An Efficient Semi-Unsupervised Gene Selection Method via Spectral Biclustering

Bing Liu, Chunru Wan, *Member, IEEE*, and Lipo Wang*, *Senior Member, IEEE*

***Abstract*—Gene selection is an important issue in microarray data processing. In this paper, we propose an efficient method for selecting relevant genes. First, we use spectral biclustering to obtain the best two eigenvectors for class partition. Then gene combinations are selected based on the similarity between the genes and the best eigenvectors. We demonstrate our semi-unsupervised gene selection method using two microarray cancer data sets, i.e., the lymphoma and the liver cancer data sets, where our method is able to identify a single gene or a two-gene combinations which can lead to predictions with very high accuracy.**

***Index Terms*—Gene ranking, semi-unsupervised gene selection, spectral biclustering.**

I. Introduction

TYPICALLY, the gene expression data sets contain thousands of genes while the number of tissue samples ranges from tens to hundreds. When analyzing expression profiles using machine learning methods, an important issue is gene selection for the target phenotypes. Of the thousands of genes, only a small number of them show strong correlation with a certain phenotype. For instance, for a two-way cancer/noncancer diagnosis, 50 such informative genes are usually sufficient [1].

From a machine learning point of view, gene selection is a typical feature selection problem. Successful gene selection will lead to reduction of classifier complexity and computational burden. A small number of input attribute will also facilitate visualization and interpretation of the classification results.

From biological and clinic points of view, finding the small number of important genes can help medical researchers to concentrate on these genes and investigate the mechanisms for cancer development and treatment. It may also bring down the cost for laboratory tests, because a patient needs to be tested on only a few genes, rather than thousands of genes. Furthermore, it may become possible to obtain simple rules for doctors to make diagnosis without even using a classifier or a computer, for example, in cases where values for only one or two genes are required as shown in this paper.

In the context of classification, gene selection methods fall into two categories: filter methods and wrapper methods [2]. In the filter methods, genes are selected based on their relevance to certain classes. Filter methods include, for example, statistical tests (t-test) [1], [3], [4], information gain, PCC-SNR-ECF [5], and Markov blanket based on conditional independence [6]. In wrapper methods, gene selection is "wrapped around" a particular learning algorithm: one can often obtain a very small subset of genes with relatively high accuracy [7]–[9], because the characteristics in the gene set correlate strongly with those of the learning algorithm. Both filter and wrapper methods are supervised in nature: they depend on class information. This presents a problem for clustering analysis for discovering new subtypes or phenotypes: to uncover new subtypes we need a subset of most relevant genes while the selection of these genes depends on *a priori* knowledge of cluster structure that we seek in the first place. The only solution to this problem is to perform unsupervised gene selection, i.e., selecting genes without prior subtype knowledge. One approach is iterative feature filtering [10]: given the initial clustering, one can use supervised methods to select relevant genes, which are used in turn to obtain improved clustering. Another approach is to use a two-way ordering mechanism to approximately identify irrelevant genes and discard them [11].

However, the gene subsets selected by the previous methods are usually too large. For example, for the lymphoma data set [12], the best published result shows that 48 genes are necessary in order to obtain 100% classification accuracy [13]. Such a large number of genes makes it difficult to identify which genes are responsible for the disease and, in addition, it may lead high costs in carrying out the cancer diagnostic tests. In this paper, we study a new and efficient semi-unsupervised gene selection method which results in much smaller gene subsets without prior subtype knowledge.

Kluger [14] first proposed spectral biclustering for processing gene expression data. But Kluger's focus is mainly on unsupervised clustering, not on gene selection. Inspired by his research, we propose a novel approach to gene selection based on spectral biclustering analysis. We attempt to select the most relevant genes with help of the best class partitioning eigenvectors. Compared with previous methods, our method can make accurate predictions with much smaller gene subsets.

Different from traditional gene selection algorithms [1], [3]–[6], [9], the proposed semi-unsupervised gene selection method can find much smaller and informative gene subsets without class information *a priori*. Hence, our method is most suitable for finding important genes from cancer datasets where class labels are unknown for each patient. The experiments on two microarray datasets showed that the selected genes are statistical important from a clinical application point of view.

The organization of this paper is as follows. In the next section, we describe the method of spectral biclustering. In

Manuscript received October 21, 2004; revised July 28, 2005. *Asterisk indicates corresponding author.*

B. Liu and C. Wan are with the School of Electrical and Electronic Engineering, Nanyang Technology University, Singapore 639798.

*L. Wang is with the School of Electrical and Electronic Engineering, Nanyang Technology University, Singapore 639798 (e-mail: elpwang@ntu.edu.sg).

Digital Object Identifier 10.1109/TNB.2006.875040

Section III, we propose an efficient semi-unsupervised gene selection algorithm. The experimental results and discussion are presented in Section IV. Section V concludes the paper.

II. Spectral Biclustering

Spectral biclustering [14] can be carried out in the following three steps: data normalization, bistochastization, and seeded region growing clustering.

The raw data in many cancer gene-expression datasets can be arranged in a matrix. In this matrix, denoted by A, the rows and columns represent the genes and the different conditions (e.g., different patients), respectively. Then we carry out the data normalization as follows [15]. Take the logarithm of the expression data. Perform five to ten cycles of subtracting either the mean or median of the rows (genes) and columns (conditions) and then perform five to ten cycles of row-column normalization.

Since gene expression microarray experiments can generate data sets with multiple missing values, the k-nearest neighbor (KNN) algorithm is used to fill those missing values [16].

Define $\bar{A}_{i.} = (1/m)\sum_{j=1}^{m} A_{ij}$ to be the average of ith row, $\bar{A}_{.j} = (1/n)\sum_{i=1}^{n} A_{ij}$ to be the average of jth column, and $\bar{A}_{..} = (1/mn)\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}$ to be the average of the whole matrix, where m is the number of genes and n the number of conditions.

Bistochastization may be done as follows. First, we define a matrix of interactions $K = (K_{ij})$ by $K_{ij} = A_{ij} - \bar{A}_{i.} - \bar{A}_{.j} + \bar{A}_{..}$. Then we compute the singular value decomposition (SVD) of the matrix K as given by $K = U\Lambda V^T$, where Λ is a diagonal matrix of the same dimension as K and with nonnegative diagonal elements in decreasing order, U and V are $m \times m$ and $n \times n$ orthonormal column matrices. We denote the ith column of the matrix V is denoted by $\vec{v}_i$. Therefore, we can obtain a scatter plot of experimental conditions of the two best class partitioning eigenvectors $\vec{v}_1$ and $\vec{v}_2$. The $\vec{v}_1$ and $\vec{v}_2$ are often chosen as the eigenvectors corresponding to the largest and the second largest eigenvalues, respectively. The main reason is that they can capture most of the variance in the data and provide the optimal partition of different experimental conditions [2], [14]. In general, we can obtain an s-dimensional scatter plot by using s eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_s$ (with largest eigenvalues).

Define $P = [\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_s]^T$, which has a dimension of $n \times s$. The rows of matrix P stand for different conditions, which will be clustered using SRG.

Seeded region growing clustering is carried out as follows [17]. It begins with some seeds (initial state of the clusters). At each step of the algorithm, we consider all as-yet unallocated samples, which border with at least one of the regions. Among them one sample, which has the minimum difference from its adjoining cluster, is allocated to its most similar adjoining cluster.

With the result of clustering, we can predict the distinct types of cancer data with very high accuracy [14]. In the next section, we will use such clustering result to select the best gene combinations.

III. Semi-Unsupervised Gene Selection

The proposed semi-unsupervised gene selection method includes two steps: gene ranking and gene combination selection.

As stated above, we have obtained the s best class partitioning eigenvectors $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_s$. Now we will use these s eigenvectors to rank and preselect genes.

The proposed semi-unsupervised gene selection method is based on the following two assumptions.

- The genes which are most relevant to the cancer should capture most variance in the data.
- Since $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_s$ may reveal the most variance in the data, the genes "similar" to $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_s$ should be relevant to the cancer.

The gene ranking and preselecting process can be summarized as follows. After defining the ith gene profile as $\vec{g}_i = (a_{i1}, a_{i2}, \ldots, a_{i,n})$, we use cosine measure [18] to compute the correlation (similarity) between each gene profile (e.g., $\vec{g}_i$) and the eigenvectors (e.g., $\vec{v}_j$, $j = 1, 2, \ldots, s$) as

$$R_{i,j} = \frac{(\vec{g}_i)^T \vec{v}_j}{\|\vec{g}_i\|_2 \cdot \|\vec{v}_j\|_2}, \qquad i = 1, 2 \ldots, n, \quad j = 1, 2, \ldots s \tag{1}$$

where $\|\cdot\|_2$ means vector 2—*norms*. Seen from (1), a large absolute of $R_{i,j}$ indicates a strong correlation (similarity) between ith gene and jth eigenvector. Therefore, we can rank genes as the absolute correlation values $|R_{i,j}|$ for each eigenvector. For jth eigenvector we can preselect the top l genes, denoted by G_j, according to the corresponding $|R_{i,j}|$ value for $j = 1, 2, \ldots, s$. The value l can be empirically determined. Thus, for each eigenvector of $\vec{v}_1, \ldots \vec{v}_s$, we obtain a set of genes with largest values of the Cosine Measure.

Since we have obtained the accurate cluster results by spectral biclustering, we will use them to select the best gene combinations. First, we choose all possible combinations (l^s) of s genes, with each gene corresponding to one eigenvector. Then, for each gene combination, we classify conditions with a supervised classification method: support vector machine (SVM) [19]. Therefore, the best gene combination must be the one which obtains the highest accuracy over all conditions.

IV. Experimental Results

A. Results

We now demonstrate our proposed semi-unsupervised gene selection method using two microarray data sets: the lymphoma data set [12] and the liver cancer data set [20].

The lymphoma microarray data has three subtypes of cancer, i.e., CLL, FL, and DLCL. When applying the proposed method to this data set, we obtained the clustering result with two best partition eigenvectors $\vec{v}_1$ and $\vec{v}_2$ as shown in Fig. 1. Seen from Fig. 1, the three classes are correctly divided. Then we select two sets of $l = 20$ genes according to $|R_{i,1}|$ and $|R_{i,2}|$ respectively. (Here we have set s to be two.) From the two sets of 20 genes each, we choose the two-gene combinations that can best divide the lymphoma data. We have found two pairs of genes: 1) Gene 1622X and Gene 2328X, and 2) Gene 1622X and Gene 3343X, which perfectly divide the lymphoma data. Since the results are similar to each other, we only show the result of one group in Fig. 1. Gene ID and gene names of the selecting genes in the lymphoma data set are shown in Table I, where we also show the group and the rank of genes.

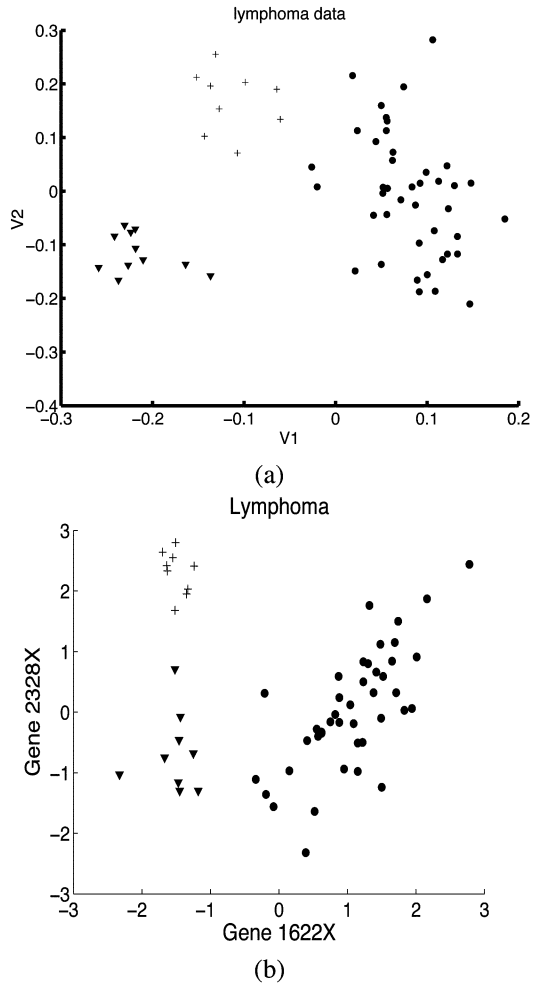


Fig. 1. Lymphoma. (a) Scatter plot of experimental conditions of the best two class partitioning eigenvectors V1, V2. (b) Scatter plot of experimental conditions of the best two-gene combination. CLL samples are denoted by pluses, DLCL by circles, and FL by triangles.

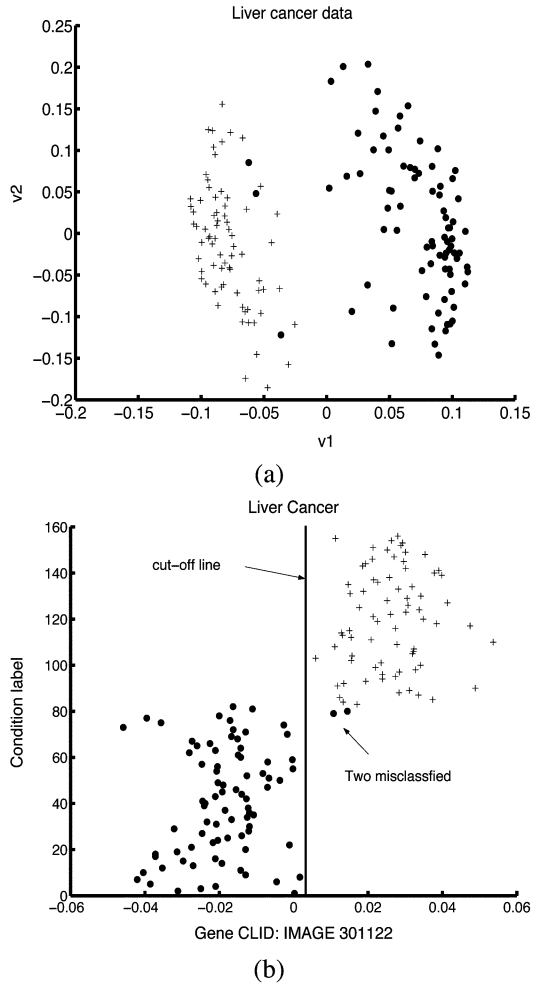


Fig. 2. Liver cancer. (a) Scatter plot of experimental conditions of the best two class partitioning eigenvectors V1, V2. (b) Scatter plot of experimental conditions with the best gene. HCC samples are denoted by circles and nontumor liver by pluses.

TABLE I
GENE IDS (CLIDS) AND GENE NAMES IN THE TWO MICROARRAY DATA SETS

Data Set	Gene ID /CLID	Gene Name	Gene Rank G1	Gene Rank G2
Lymphoma	GENE 1622X	*CD63 antigen (melanoma 1 antigen); Clone=769861	3	/
	GENE 2328X	*FGR tyrosine kinase; Clone=728609	/	3
	GENE 3343X	*mosaic protein LR11=hybrid; receptor gp250 precursor; Clone=1352833	/	4
Liver Cancer	IMAGE: 301122	116682 ECM1 extracellular matrix protein 1 Hs.81071 N79484	7	/

Note that here "/" indicates that the gene does not rank in the top 20 in this group.

We apply the method to the liver cancer data with two classes, i.e., nontumor liver and HCC. The clustering result with the two best partition eigenvectors $\vec{v}_1$ and $\vec{v}_2$ is shown in Fig. 2. We can see there are three samples misclassified and the clustering accuracy is 98.1%. Actually, we can set $s = 1$ so that the scatter plot is on a single axis. Then we select top 20 genes with the largest $|R_{i,1}|$. From the top 20 genes, we found one gene that can divide the liver cancer data well with accuracy of 98.7%. The result is shown in Fig. 2. Gene CLID and gene name of selecting gene in liver cancer data set are shown in Table I.

B. Comparisons With Published Results

We used the paired t-test method to show the statistical difference between our results and other published results. In general, given two paired sets X_i and Y_i $(i = 1, 2, \ldots, n)$ of n measured values, the paired t-test can be employed to compute a p-value between X_i and Y_i and determines whether they differ from each other in a statistically significant way under the assumptions that the paired differences are independent and identically normally distributed. The p-value is defined as follows:

$$p = (\bar{X} - \bar{Y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{X}_i - \hat{Y}_i)^2}} \tag{2}$$

TABLE II
COMPARISION OF GENERALIZATION ABILITY

Data Set	Method	Number of Genes	Test Rate (%)	(p_1, p_2)
Lymphoma	Hierarchical clustering [12]	4026	100 ± 0	$(0, 0.9937)$
	Nearst Shrunken [13]	81	100 ± 0	$(0, 0.9937)$
	Spectral Biclustering	2 ± 0	99.92 ± 0.37	(NA, NA)
Liver Cancer	Hierarchical clustering [20]	1648	98.10 ± 0.11	$(0, 0.9973)$
	Spectral Biclustering	1 ± 0	98.70 ± 0.08	(NA, NA)

Both the mean test accuracy and standard deviations are shown in this table. The average number of genes, as well as the standard deviations, are also reported in this table. The p_1 and p_2 value indicated the statistical significance on number of genes and the test accuracy between the spectral biclustering method and other published algorithms, respectively.

where $\hat{X}_i = X_i - \bar{X}$, $\hat{Y}_i = Y_i - \bar{Y}$, and $\bar{X}$, $\bar{Y}$ are the mean values for X_i and Y_i, respectively. Hence, all $p \in [0, 1]$, with a high p-value indicating statistically insignificant differences and a low p-value indicating statistically significant differences between X_i and Y_i.

We shuffled the order of cancer subtypes and carried out the experiments 20 times for each data set. We built the classifiers with the SVM and applied tenfold cross validation to evaluate the generalization ability. Each time we obtained the same gene selection result for each data set, but slightly different classification accuracies. We computed the p-values for both numbers of genes and classification accuracies for both data sets in Table II, which showed that the differences between the numbers of genes used in our method and other methods are statistically significant, whereas the differences between the classification accuracies between our method and other methods are not statistically significant.

For the lymphoma data set, we compared the generalization ability of one of our two gene sub sets (gene 1622X and gene 2328X) with 81 genes selected in [13] and the total 4026 genes [12]. Seen from Table II, our two-gene subsets produced a comparable generalization ability compared with 81 genes [13] and the total 4026 genes [12]. Hence, this represents a remarkable reduction in the number of genes required to achieve the same classification accuracy over the best published results in [13].

For the liver cancer data set, the single best gene which we have selected showed a slightly higher classification accuracy compared with the 1648 genes used in [20].

V. CONCLUSION

We have proposed an efficient semi-unsupervised gene selection method. We first use the best class partitioning eigenvectors as a result of spectral biclustering, to preselect genes. Then from these genes, we can select the best s-gene combinations, which can accurately divide the cancer data. Compared with previous work, our method can make accurate prediction with much smaller gene subsets.

The proposed gene selection method is based on the spectral biclustering algorithm proposed by Kluger [14]. However, Kluger focused on unsupervised clustering, i.e., finding distinctive "checkboard" patterns in matrices of gene expression data, not on gene selection. Different from [14], we focused on gene selection by combining spectral biclustering [14] with gene importance ranking, and we have greatly reduced the number of genes needed to predict particular types of tumors.

Gene selection is an important issue in microarray data processing. Selecting a small number of informative genes will lead to a great reduction of computational burden in cancer classification. Furthermore, finding a small number of important genes can help medical researchers and doctors concentrate on these genes, investigate the mechanisms for cancer development and treatment, and even make diagnosis using some very simple rules. For example, for lymphoma, we have discovered the following simple diagnostic rule from Fig. 1(a): for lymphoma patients, if the expression value of gene 1622X > -0.8, then the patient has DLCL. If the expression value of gene 1622X< -0.8 *and* the expression value of gene 2328X> 1.1, then the patient has CLL; otherwise, the patient has FL. For liver cancer, from Fig. 2(b) we found that if the expression value IMAGE 301 122 < 0.003, then the doctor can make the diagnosis that the patient has HCC; otherwise, the patient does not have HCC.

REFERENCES

[1] T. Golub and D. Slonim *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[2] D. P. Berrar, W. Dubitzky, and M. Granzow, *A Practical Approach to Microarray Data Analysis*. Norwell, MA: Kluwer.

[3] U. Alone *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[4] C. Ding, "Analysis of gene expression profiles: class discovery and leaf ordering," in *Proc. RECOMB 2002*, pp. 127–136.

[5] L. Guh, Q. Song, and N. Kasabov, "A novel feature selection method to improve classification of gene expression data," in *Proc. 2nd Asia-Pacific Bioinformatics Conf.*, 2004, pp. 161–166.

[6] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. Int. Conf. Machine Learning*, 1996, pp. 284–292.

[7] W. Li and Y. Yang, "How many genes are needed for a discriminant microarray data analysis?," in *Critical Assessment of Techniques for Microarray Data Mining Workshop*, 2000, pp. 137–150.

[8] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Res.*, vol. 11, pp. 1878–1887, 2001.

[9] F. Chu, W. Xie, and L. P. Wang, "Gene selection and cancer classification using a fuzzy neural network," in *Proc. IEEE Annu. Meeting North Amer. Fuzzy Information Processing Soc.*, 2004, vol. 2, pp. 555–559.

[10] E. Xing and R. Karp, "Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts," *Bioinformatics*, vol. 1, no. 1, pp. 1–9, 2001.

[11] C. Ding, "Unsupervised feature selection via two-way ordering in gene expression analysis," *Bioinformatics*, vol. 19, no. 10, pp. 1259–1266, 2003.

[12] A. A. Alizadeh *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, Feb. 2000.

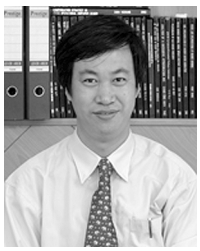
[13] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids, with application to dna microarrays," *Statist. Sci*, 2003.

[14] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray cancer data: co-clustering genes and conditions," *Genome Res.*, vol. 13, pp. 703–716, 2003.

[15] M. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 9, pp. 933–942, 1998.
[16] O. Troyanskaya and M. Cantor *et al.*, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001.
[17] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.
[18] A. Strehl, "Relationship-based clustering and cluster ensembles for high-dimensional data mining" Ph.D. dissertation, Univ. Texas, Austin, 2002 [Online]. Available: http://www.lans.ece.utexas.edu/~strehl/diss/
[19] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
[20] X. Chen and S. T. Cheung *et al.*, "Gene expression patterns in human liver cancers," *Mol. Biol. Cell*, vol. 13, pp. 1929–1939, 2002.

Bing Liu received the Bachelors degree from Xi'an Jiaotong University, China, and is currently pursuing graduate studies.

His research interests are neural networks and bioinformatics.



Chunru Wan (M'97) received the B.Eng. and M.Eng. degrees in electrical and electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1985 and 1988, respectively, and the Ph.D. degree in computer studies from Loughborough University, Loughborough, U.K. in 1996.

From 1988 to 1990, he was a researcher in Institute of Acoustic Engineering, College of Marine Engineering, Northwestern Polytechnical University. From 1990 to 1992, he was associated with the Department of Electrical and Electronic Engineering of Loughborough University. From 1992 to 1993, he was with the Department of Computer Studies of Loughborough University to pursue his Ph.D. degree. From 1993 to 1998, he worked in DSO National Laboratories as a Senior Engineer. Since 1998, he has been an Assistant Professor in the School of Electrical and Electronic Engineering of Nanyang Technological University, Singapore, while remaining an Adjunct Principal Member of Technical Staff in DSO National Laboratories. He is a member of the Editorial Board of the journal *Soft Computing*. His research interests are in signal processing, sonars, communications, parallel computing, underwater acoustics and computational mathematics.

Lipo Wang (M'97–SM'98) is with the School of Electrical and Electronic Engineering, Nanyang Technology University, Singapore. He is Area Editor of the journal *Soft Computing* (since 2002). He serves on the Editorial Board of five additional international journals and was on the Editorial Board of three other journals. He is author or coauthor of over 60 journal publications, 12 book chapters, and 90 conference presentations. He holds a U.S. patent in neural networks. He has authored two monographs and edited 16 books. His research interests include computational intelligence, with applications to data mining, bioinformatics, and optimization.

Dr. Wang has been keynote/panel speaker for several international conferences. He is Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS (since 2002), IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION (since 2003), and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (since 2005). Dr. Wang is Vice President—Technical Activities, IEEE Computational Intelligence Society (2006-2007) and served as Chair of Emergent Technologies Technical Committee (2004-2005). He has been on the Governing Board of the Asia-Pacific Neural Network Assembly since 1999 and served as its President in 2002/2003. He was Founding Chair of both the IEEE Engineering in Medicine and Biology Chapter Singapore and IEEE Computational Intelligence Chapter Singapore. He serves/served as General/Program Chair for 11 international conferences and as member of steering/advisory/organizing/program committees of over 110 international conferences.