Ensemble Dependence Model Based Cancer Classification using Gene Microarray Data

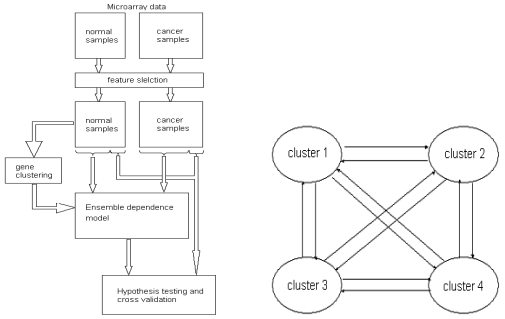
Peng Qiu, [2]Z. Jane Wang and K. J. Ray Liu
Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
[2]Department of Electrical and Computer Engineering, University of British Columbia, Canada

***Abstract*— 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 level. Currently, various clustering methods have been proposed in the literature to classify cancer and normal samples based on microarray data, and they are dominantly data-driven approaches. In this paper, we propose an alternative approach, a model-driven approach. We propose an ensemble dependence model, aiming 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 five cancer data sets, and it is noted that the proposed method yields very promising performance. We further analyze the eigen domain of the proposed method, and discovered different patterns between cancer and normal samples.**

I. Background and Proposed Scheme

Current methods for the classification of microarray gene expression data can be mainly divided into two categories. One is based on clustering, which can be used to distinguish cancer and normal samples, and subtypes of cancers. Example schemes include Hierarchical clustering, Local Maximum clustering, Self-Organizing Map, and K-means clustering. These clustering methods are mainly data-driven approaches. Usually, they do not require much prior assumption, i.e., the underlying model. However, determining the number of clusters is a challenging problem itself, and there lacks of widely-accepted measures to evaluate the clustering performance. The other category is mainly based on machine-learning approach. Motivated by the success of machine learning algorithms in image and speech processing, many researches have been reported to apply them to microarray data analysis. For example, support vector machine and neural network analysis. Machine learning methods generally yield better results than that of the traditional clustering methods.

In this paper, we propose an ensemble dependence model based classification approach, as illustrated in Fig 1(a). It includes four main components, feature selection, gene clustering, ensemble dependence model and hypothesis testing. Due to 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 in a right way, the noise level in the resulting cluster expression will be reduced, and we will be able to reveal the ensemble dynamics of gene clusters.



(a) (b)

Fig. 1. (a) classification procedure; (b) ensemble dependence model

Since not all genes' expression profiles are informative in understanding the difference between cancer and normal samples, feature selection is needed to exclude irrelevant genes. As mentioned above, gene clustering is performed to group together genes with similar expression. To average out experiment noise and enhance genes' common expression within each cluster, 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, as shown in Fig 1(b). Linear dependence relationship is studied, in Equation (1), where a_{ij} represents an inter-cluster dependence relationship. The so-called self-regulation is assumed to be zero, i.e. $a_{ii} = 0,\ i = 1, 2, 3, 4$. Because cluster average is used to represent each cluster, intra-cluster dependence relationship is averaged out.

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

or equivalently defined as

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

where, $\mathbf{A}$ matrix is what we call the dependence matrix; x_i, $i = 1, 2, 3, 4$ are the expression profiles for each gene cluster. The noise-like term $\mathbf{N}$ is contributed by model mismatch and microarray measurement uncertainty. For simplicity, the noise-like term is modelled as Gaussian random vector.

Given the gene-clustering result, cluster expression profiles can be easily obtained by taking the cluster average. Then, the dependence matrix $\mathbf{A}$ can be estimated row by row,

	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%
prostate cancer	100%	93.3%	97.5%
cervical cancer	100%	75%	93.9%
lung cancer	100%	66.7%	95.35%

TABLE I

CORRECT CLASSIFICATION RATE OF ENSEMBLE DEPENDENCE MODEL FOR DIFFERENT DATA SETS (500 GENES 4 CLUSTERS)

	Golub's approach 100 genes	Golub's approach 500 genes	T-test 3319 genes	All features 6688 genes
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%

TABLE II

CLASSIFICATION PERFORMANCE COMPARISON ON GASTRIC CANCER DATA SET. "#/#" MEANS "CORRECT CLASSIFICATION RATE FOR CANCER SAMPLES / CORRECT CLASSIFICATION RATE FOR NORMAL SAMPLES"

based on the least squares (LS) criterion. For each data set, after feature selection and gene-clustering, the dependence matrix and noise distribution for cancer and normal cases are estimated separately. ($\mathbf{A_c}$, $\mathbf{N_c}$, $\mathbf{A_n}$, $\mathbf{N_n}$). These two models form a hypothesis-testing problem:

$$\begin{aligned} H_1 &: \mathbf{X} = \mathbf{A_c}\mathbf{X} + \mathbf{N_c}. \\ H_0 &: \mathbf{X} = \mathbf{A_n}\mathbf{X} + \mathbf{N_n}. \end{aligned} \quad (3)$$

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

$$Pr(\mathbf{X}|H_1) = -0.5\log((2\pi)^k|\mathbf{V_c}|) - 0.5(\mathbf{X}-\mathbf{M_c})^T{\mathbf{V_c}}^{-1}(\mathbf{X}-\mathbf{M_c}), \quad (4)$$

$$Pr(\mathbf{X}|H_0) = -0.5\log((2\pi)^k|\mathbf{V_n}|) - 0.5(\mathbf{X}-\mathbf{M_n})^T{\mathbf{V_n}}^{-1}(\mathbf{X}-\mathbf{M_n}), \quad (5)$$

where, k is the number of clusters, $\mathbf{V_c}$, $\mathbf{M_c}$, 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.

II. RESULTS AND CONCLUSION

Five public-available data sets [1] are investigated in this work. We applied t-test and approach in [2] for feature selection, employ the Gaussian mixture model to group selected genes into several clusters, and apply the proposed classification scheme to perform leave-one-out cross-validation. From Table I, we can see that, the proposed scheme yields high classification accuracy. In this work, the number of clusters is heuristically chosen to be four, based on the classification performance from different choices of cluster number, Table II.

The dependence matrix of cancer and normal cases do not have clear different patterns entry-wisely. However, in the eigenvalue domain, we observe two clear different patterns. In general, the eigenvalues for normal dependence matrix have larger absolute value than that of cancer case. The difference is most distinguishing at the smallest eigenvalue. To explain the observations in eigen-domain, an ideal case is defined, where there is no noise-like term in Equation (1), meaning the four cluster expression profiles are completely linearly dependent. In this case, the dependence matrix will have a special structure, with eigenvalues 1,1,1,-3. From the ideal case, 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, and the corresponding eigenvalues are calculated. From the ideal case, as the cluster expression profiles suffer more and more noisy variation, the eigenvalues of their dependence matrix will follow the trends shown in Fig 2. Compared with observations in experiment data, it can be suggested that, the cluster expression profiles in cancer samples are correspondent to a much larger variation level than that of normal samples. Moreover, the transition in between cancer and normal in Fig 2 suggests that the proposed model may have potential to predict cancer development.

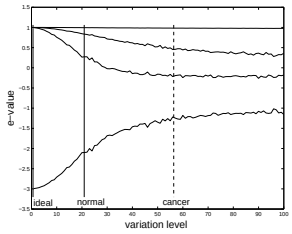


Fig. 2. The horizontal axis is variation level, which indicates how much noise is added. As the cluster expression profiles become more noisy, the eigenvalues patter will follow the above curves.

In this study, we developed an ensemble dependence model to classify cancer and normal microarray data. The proposed method yields high classification accuracy in real data sets. An interesting observation is noted in the eigen domain analysis: there are two different patterns in the eigen domain of the dependence models representing the cancer and normal cases. Moreover, the transition pattern in between shows potential in diagnosis usage.

REFERENCES

[1] Qiu,P., Wang,J.Z. and Liu,K.J.R. "Ensemble Dependence Model for Classification of Cancer and Normal Patterns Using Gene Expression Data", to appear in *bioinformatics*

[2] Slonim,D., Tamayo,P., Mesirov,J., Golub,T. and Lander,E., "Class prediction and discovery using gene expression data", *Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, 263-272, 2000.