cluster expression profiles more and more noisy. At each variation level, a dependence matrix is estimated as in the 'Ensemble dependence model' section, and the corresponding eigenvalues are calculated. Compared with the ideal case, as the cluster expression profiles suffer more and more noisy variations, the eigenvalues of their dependence matrix will change and follow the trends shown in Figure 4. Compared with Figure 3, it can be suggested that the cluster expression profiles in cancer samples correspond to a much larger variation level than those of the normal samples, which means the gene clusters' behavior in cancer samples is much more noisy than in normal samples. Here

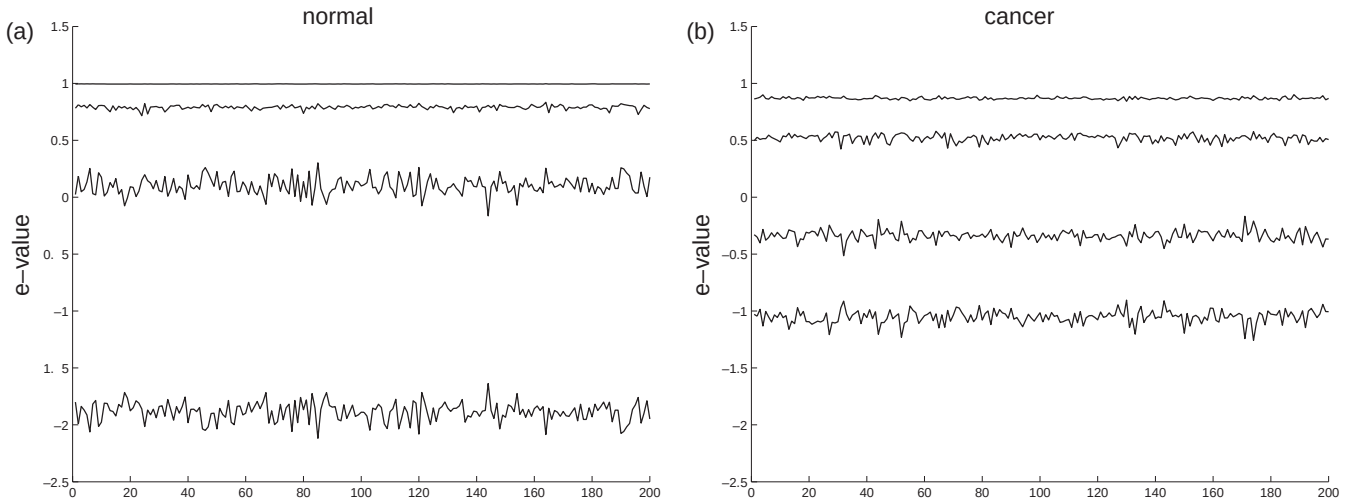

Fig. 3. Eigenvalue pattern of the gastric dataset. (**a**) Shows the four eigenvalues of normal dependence matrices, estimated using 200 different subsets of gastric cancer data. (**b**) Shows the four eigenvalues of cancer dependence matrices, estimated using 200 different subsets of gastric normal data.

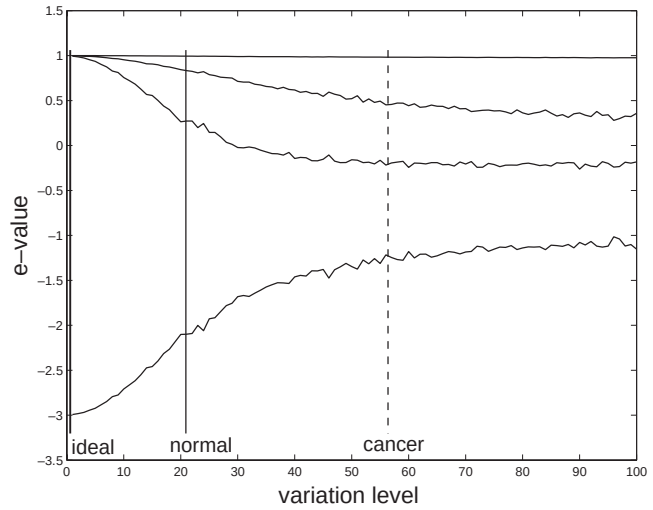

Fig. 4. The horizontal axis is variation level, which indicates how noisy the four cluster expression profiles are. As the cluster expression profiles become more noisy because of diseases, the eigenvalues of the corresponding dependence matrix will change, following the curves.

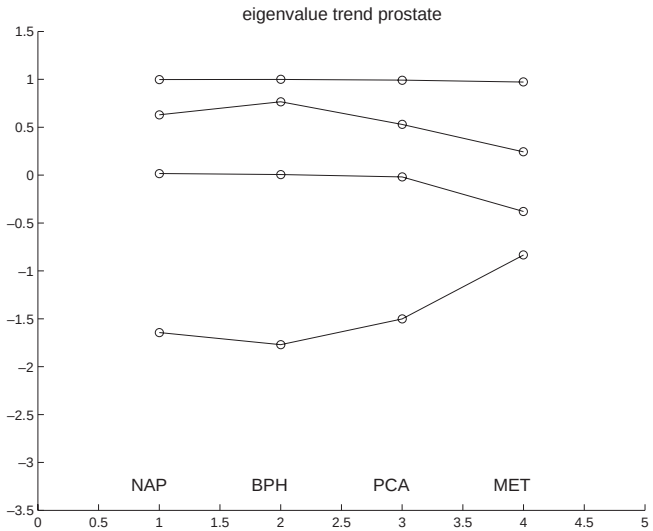

Fig. 5. Trend of eigenvalue change in the four-stage prostate dataset.

we propose explaining this intuitively. In the normal samples, the gene clusters' dependence relationship is clearer, and gene clusters work more cooperatively to maintain genetic stability. On the other hand, in the cancer case, the dependence relationship between gene clusters is overwhelmed by a large variation caused by diseases, which thus makes gene clusters work less cooperatively and makes the cell system become worse and worse. Moreover, the transition stage between normal and cancer patterns suggests that the resulting eigenvalue pattern from the proposed models can be used as a feature to predict the early stage of cancer development, i.e. whether a sample is in transition from healthy to cancer.

To support the above argument, we use the prostate cancer dataset as an example. As mentioned earlier, it contains four stages of data, NAP, BPH, PCA and MET, which can be simply regarded as being from normal (NAP and BPH), to early cancer stage (PCA), to cancer in late stage (MET). The dependence matrix and eigenvalues of each stage are calculated. As shown in Figure 5, the overall trend of eigenvalues from normal to cancer follows the trend in Figure 4, which verifies the above argument. However, since what Figure 4 shows are statistical-mean curves, there is a certain probability of error, especially under limited learning data size. One possible solution in practical clinical use is to gather more samples from a single person with the hope of averaging out statistical error.

Performance analysis

As indicated in Equation (1), a linear model is assumed, and the noise-like term could be contributed both by the model mismatch and by the microarray experiment process. One may argue that a linear

assumption may not truly fit the classification problem of interest here. In this subsection, we plan to examine the effect of model mismatch on classification accuracy.

Based on the observations about eigenvalue patterns above, we suggest that the underlying true models for the normal and cancer cases can be described as

$$\text{Normal case: } \mathbf{X} = \mathbf{A_{ideal}}\mathbf{X_0} + \mathbf{N_n}, \tag{14}$$

$$\text{Cancer case: } \mathbf{X} = \mathbf{A_{ideal}}\mathbf{X_0} + \sigma\,\mathbf{\Delta X} + \mathbf{N_c}, \tag{15}$$

where $\mathbf{N_c}$ and $\mathbf{N_n}$ are white Gaussian random vectors whose variance is chosen to yield a similar eigenvalue pattern to that observed in the real datasets. The expressions $\mathbf{X_0}$ are generated according to the ideal case model, i.e. $\mathbf{X_0} = \mathbf{A_{ideal}}\mathbf{X_0}$; $\mathbf{A_{ideal}}$ is a matrix with the structure defined in Equation (13); $\mathbf{N}$ is a Gaussian noise term; $\mathbf{\Delta X}$ represents the unit variation and σ is a factor representing the variation level. The vector $\mathbf{\Delta X}$ represents different types of model mismatch (e.g. non-linear feature of the model). Here we consider one form of a second-order non-linear variation vector $\mathbf{\Delta X}$ whose i-th element is defined as $\Delta X(i) = X_0(i)^2 - b_0$, with b_0 being the mean of the $X_0(i)^2$. In our proposed scheme, the models representing the normal and cancer hypotheses are described in Equation (7). Therefore, there is model mismatch between the model in Equation (7) and the above underlying true model. Specifically, a linear model $\mathbf{A_n}$ as in Equation (7) is estimated to approximate the model in Equation (14), which is based on both $\mathbf{A_{ideal}}$ and an unknown vector $\mathbf{X_0}$; a simple linear model $\mathbf{A_c}$ is estimated to approximate the model in Equation (15), which contains a non-linear element. The larger the factor σ, the higher the level of non-linearity observed in the true model. Our purpose is to evaluate the effects of this mismatch on the proposed scheme by examining the classification performance loss, compared with the hypothesis-testing approach, assuming the true models in Equations (14) and (15) are known. Clearly, the classification accuracy of the latter approach serves as a performance bound since, in practice, the information of $\mathbf{A_{ideal}}$, $\mathbf{X_0}$ and $\mathbf{\Delta X}$ is not available.

We now describe how to generate simulated samples. Based on the estimated distribution of normal gastric samples from experiment data and the ideal case model, we simulated 1000 $\mathbf{X_0}$ samples. Based on these generated $\mathbf{X_0}$ samples, half of them are used to generate normal samples, with the noise term $\mathbf{N}$ added, according to the model in Equation (14). The other half is used to generate cancer samples, as in the model in Equation (15), which simulates the model mismatch in the cancer case.

In the model-learning part of the proposed scheme, because of the unknown vector $\mathbf{X_0}$, neither estimated $\mathbf{A_c}$ nor estimated $\mathbf{A_n}$ will be the same as $\mathbf{A_{ideal}}$. And in the simulation setting, the model for the cancer case suffers more model mismatch than the mismatch in the normal case since it includes a non-linear component $\sigma\,\mathbf{\Delta X}$. In other words, the estimated $\mathbf{A_c}$ is used for taking into account both the unknown $\mathbf{X_0}$ and the non-linear variation $\sigma\,\mathbf{\Delta X}$.

In Figure 6, we report the classification performances of the hypothesis-testing approaches, using the proposed model in Equation (7) and the underlying true model as in Equation (14) and (15), where the classification accuracy rate P_d versus the factor σ used in Equation (15) is plotted. From Figure 6, we can see that, as the factor σ increases, meaning that the underlying model for the cancer case drifts away from the model for the normal case, the classification error reduces. We also notice that there is a performance

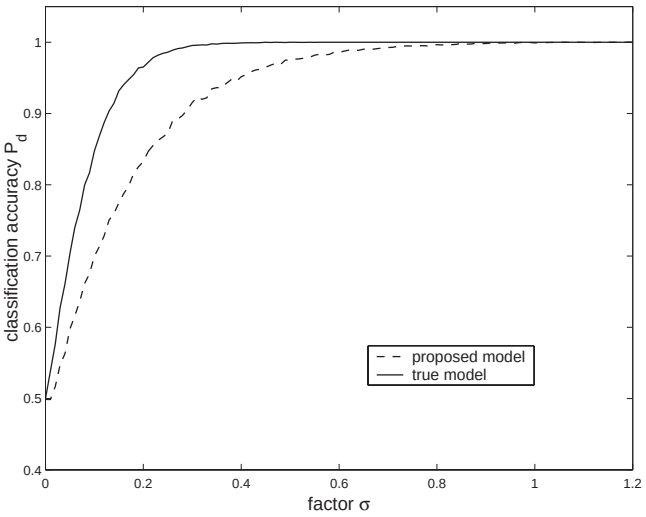

Fig. 6. Correct classification accuracy rate P_d versus the factor σ in model (15). The larger σ, the more non-linear the underlying model in Equation (15) is.

loss, though not significant, when applying the proposed scheme in Equation (7), compared with the performance bound achieved by assuming the true underlying non-linear model. The peak performance loss, 0.1589, occurs around $\sigma = 0.14$, where the correct classification rate for the proposed scheme in Equation (7) is 0.7284 and the maximum classification accuracy is 0.8873 when we assume that we know the underlying non-linear model exactly. The performance of the proposed model follows the same tendency as the true non-linear model. Since it is not practical to estimate the underlying model where $\mathbf{A_{ideal}}$, $\mathbf{X_0}$ and $\sigma\,\mathbf{\Delta X}$ are all unknown, it is encouraging to see that the proposed simple model does not demonstrate a significant classification performance loss. It is also interesting to observe that, when σ is large enough, meaning that the underlying model is highly non-linear, the proposed scheme in Equation (7) provides high classification accuracy which almost coincides with that of using the true non-linear model. Therefore, although the true underlying model in Equation (15) for the cancer case is not linear, the proposed linear model may not be a good approximation for the underlying true non-linear model; however, it works well for classification purposes.

CONCLUSION

We developed an EDM to classify cancer and normal samples, using microarray gene expression data. The results on real datasets show that the proposed method yields high accuracy in identifying cancer and normal samples. We also compared the proposed approach with the widely applied SVM algorithm. Although these two algorithms show similar performance, our algorithm presents a fundamental departure from the existing SVM approach to classification because it develops a more plausible EDM by taking genes' group behaviors and interactions into account, and thus may have potential to classify intransigent data at which other classifiers balk.

An interesting observation is noted in the eigen domain analysis: two distinguishing patterns of the eigenvalues of the dependence models are noted for the cancer and normal hypotheses. By examining one prostate cancer dataset, we also illustrated the 'expected'

changes in the eigenvalue pattern from the ideal case to the normal case, and further to the cancer case. This example suggests that the eigenvalue pattern changes gradually from a healthy status to cancer status. Since such an eigenvalue is an indicator of genes' ensemble dependence (cooperative) status, the eigenvalue pattern is promising for serving as a feature for the prediction of the transition from the healthy stage to the cancer stage and the early stage of cancer development, and thus for potential cancer diagnosis usage. Therefore, we plan to further explore and verify this promising approach in future study.

REFERENCES

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Antoniadis,A. *et al*. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.

Chang,J. *et al*. (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Mechanisms of Disease*, **362**, 362–369.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Cell. Biol.*, **13**, 1929–1939.

Chen,X. *et al*. (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Cell. Biol.*, **14**, 3208–3215.

Choisy,C. and Belaid,A. (2001) Handwriting recognition using local methods for normalization and global methods for recognition. In *Proceedings of Sixth International IEEE Conference on Document Analysis and Recognition*, 23–27.

Dhanasekaran,S. *et al*. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

Dong,X. and Zhaohui,W. (2001) Speaker recognition using continuous density support vector machines. *Electronics Letters*, **37**, 1099–1101.

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*, 2nd edn. John Wiley & Sons, Inc.

Eisen,M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14683–14688.

Furey,T. *et al*. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Garber,M.E. *et al*. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 12784–12789.

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Ismail,S. *et al*. (2000) Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer Res.*, **60**, 6744–6749.

Jonsson,K. *et al*. (2002) *Journal of Image and Vision Computing*, **20**, 369–375.

Kohonen,T. (1990) The self-organizing map. *Proceedings of the IEEE*, volume 78, Issue 9, Sept. 1990, pp. 1464–1480.

Kohonen,T. (1997) *Self-organizing Maps*. Springer, Berlin.

Lockhart,D. and Winzeler,E. (2000) Genomics, gene expression and DNA arrays, *Nature*, **405**, 827–846.

O'Neill,M. and Song,L. (2003) Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics*, **4**, 28–41.

Poor,H.V. (1994) *An Introduction to Signal Detection and Estimation*. Springer Texts in Electrical Engineering.

Rifkin,R. *et al*. (2003) An analytical method for Multi-class molecular cancer classification. *SIAM Rev.*, **45**, 706–723.

Slonim,D., Tamayo,P., Mesirov,J., Golub,T. and Lander,E. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB)*, Universal Academy Press, Tokyo, Japan, 263–272.

Steinhoff,C., Müller,T., Nuber,U.A. and Vingron,M. (2003) Gaussian mixture density estimation applied to microarray data. *RECOMB, LNCS* (Lecture Notes in Computer Sciences), **2810**, pp. 418–429.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 218–285.

Van't Veer,L. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wu,X., Chen,Y., Bernard,R, Yan,A. (2004) The local maximun clustering method and its application in microarray gene expression data analysis. *EURASIP J. Appl. Signal Proc.*, **1**, 51–61.

Wong,Y.F. *et al*.(2003) Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. *Clin. Cancer Res.*, **9**, 5486–5492.

Young,R., (2000) Biomedical discovery with DNA arrays. *Cell*, **102**, 9–15.