JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 26, 649-658 (2010)

A Novel Spectral Clustering Method Based on Pairwise Distance Matrix

CHI-FANG CHIN[1], ARTHUR CHUN-CHIEH SHIH[2] AND KUO-CHIN FAN[1,3]
[1]*Institute of Computer Science and Information Engineering*
National Central University
Chungli, 320 Taiwan
E-mail: annking@iis.sinica.edu.tw
[2]*Institute of Information Science*
Academia Sinica
Taipei, 115 Taiwan
E-mail: arthur@iis.sinica.edu.tw
[3]*Department of Informatics*
Fo Guang University
Ilan, 262 Taiwan
E-mail: kcfan@csie.ncu.edu.tw

In general, the similarity measure is indispensable for most traditional spectral clustering algorithms since these algorithms typically begin with the pairwise similarity matrix of a given dataset. However, a general type of input for most clustering applications is the pairwise distance matrix. In this paper, we propose a distance-based spectral clustering method which makes no assumption on regarding both the suitable similarity measure and the prior-knowledge of cluster number. The kernel of distance-based spectral clustering is that the symmetric LoG weighted matrix constructed by applying the Laplace operator to the pairwise distance matrix. The main difference from the traditional spectral clustering is that the pairwise distance matrix can be directly employed without transformation as a similarity pairwise matrix in advance. Moreover, the inter-cluster structure is embedded and the intra-cluster pairwise relationships are maximized in the proposed method to increase the discrimination capability on extracting clusters. Experiments were conducted on different types of test datasets and the results demonstrate the correctness of the extracted clusters. Furthermore, the proposed method is also verified to be robust to noisy datasets.

Keywords: spectral clustering, laplace operator, LoG weighted matrix, pairwise distance matrix, PoD histogram

1. INTRODUCTION

Clustering technique plays an important role for data analysis in many fields. Abundant of clustering algorithms have been reported in literature. However, the dependence of clustering results on data distribution and the prerequisite of a prior-knowledge of cluster number are two open problems [1-3]. Over the past years, the research on spectral clustering has received tremendous attentions [4-8]. The most recent survey is given in [8]. The prototype of spectral clustering was first presented by Donath-Hoffman [9] and Fiedler [10, 11]. They introduced the original idea of using eigenvalue decomposition for graph partition. The kernel of spectral clustering is a graph Laplacian matrix, which is derived from a similarity graph constructed from a set of data points. Fiedler [10] had

Received March 14, 2008; revised May 23, 2008; accepted June 5, 2008.
Communicated by H. Y. Mark Liao.

proven that the algebraic connectivity is the second smallest eigenvalue of the graph Laplacian matrix. In summary, a clustering result generated by spectral clustering can be seen as performing eigenvalue decomposition on the graph Laplacian matrix derived from a similarity matrix of input data. Hence, the similarity measure is indispensable for most traditional spectral clustering algorithms since these algorithms typically begin with the pairwise similarity matrix of a given dataset. However, it is difficult in choosing an appropriate similarity measure for a given dataset in advance. To remedy this problem, we propose a distance-based spectral clustering in this paper, which starts with a pairwise distance matrix for the further eigenvector analysis.

For the given dataset, the pairwise distance matrix can be transformed to a LoG weighted matrix by applying the Laplace operator on the standard deviation of each row. Here, the Laplace operator is the second derivation of a Gaussian function. If two data points are neighboring and have similar Laplacian of Gaussian (LoG) weights to the other corresponding data points, the inner product of their row vectors in the LoG weighted matrix should be much larger than zero. In contrast, if two data points are far from each other and have entirely different LoG weights to other corresponding points, the inner product of their row vectors would be much less than zero. Thus, if the LoG weighted matrix is diagonalizable, we can use the eigenvector v with the largest eigenvalue to maximally spread out the data points with different LoG weight distributions in the LoG weighted matrix. The diagonalization of the LoG weighted matrix by the magnitude order of the elements of v is an optimization process so that the intra-cluster inner product of pairwise relationships is the maximum. Moreover, the reordered distance matrix will form block-wise structures of which each block-wise structure represents at least one cluster. We therefore propose a series of processes to extract two most distinguishable block-wise structures in the reordered distance matrix iteratively. As the results show, our proposed method demonstrates the promising performance in the correctness in extracting clusters under different kinds of datasets including the clusters of marginal connection, arbitrary shapes, and different levels of noises.

The remainder of the paper is organized as follows. The systematic description of the proposed distance-based spectral clustering method is presented in section 2. The rationale behind the proposed method is also addressed in the section. In section 3, experimental results are illustrated to demonstrate the feasibility and validity of our proposed method to various kinds of datasets. Finally, conclusions are made in section 4.

2. THE PROPOSED METHOD

The proposed distance-based spectral clustering method consists of three stages including LoG weighted matrix calculation, eigenvalue decomposition, and cluster extraction. The detail description of each stage will be elaborated in this section. The rationale behind the proposed method will also be given after the addressing of Laplacian of Gaussian (LoG) weighted matrix.

2.1 Laplacian of Gaussian (LoG) Weighted Matrix Calculation

In our proposed method, the first stage is to construct the pairwise distance matrix

from the given datasets and calculate its corresponding LoG weighted matrix based on the data adaptive standard deviations. Let $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ be a set of n input data points belonging to the d-dimensional space R^d. The constructing of pairwise distance matrix $\boldsymbol{D} \in R^{n\times n}$ is defined by $\boldsymbol{D}_{p,q} = d(\boldsymbol{x}_p, \boldsymbol{x}_q)$ for $p \neq q$ and $\boldsymbol{D}_{p,q} = 0$, where $d(\boldsymbol{x}_p, \boldsymbol{x}_q)$ represents certain distance function. The pairwise distances can be measured by using any popular distance measure. In our work, Euclidean distance is adopted as the distance measure. An important consideration in calculating the LoG weighted matrix is the estimation of data adaptive standard deviations of the corresponding pairwise distances for each data point to the other data points. The LoG weighted matrix is derived by applying the Laplace operator with adaptive standard deviation to the values in each row of pairwise distance matrix, which is defined by

$$L_{p,q} = \frac{1}{\sqrt{2\pi}\sigma_p^{\,3}}\left(1-\frac{\left|x_p - x_q\right|^2}{\sigma_p^{\,2}}\right)\left(e^{\frac{-\left|x_p-x_q\right|^2}{2\pi\sigma_p^{\,2}}}\right) = \frac{1}{\sqrt{2\pi}\sigma_p^{\,3}}\left(1-\frac{D_{p,q}^{\,2}}{\sigma_p^{\,2}}\right)\left(e^{\frac{-D_{p,q}^{\,2}}{2\pi\sigma_p^{\,2}}}\right) \quad (1)$$

where σ_p denotes the standard deviation of the elements in each row of the matrix $\boldsymbol{D}$.

To incorporate the mutual relationships between any data pairs into the LoG weighted matrix, a modification is made to the LoG weighted matrix further and defined as $\boldsymbol{L}^{\boldsymbol{S}}_{\boldsymbol{p,q}} = \frac{1}{2}\left(\boldsymbol{L}_{\boldsymbol{p,q}} + \boldsymbol{L}_{\boldsymbol{q,p}}\right)$. The resulting matrix is called a symmetric LoG weighted matrix. The characteristics and properties of LoG weighted matrix $\boldsymbol{L}$ still hold for symmetric LoG weighted matrix $\boldsymbol{L}^{\boldsymbol{S}}$.

The rationale behind the proposed method can be stated briefly as follows: since Laplacian is a well-known operator which is good at boundary detection in image analysis [11], we make use of the zero-crossing of Laplacian to be treated as the boundary between clusters for embedding the local inter-cluster information in the LoG weighted matrix from the view of each data point. That is, the Laplace operator with the estimated adaptive standard deviation based on the values of each row in the pairwise distance matrix to construct the LoG weighted matrix with an eye to achieving the inter-cluster structure. The location of zero crossing can be considered as the boundary for each data point with the other ones on estimating which data points belong to the same cluster and which data points do not belong to the same cluster. For each data point, the local inter-cluster structure can be revealed by positive or negative LoG magnitudes, where the local cluster structure are successfully embedded in the LoG weighted matrix for increasing the discrimination capability.

To facilitate the analysis of distance-based spectral clustering, assume $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T \in R^{1\times n}$ be the cluster relationship indicator and $\boldsymbol{L}$ be the LoG weighted matrix. When two data points are neighboring, the inner product of their corresponding row vectors in the LoG weighted matrix should be much higher than zero because they have similar LoG weights to the other data points. In contrast, when two data points are far from each other, their inner product would be much less than zero because the LoG magnitudes have different signs for most corresponding elements in the row vectors. Thus, we define the cluster relationship indicators between two arbitrary rows as follows:

$$y_i y_j = \left\langle \boldsymbol{L}_i, \boldsymbol{L}_j \right\rangle = \begin{cases} > 0, & \text{where } y_i \text{ and } y_j \text{ belong to the same cluster,} \\ \leq 0, & \text{where } y_i \text{ and } y_j \text{ belong to different clusters,} \end{cases} \quad (2)$$

where $\boldsymbol{L}_i$ and $\boldsymbol{L}_j$ represent the ith row and jth row in the LoG weighted matrix, respectively, and <A, B> is the inner product of vectors A and B. A reasonable criterion for choosing a good bipartition of the input dataset on maximizing the intra-cluster inner product of pairwise relationships is to maximize the following objective function

$$\sum_{i,j} y_i y_j \boldsymbol{L}_{i,j} = \sum_i y_i \sum_j \boldsymbol{L}_{i,j} y_j = \boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}. \quad (3)$$

Therefore, the vector $\boldsymbol{Y}$ that maximizes the objective function is given by the maximum eigenvalue solution to the generalized eigenvalue problem by $\boldsymbol{LY} = \lambda \boldsymbol{Y}$. Obviously, the diagonalization of the LoG weighted matrix by the magnitude order of the elements of eigenvector with the largest eigenvalue is an optimization process to maximize the intra-cluster inner product of pairwise relationships.

2.2 Eigenvalue Decomposition

The goal of eigenvalue decomposition is to find the eigenvector with the largest eigenvalue of the symmetric LoG weighted matrix selected to maximize the intra-cluster inner product of pairwise relationships, such that the indices of data points with similar pairwise distances will be reordered to other points as close as possible in the reordered matrix. In our method, the eigenvector, V_1, corresponding to the largest eigenvalue, λ_1, of $\boldsymbol{L}^{\boldsymbol{S}}$ is utilized to reorder the distance matrix $\boldsymbol{D}$. Considering the eigenvector $V_1 = (v_1^1, v_2^1, \ldots, v_n^1)^T$ corresponding to the largest eigenvalue of $\boldsymbol{L}^{\boldsymbol{S}}$, the order is defined by index permutation $\varphi(1, 2, \ldots, \boldsymbol{n}) = (\varphi_1, \varphi_2, \ldots, \varphi_n)$. For the eigenvector V_1, its permutation vector can be represented as $\varphi(V_1) = \left(v_{\varphi_1}^1, v_{\varphi_2}^1, \ldots, v_{\varphi_n}^1\right)^T$, where $v_{\varphi_1}^1 \leq v_{\varphi_2}^1 \leq \cdots \leq v_{\varphi_n}^1$. Thus, each element of the reordered distance matrix $\boldsymbol{R}$ is represented as $\boldsymbol{R}_{\varphi_p, \varphi_q} = \boldsymbol{D}_{p,q}$.

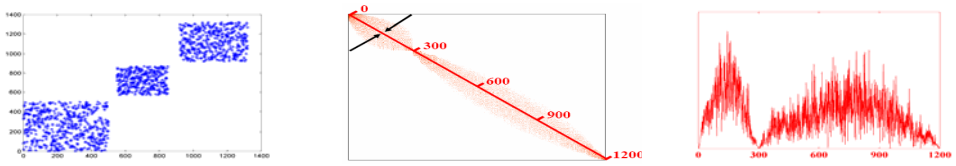
2.3 Cluster Extraction

After reordering the distance matrix, the data points belonging to the same cluster will possess close indices in the reordered distance matrix to form the nearly block-like structure with each block representing at least one cluster. To decompose the block-like structure in the reordered distance matrix, each element that represents a pairwise distance value is binarized by a cut-off threshold which is set empirically for selecting the data pairs having strong relationships. Here, the reordered distance matrix after binarization is denoted by $\boldsymbol{R}^{\boldsymbol{B}}$.

Then, we project the non-zero elements in the binarized reordered distance matrix along the diagonal direction into a one-dimensional histogram. We call it the PoD (Projection-on-Diagonal) histogram in which the order of data points is the same as the index permutation as described in section 2.2. The PoD histogram f at position i is defined as the sum of all elements along the anti-diagonal direction produced by the value of closeness relationship g_{close} at the ith position in the diagonal direction; that is,

$$f(i) = \sum_{j=i}^{j<2i} \boldsymbol{R}^{\boldsymbol{B}}_{j,2i-j} \cdot g_{close}(\boldsymbol{R}^{\boldsymbol{B}}_{j,2i-j}, \boldsymbol{R}^{\boldsymbol{B}}_{i,i}) \tag{4}$$

where $g_{close}(\boldsymbol{R}^{B}_{j,2i-j}, \boldsymbol{R}^{B}_{i,i})$ represents the closeness relationship between $\boldsymbol{R}^{B}_{j,2i-j}$ and $\boldsymbol{R}^{B}_{i,i}$ measured by an inverse proportion of their index difference. The less closeness relationship between a pair of data points in the reordered distance matrix, the less the value will be for these two data points. Here, let us use an illustrative example as shown in Fig. 1 to explain the basic concept of projecting the binarized reordered distance matrix into the PoD histogram. The original dataset is shown in Fig. 1 (a) and its corresponding binarized reordered distance matrix is shown in Fig. 1 (b). All elements along the anti-diagonal direction projecting into the main diagonal according to Eq. (4) are shown in Fig. 1 (b) where the arrows represent the projecting direction to a value on the main diagonal in the reordered distance matrix. It is noteworthy that the scale of the main diagonal in the reordered distance matrix is the same as the scale of x-axis in the PoD histogram.



(a) (b) (c)

Fig. 1. An example illustrating the projection from the binarized reordered distance matrix into the PoD histogram; (a) The original dataset; (b) Its corresponding binarized reordered distance matrix where all elements along the anti-diagonal direction are projected into the main diagonal; (c) The PoD histogram where the scale of the x-axis is the same as the main diagonal in the reordered distance matrix.

In practice, if multiple clusters exist in the input dataset, the PoD histogram will exhibit two or more significantly distinguishable distributions. To simplify the problem of distribution partition, we partition two sets of distributions by finding the most significant splitting point one at a time. A significant splitting point x_k, whose index permutation $\varphi_k = i$, is defined as the point that can apparently separate two significant distributions. That is, if the data points whose corresponding index permutations are smaller than i, the data points and the splitting point x_k belong to the same cluster. Otherwise, the data points and the splitting point x_k belong to different clusters. It is obvious that an optimal splitting point should locate at the point of the global minimum between two distributions in the PoD histogram. Since the PoD histogram is usually not smooths but saw-toothed, it is necessary to convolve the PoD histogram with a Gaussian function to filter out the high frequency signal to obtain the smoothing PoD histogram. The global minimum in the original PoD histogram can thus be found after the convolution process. In general, the position of estimated splitting point at the smoothing PoD histogram may not exactly locate at the actual position in the original saw-toothed PoD histogram due to the convolution effect. To remedy this problem, we propose a nearest-neighbor (NN) rule to correct the position of splitting point from an estimated position to the actual position, *i.e.*, the

so-called splitting point correction process. Since data points belonging to the same cluster must be close to each other, the nearest neighbor for a data point in the index of the PoD histogram must also belong to the same cluster after the bipartition process. The NN rule is used to verify the neighboring data points within a quite small window size across the position of the estimated splitting point. If not all of the points within the window satisfy the NN rule that belong to the same distribution, its neighboring position will be considered as the candidate splitting point to repeat the verification of NN rule until the data points locating within the window all satisfy the NN rule for the current splitting point. After the performing of splitting point correction process, the splitting point that passes the verification of NN rule is considered as the one which is able to accurately bipartition the current dataset.

3. EXPERIMENTAL RESULTS

Using the above proposed method, a given dataset can be firstly partitioned into two groups with each group including one or more than one clusters. Then, the same procedures are applied to each partitioned group iteratively until no significant minimum is found in the PoD histograms. Finally, the data points in each undistinguishable group are assigned to the same cluster and the total number of groups is the final cluster number. The depiction of the iterative bipartition procedure to generate the result of disjoint cluster partition can be explained by using the following illustrative example as shown in Fig. 2.

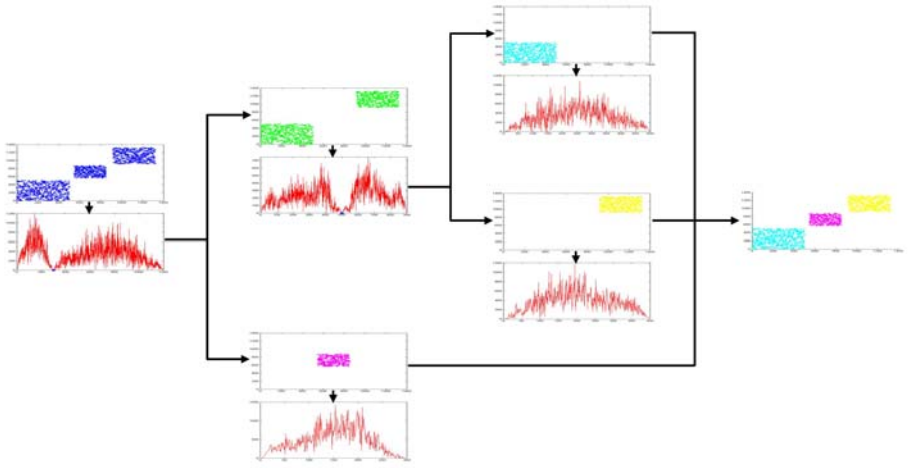
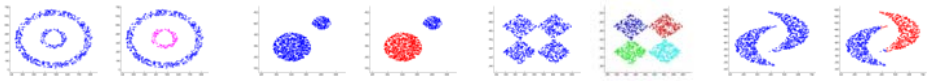
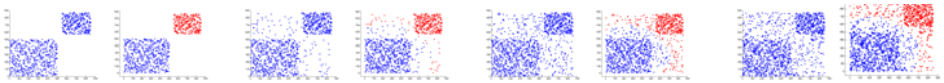


Fig. 2. An example illustrating the iterative bipartition procedure; (a) The original dataset containing three clusters and its corresponding PoD histogram; (b) The sub-datasets and the corresponding PoD histograms after the first iteration of bipartition; (c) The sub-datasets and the corresponding PoD histograms after the second iteration of bipartition; (d) The final clustering result extracted by our proposed method.



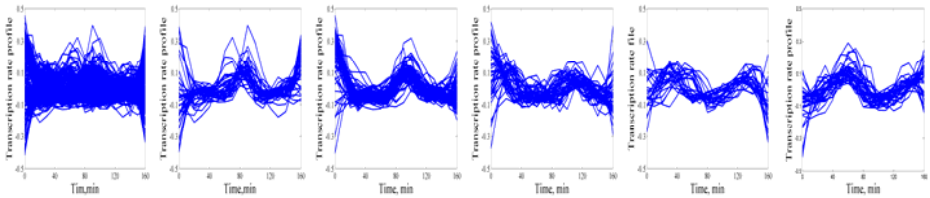
(a) (b) (c) (d)

Fig. 3. Four representative datasets with different clustering instances.



(a) (b) (c) (d)

Fig. 4. The clustering results of noisy datasets generated by our proposed method.



(a) (b) (c) (d) (e) (f)

Fig. 5. The clustering results of co-transcription rate profiles generated by our proposed method; (a) Transcription rate profiles from Yeung *et al*. that refined yeast cell cycle data includes 222 genes sampled by 17 time points; (b) Cluster of genes in Early *G*1 phase; (c) Cluster of genes in Late *G*1 phase; (d) Cluster of genes in *S* phase; (e) Cluster of genes in *G*2 phase; (f) Cluster of genes in *M* phase.

Furthermore, 29 2-D synthetic datasets including different shapes, sizes, densities, and noise levels are tested to evaluate the performances of our proposed method. For each dataset, our proposed method can successfully find the correct number of clusters. We also further analyze whether each data point assigning to the cluster is correct or not, *i.e*., the accuracy. 18 datasets are successful in data point assignment with 100% accuracy and the results for the others are also with more than 99% accuracy. Fig. 3 shows the four representative datasets used in our experiment. These datasets represent different clustering cases with different difficulties because they contain clusters of arbitrary shapes, different proximity, and varying sizes. The left sides of Figs. 3 (a)-(d) demonstrate the datasets and the clustering results generated by our proposed method are shown in the right sides of Figs. 3 (a)-(d) where the data points belonging to different clusters are represented by different colors. The result reveals that our proposed method can still successfully extract correct clusters embedded in these datasets.

Besides, noise and outlier are two main factors that usually affect the performance of clustering results. We thus examine whether noise factors affect the results generated by our method or not. The left side of Fig. 4 (a) shows the original dataset of two clusters without noise and the clustering result generated by our method is shown in the right side of Fig. 4 (a). The datasets added with different proportions of random points with 20%, 50%, and 70% are shown in the left sides of Figs. 4 (b), (c), and (d), respectively. The clustering results generated by our proposed method are shown in the right sides of Figs. 4 (b)-(d). The results demonstrate that our proposed method is robust to noises.

Furthermore, we apply the proposed method to a real case which groups transcription rate profiles in yeast genes to different clusters. Since Yeung *et al*. [13] refined yeast cell cycle genes including 222 genes which had been classified into five cell cycle phases-specific groups, it is advantageous to make use of these genes to verify the validity of our proposed method. Shown in Fig. 5 (a) are transcription rate profiles of 222 yeast cell cycle genes. Figs. 5 (b)-(f) illustrate the clustering results of co-transcription rates profiles generated by our proposed method in Early *G*1 phase, Late *G*1 phase, *S* phase, *G*2 phase, and *M* phase, respectively. Comparing with the five cell cycle phases-specific groups classified by Yeung *et al*., the results generated by our proposed method exhibit the promising performance with a 98.6% true positive rate.

4. CONCLUSIONS

In this paper, we tackle the problems occurring in traditional spectral clustering on which the clustering result is highly dependent on the choosing of similarity measure. The existence of a suitable similarity measure is seldom addressed in the literatures on traditional spectral clustering. Hence, we propose a distance-based spectral clustering which makes no assumption on regarding the selection of suitable similarity measure. That is, a pairwise distance matrix can be directly employed to construct a weighted graph without a similarity measure. The kernel of distance-based spectral clustering is the construction of LoG weighted matrix by applying the Laplace operator to the pairwise distance matrix. Since the adaptive pairwise relationships of data points are considered and the standard deviation for each row is adaptive, the local inter-cluster structure can thus be embedded in the LoG weighted matrix to increase the discrimination capability in extracting clusters. Besides, the maximization of intra-cluster pairwise relationships can be achieved by the generalized eigenvalue decomposition problem. In addition to the noiseless datasets, the experimental results also demonstrate that our proposed method possesses good capability on handling noises even the case with severe noises.

REFERENCES

1. R. T. Ng and J. Han, "Efficient and effective clustering method for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Databases*, 1994, pp. 144-155.
2. K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, Vol. 12, 2001, pp. 181-201.
3. R. Xu and D. Wunsch II, "Survey of clustering algorithm," *IEEE Transactions on Neural Networks*, Vol. 16, 2005, pp. 645-678.
4. F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, RI, 1997.
5. Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, 1999, pp. 975-982.
6. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, Vol. 22, 2000, pp. 888-905.
7. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proceedings of Neural Information Processing Systems*, 2002, pp. 849-856.
8. U. von Luxburg, "A tutorial on spectral clustering," Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics, 2006.
9. W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM Journal of Research and Development*, Vol. 17, 1973, pp. 420-425.
10. M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal,* Vol. 23, 1973, pp. 298-305.
11. M. Fiedler, "Laplacian of graphs and algebraic connectivity," *Combinatorics and Graph Theory*, Banach Center Publications, Warsaw, Vol. 25, 1989, pp. 57-70.
12. D. Marr, *Vision*, W. H. Freeman and Company, New York, 1985.
13. K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, Vol. 17, 2001, pp. 309-318.

Chi-Fang Chin (金繼昉) received her M.S. degree in Computer Science and Information Engineering from National Central University, Chungli, in 1996. From 1996 to 2000, she worked as a Lecturer at Nanya Institute of Technology, Chungli. She is now a Ph.D. candidate in the Institute of Computer Science and Information Engineering, National Central University, Chungli, Taiwan. Her research interests include clustering, machine learning, image processing, and bioinformatics.

Arthur Chun-Chieh Shih (施純傑) was born in Taipei, Taiwan, on September 26, 1966. He received the B.S. degree in Electrical Engineering from the Chinese Culture University, Taipei, in 1992, the M.S. degree also in Electrical Engineering from National Chung Cheng University, Chiayi, Taiwan, in 1994 and a Ph.D. degree in Computer Science and Information Engineering from National Central University, Chungli, Taiwan, in 1998. From October 1998 to July 2002, he worked for the Institute of Information Science, Academia Sinica, Taiwan, and the Department of Ecology and Evolution, the University of Chicago, Chicago, IL, as a Postdoctoral Fellow. He joined the Institute of Information Science, Academia Sinica, as an Assistant Research Fellow in July 2002 and became an Associate Research Fellow in 2008. His current research interests include molecular evolution, bioinformatics, and multimedia signal processing.

Kuo-Chin Fan (范國清) (S'88-M'88) was born in Hsinchu, Taiwan, R.O.C., on June 21, 1959. He received the B.S. degree in Electrical Engineering from National Tsing Hua University, Hsinchu, in 1981 and the M.S. and Ph.D. degrees from the University of Florida, Gainesville, in 1985 and 1989, respectively. In 1983, he joined the Electronic Research and Service Organization (ERSO), Taiwan, as a Computer Engineer. From 1984 to 1989, he was a Research Assistant with the Center for Information Research, University of Florida. In 1989, he joined the Institute of Computer Science and Information Engineering, National Central University, Chungli, Taiwan, where he became a Professor in 1994. From 1994 to 1997, he was Chairman of department. Currently, he is the Director of the Computer Center. His research interests include image analysis, optical character recognition, and document analysis. Dr. Fan is a member of SPIE.