Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Gene Expression Clustering: a Novel Graph Partitioning Approach

Yanhua Chen, Ming Dong, Manjeet Rege

Abstract— **In order to help understand how the genes are affected by different disease conditions in a biological system, clustering is typically performed to analyze gene expression data. In this paper, we propose to solve the clustering problem using a graph theoretical approach, and apply a novel graph partitioning model - Isoperimetric Graph Partitioning (IGP), to group biological samples from gene expression data. The IGP algorithm has several advantages compared to the well-established Spectral Graph Partitioning (SGP) model. First, IGP requires a simple solution to a sparse system of linear equations instead of the eigen-problem in the SGP model. Second, IGP avoids degenerate cases produced by spectral approach to achieve a partition with higher accuracy. Moreover, we integrate unsupervised gene selection into the proposed approach through two-way ordering of gene expression data, such that we can eliminate irrelevant or redundant genes in the data and obtain an improved clustering result. We evaluate our approach on several well-known problems involving gene expression profiles of colon cancer and leukemia subtypes. Our experiment results demonstrate that IGP constantly outperforms SGP and produces a better result that is closer to the original labeling of sample sets provided by domain experts. Furthermore, the clustering accuracy is improved significantly when IGP is integrated with the unsupervised gene (feature) selection.**

I. Introduction

In order to help understand how the genes are affected by different disease conditions in a biological system, clustering is typically performed to analyze gene expression data. The purpose of the clustering analysis has two sides. First of all, grouping genes with similar expression patterns based on the samples can help unravel relations between genes, deduce the function of genes and reveal the underlying regulatory gene network. Secondly, samples can be clustered into homogeneous groups corresponded to particular phenotypes, such as cancer or non-cancer types, so that more systematic characterization can be developed and new subtypes may be discovered. Usually, the second type of clustering is more difficult than the first one because of the small sample volume and high dimensionality of gene (feature) space.

In the literature, many clustering approaches have been applied to the analysis of gene expression data (refer to [1] for a comprehensive review and evaluations). For example, some conventional clustering algorithms, such as hierarchical clustering [2], self-organizing maps (SOM) [3] and simulated annealing [4], were applied in the early stage and proven to be useful in gene clustering. Recently, graph theoretic approaches [5], [6], [7] have drawn attention from research community and been applied to gene expression analysis. The experimental studies have shown that graph partitioning algorithms provide better performance than the conventional clustering methods on gene expression data. In particular, of these graph theoretic approaches, Spectral Graph Partitioning (SGP) [30] is the most popular one. It was shown that the optimization of the SGP objective function eventually leads to the solution of a eigen-problem.

In this paper we propose to apply a novel graph partitioning model - Isoperimetric Graph Partitioning (IGP), to group biological samples from gene expression data. In IGP, samples clustering is achieved by partitioning the graph with smallest isoperimetric ratio. The isoperimetric partitioning criterion used in IGP is intuitively similar to the one in the SGP model. However, IGP has its own advantages. First, IGP requires a simple solution to a sparse system of linear equations instead of the eigen-problem in the SGP model [17]. Second, IGP avoids degenerate cases produced by spectral approach to achieve a partition with higher accuracy [16]. Moveover, in order to address the challenges in sample clustering (curse of dimensionality), we integrate unsupervised gene selection into the IGP model through two-way ordering of gene expression data, so that we can eliminate irrelevant or redundant genes in the data and obtain an improved clustering result.

The remainder of the paper is organized as follows. Section II gives a brief overview of graph theory and some related work on graph partitioning. The details of IGP algorithm with gene selection via two-way ordering applied to sample clustering in gene expression data are discussed in Section III. Section IV presents our experimental setup and results. Finally, we conclude in Section V.

II. Related Work

In this section, we introduce some relevant terminology on graph theory and review related work in graph partitioning for clustering. The notations used in this paper are as follows: lowercase-bold letters such as $\mathbf{x}$ denote column vector, capital-bold letters such as $\mathbf{A}$ denote a matrix, and capital-italic letters such as V denote graph vertex and edge sets.

A. Graph Partitioning

A **graph** is a pair $G=(V, E)$ with vertices (nodes) $v \in V$ and edges $e \in E \subseteq V \times V$. An edge, e, spanning two vertices, v_i and v_j, is denoted by e_{ij}. Let $|V|$ be the number of vertices and $|E|$ be the number of edges in the graph, where $|.|$ denotes cardinality. A **weighted graph** has a value assigned to each edge called a weight. The weight of an edge e_{ij} is denoted by $w(e_{ij})$. If there is no edge between node i and j, then $w(e_{ij})$ is zero, otherwise the weight is typically nonnegative. We

Yanhua Chen, Ming Dong, and Manjeet Rege are with Machine Vision and Pattern Recognition Laboratory, Department of Computer Science, Wayne State University, Detroit, MI 48202, USA (phone: 313-577-0725; Fax: 313-577-6868; email: aw6272, mdong, rege@wayne.edu).

1-4244-1380-X/07/$25.00 ©2007 IEEE

begin with definitions of a few graph terminologies that play an important role in the paper.

DEFINITION 1. The **degree** of a vertex v_i, denoted by d_i, is,

$$d_i = \sum_{e_{ij}} w(e_{ij}) \; \forall e_{ij} \in E. \tag{1}$$

DEFINITION 2. The **adjacency** or **affinity matrix A** of the graph is defined by,

$$\mathbf{A}_{ij} = \begin{cases} w(e_{ij}) & \text{if } e_{ij} \text{ exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

DEFINITION 3. The **degree matrix D** of the graph is a diagonal matrix as,

$$\mathbf{D}_{ij} = \begin{cases} d_i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

DEFINITION 4. The $|V| \times |V|$ matrix, **Laplacian matrix L**, of a graph is defined as,

$$\mathbf{L}_{v_i v_j} = \begin{cases} d_i & \text{if } i = j, \\ -w(e_{ij}) & \text{if } e_{ij} \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that the notation $\mathbf{L}_{v_i v_j}$ is used to indicate that the matrix **L** is being indexed by vertices v_i and v_j. Some of the properies of **L** are discussed in [29],

1) $\mathbf{L} = \mathbf{D} - \mathbf{A}$.
2) **L** is a symmetric positive semi-definite matrix. Thus, all eigenvalues of **L** are real and non-negative, and **L** has a full set of n real and orthogonal eigenvectors.
3) Let vector $\mathbf{e} = [1, 1, ...1]^T$, then $\mathbf{Le} = \mathbf{0}$. Thus 0 is an eigenvalue of **L** and **e** is the corresponding eigenvector.

DEFINITION 5. Given a partitioning of the vertex set V into two subsets V_1 and V_2, the **graph cut** is defined as,

$$cut(V_1, V_2) = \sum_{i \forall V_1, j \forall V_2} \mathbf{A}_{ij}. \tag{5}$$

DEFINITION 6. The definition of a cut can be extended to **k-partitioning of the graph**, which is formally defined as,

$$cut(V_1, V_2, ..., V_k) = \sum_{i<j} cut(V_i, V_j). \tag{6}$$

It is well known that the graph partitioning problem is NP-complete [18]. A wide variety of graph partitioning algorithms have been explored in recent years in a number of fields such as parallel processing [19], VLSI circuit design[20], image segmentation[14], [11], and data clustering [13] or co-clustering [21], [22], [23]. Of the various graph partitioning algorithms proposed such as geometric partitioning [24], inertial partitioning or coordinate partitioning [25], spectral partitioning [30] has been the most popular and widely applied.

B. *Spectral Graph Partitioning*

SGP is based on the early works of Donath and Hoffman [26] who proposed using the eigenvectors of adjacency matrices of the graphs to find partitions. Subsequently, Fiedler [27], [28] proposed an effective heuristic solution which associated λ_2 with connectivity of graph and suggested partitioning graph by dividing vertices according to the corresponding eigenvector $\mathbf{u}_2 = \{u_1, u_2, ..., u_k\}$. λ_2 and $\mathbf{u}_2$ are commonly referred to as the Fiedler value and Fielder vector, respectively of a graph. A splitting value s partitions the vertices of graph into two sets, $u_i > s$ and $u_i \leq s$.

Recently, Shi and Malik applied the SGP model to the problem of image segmentation [14]. The objective function with normalized cuts used in this work is defined as,

$$\min_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{\mathbf{x}^T \mathbf{D} \mathbf{x}}, \text{subject to } \mathbf{x}^T \mathbf{D} \mathbf{e} = 0, \tag{7}$$

where an indicator vector or partition vector, **x**, takes a binary value at each node. If relaxing **x** from discrete to continuous, it can be shown that the solution to equation (7) is the eigenvector corresponding to the second smallest eigenvalue of the generalized eigen-problem [16], [30],

$$\mathbf{Lx} = \lambda \mathbf{Dx}. \tag{8}$$

Subsequently, partitions can be obtained by running a clustering algorithm such as k-means on **x**.

III. Clustering Using Isoperimetric Graph Partitioning Model

In this section, we first briefly introduce the IGP model in Section A, followed by a discussion of our proposed approach to clustering samples of gene expression data in Sections B and C. Section D provides a summary of our algorithm.

A. *Isoperimetric Graph Partitioning Model for Gene Analysis*

The IGP model is developed based on the optimization of an isoperimetric constant. The **isoperimetric constant** h of a manifold is defined as [9],

$$h = \inf_S \frac{|\partial S|}{Vol_S}, \tag{9}$$

where S is a region in the manifold, Vol_S denotes the volume of region S, $|\partial S|$ is the area of the boundary of region S, and h is the infimum of the ratio over all possible S. For a compact manifold, $Vol_S \leq \frac{1}{2} Vol_{Total}$ and, for a noncompact manifold, $Vol_S < \infty$ [10].

Here we assume the graph is a compact manifold which has finite nodes. For clustering samples of gene expression data, we can use a vertex (node) to represent tissue sample in a graph and the similarity between two tissue samples v_i and v_j becomes the weight w_{ij} on the edge between them. Then we use Pearson Correlation $c(i,j)$ to define the pairwise similarity of two samples. Thus, the matrix $w(i,j) = c(i,j)$ defines a similarity weighted graph. We wish to partition the vertices (samples) into two subgraphs S and $\bar{S}$ with minimum h so that the isoperimetric constant, h_G, becomes

$$h_G = \inf_S \frac{|\partial S|}{Vol_S}, \tag{10}$$

where $S \subset V$ and $Vol_S \leq \frac{1}{2} Vol_V$.

The numerator, i.e., the boundary of a set S, is defined as $\partial S = \{w_{ij} | v_i \in S, v_j \in \bar{S}\}$, where $\bar{S}$ denotes the set complement and

$$|\partial S| = \sum_{e_{ij} \in \partial S} w_(e_{ij}). \tag{11}$$

The denominator determines a notion of volume for a graph which can be defined as

$$Vol_S = \sum_i d_i \; \forall v_i \in S. \tag{12}$$

Thus, for a given S, we define the ratio of its boundary to its volume as the **isoperimetric ratio**, denoted by $h(S)$. The **isoperimetric sets** for a graph, G, are any sets S and $\bar{S}$ for which $h(S)$ is minimized (note that the isoperimetric sets may not be unique for a given graph). Then the specification of a set and its complement satisfying the constraint in Equation (10) may be considered as a partition of the graph G, which divides the samples into two clusters. In our work, a good partition of samples is considered to the one with a low isoperimetric ratio. Our goal is to maximize Vol_S while minimizing $|\partial S|$. Unfortunately, finding isoperimetric sets is an NP-hard problem [10]. Grady and Schwartz present a heuristic algorithm for finding optimal partition of a graph with low isoperimetric ratio that runs in low-order polynomial time, and apply it on image segmentation and data clustering [11], [31], [32]. More recently, co-clustering framework based on isoperimetric graph partitioning has been proposed in [23]. In this paper, we propose to apply the IGP model to clustering samples in gene expression data.

B. Algorithm Derivation

If we plug Laplacian matrix $\mathbf{L}$ and indicator vector $\mathbf{x}$ into equation (11) and (12), we obtain,

$$|\partial S| = \mathbf{x}^T \mathbf{L} \mathbf{x}, \tag{13}$$

and

$$Vol_S = \mathbf{x}^T \mathbf{d} = k, \tag{14}$$

where $0 < k < \frac{1}{2} \mathbf{e}^T \mathbf{d}$ is an arbitrary constant. The choice of k becomes irrelevant to the final formulation. Thus, the isoperimetric constant in Equation (11), G, may be rewritten in terms of the indicator vector as,

$$h_G = \min_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{\mathbf{x}^T \mathbf{d}}, \tag{15}$$

subject to the constraints of Equation (13) and (14). Given an indicator vector, $\mathbf{x}$, $h(\mathbf{x})$ represents the isoperimetric ratio associated with the partition specified by $\mathbf{x}$. For gene expression analysis, $\mathbf{x}$ can be considered as a partition of samples in gene expression data. For example, S could indicate cancer sample set, while $\bar{S}$ indicate non-cancer sample set.

In order to solve Equation (15), we introduce a Lagrange multiplier Λ and relaxation of the binary definition of $\mathbf{x}$ to take nonnegative real values by minimizing the following cost function,

$$Q(\mathbf{x}) = \mathbf{x}^T \mathbf{L} \mathbf{x} - \Lambda(\mathbf{x}^T \mathbf{d} - k). \tag{16}$$

Since $\mathbf{L}$ is positive semi-definite and $\mathbf{x}^T$ is nonnegative, $Q(\mathbf{x})$ will be at a minimum for any critical point. Differentiating $Q(\mathbf{x})$ with respect to $\mathbf{x}$ and setting to a minimum yields,

$$2\mathbf{L}\mathbf{x} = \Lambda \mathbf{d}. \tag{17}$$

Thus, the problem of finding the $\mathbf{x}$ that minimizes $Q(\mathbf{x})$ reduces to solving a linear system. This linear system solution is desirable and efficient instead of solving an eigenvector problem in Equation (8) of the SGP model. Also, the scalar multiplier 2 and the scalar Λ of Equation (17) are dropped since only the relative values of the solution are significant.

Unfortunately, the matrix $\mathbf{L}$ is singular: all rows and columns sum to zero. So, finding a unique solution to Equation (17) requires an additional constraint as suggested in [11]. Assume that we remove the node of largest degree from $\mathbf{L}$, and the corresponding rows from $\mathbf{x}$ and $\mathbf{d}$, we have,

$$\mathbf{L}_0 \mathbf{x}_0 = \mathbf{d}_0 \tag{18}$$

which is a nonsingular system of equations. Solving Equation (18) for $\mathbf{x}_0$ yields a real-valued solution that may be converted into a partition of samples. For example, if $\mathbf{x}_0$ is greater than 0, it is labeled as one type of samples, otherwise, as another type.

C. Feature Selection via Two-way Ordering of Gene Expression

When directly used for sample clustering, IGP does not provide a satisfactory result. This is mainly due to the high dimensionality of feature spaces (thousands of genes) and the fact that many genes are irrelevant or redundant. Therefore, in order to obtain an improved clustering performance, there is a need to identify genes that significantly contribute to the partition of the samples.

There are two general types of feature selection methods: supervised and unsupervised. Supervised feature selection can be categorized into filters and wrappers [12], both depend on known class information. However, supervised approaches are not suitable for clustering of samples in gene expression data because certain set of genes may correspond to new phenotypes or subtypes, i.e., the class labels are not always available. In this case, unsupervised feature selection becomes a more reasonable method to find the relevant genes. Certainly, we can select the most relevant features from all genes using prior knowledge of cluster structure if available. However, in most cases, it is hard to know any prior information before clustering, so we have to design algorithms to select the most relevant features without any prior knowledge. To this end, [6] illustrated an iterative feature filtering approach to identify the relevant genes, but it may be stuck in a local fixed point in the high dimensional parameter space. A better approach is to use two-way ordering of gene expression profiles to discard irrelevant or non-discriminate features [8].

The objective of node ordering is to ensure that more adjacent nodes are more similar while further away nodes are less similar. Original gene expression data can be expressed as a weighted bipartite graph. Each gene is g-type node and each tissue sample is a s-type node. Let $\mathbf{B}$ denote $m \times n$ gene-sample original association matrix, $\mathbf{g} = (g_1, ..., g_m)^T$ an index permutation of genes, and $\mathbf{s} = (s_1, ..., s_n)^T$ an index permutation of samples. The indicator vector is $\mathbf{p} = (\mathbf{g}, \mathbf{s})^T$, which orders both type of nodes in a high dimensionality. The symmetric weighted adjacency matrix $\mathbf{W}$ for the bipartite graph is represented by,

$$\mathbf{W} = \left(\begin{array}{cc} 0 & \mathbf{B} \\ \mathbf{B}^T & 0 \end{array} \right). \tag{19}$$

The degree matrix is denoted as $\mathbf{D} = diag(\mathbf{D}_g, \mathbf{D}_s)$, where $\mathbf{D}_g = diag(\mathbf{B}\mathbf{e}_n)$, $\mathbf{D}_s = diag(\mathbf{B}\mathbf{e}_m)$, and $\mathbf{e} = (1, ..., 1)^T$ with appropriate size. Here $\mathbf{g}$ and $\mathbf{s}$ are represented as,

$$\mathbf{z} = \left(\begin{array}{c} \mathbf{u} \\ \mathbf{v} \end{array} \right) = \mathbf{D}^{1/2}\mathbf{p} = \left(\begin{array}{c} \mathbf{D}_g^{1/2}\mathbf{g} \\ \mathbf{D}_s^{1/2}\mathbf{s} \end{array} \right). \tag{20}$$

Plugging $\mathbf{W}$, $\mathbf{D}$, $\mathbf{z}$ into $(\mathbf{D} - \mathbf{W})\mathbf{p} = \lambda\mathbf{D}\mathbf{p}$, we obtain,

$$\hat{\mathbf{B}}\mathbf{v} = \zeta\mathbf{u}, \hat{\mathbf{B}}^T\mathbf{u} = \zeta\mathbf{v}, \hat{\mathbf{B}} = \mathbf{D}_g^{-1/2}\mathbf{B}\mathbf{D}_s^{-1/2}, \tag{21}$$

where $\zeta = 1 - \lambda$. The solutions to the above equations are the singular value decomposition (SVD) of $\hat{\mathbf{B}}$, $\hat{\mathbf{B}} = \sum_k \mathbf{u}_k \zeta_k \mathbf{v}_k^T$. Therefore, we compute the second principle components $\mathbf{u}_2$, $\mathbf{v}_2$, and obtain $\mathbf{g}_2 = \mathbf{D}_g^{-1/2}\mathbf{u}_2$ and $\mathbf{s}_2 = \mathbf{D}_s^{-1/2}\mathbf{v}_2$. We sort $\mathbf{g}_2$ and $\mathbf{s}_2$ in an increasing order and use the induced permutation to order genes and tissue samples, respectively. Through this re-ordering, nondiscriminating genes will locate near the middle and can be discarded.

Based on the selective genes via two-way ordering, we can cluster samples using IGP in a feature space with reduced dimensionality. Better clustering performance is expected.

D. Summary of the Algorithm

The complete IGP algorithm for sample clustering integrated with unsupervised feature selection via two-way ordering can be described in the following steps,

1) Given bipartite graph weighted matrix $\mathbf{W}$ from original gene expression association matrix $\mathbf{B}$, form $\hat{\mathbf{B}} = \mathbf{D}_g^{-1/2}\mathbf{B}\mathbf{D}_s^{-1/2}$, where $D_g(i,i) = \sum_j \mathbf{W}_{ij}$, $D_s(i,i) = \sum_i \mathbf{W}_{ij}$.
2) Compute the second principle components $\mathbf{u}_2$ and $\mathbf{v}_2$ of $\hat{\mathbf{B}}$ according to equation (21), and get index permutation for genes $\mathbf{g}_2 = \mathbf{D}_g^{-1/2}\mathbf{u}_2$ and index permutation for samples $\mathbf{s}_2 = \mathbf{D}_s^{-1/2}\mathbf{v}_2$.
3) Sort $\mathbf{g}_2$ and $\mathbf{s}_2$ to increasing order to reorder genes and samples to get reordering matrix $\mathbf{B}'$.
4) Discard genes in the middle of matrix $\mathbf{B}'$, and calculate new weight $w(i,j)$ based on Pearson Correlation to form a new weighted graph $\mathbf{A}$ from samples.
5) Build the Laplacian matrix $\mathbf{L}$ and $\mathbf{d}$ using equations (4) and (1).
6) Remove the node of largest degree from $\mathbf{L}$ and the corresponding rows from $\mathbf{x}$ and $\mathbf{d}$ to determine $\mathbf{L}_0$, $\mathbf{x}_0$ and $\mathbf{d}_0$.
7) Solve equation (18) for indicator vector $\mathbf{x}_0$ which gives partitions of samples corresponding to the lowest isoperimetric ratio.

IV. Experiments And Results

In this section, we report the results of using IGP to cluster several well-studied gene expression data. Our data comes from Kent Ridge Bio-medical Data Set Repository [1]. We choose two datasets. The first one is related to colon tumor sample analysis; while the second one is about ALL-AML leukemia subtypes analysis. Clustering results are specified and evaluated by a contingency table $T = (t_{ij})$ and the simple Q-accuracy defined as $\sum_i t_{ii}/N$ [12], which is the sum of the diagonal elements divided by the total number of samples. In addition, we also report the isoperimetric ratio [32], which represents the quality of clustering from another perspective: smaller the isoperimetric ratio, better is the clustering.

A. Analysis of Colon Cancer Data

In our first experiment, we apply the IGP model to the expression profiles of colon tumor tissues [4]. It contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as “negative” by experts) and 22 normal (labeled as “positive” by experts) biopsies are from healthy parts of the colons of the same patients. 2000 out of around 6500 genes were selected based on the confidence in the measured expression levels.

We compare IGP with the SGP model and report our results in Table I. As shown in the first row of Table I, Q-accuracy increases from 0.5806 (SGP) to 0.6290 (IGP) when all 2000 genes are used. After we apply unsupervised feature selection via two-way ordering, the Q-accuracy increases significantly given the selected number of genes, $m = 800, 400, 200$. Moreover, the isoperimetric ratio of IGP is only half of that of SGP under all selective number of genes. Figure 1 shows that IGP constantly outperforms SGP in every level of our comparison.

TABLE I

THE COMPARISON OF Q-ACCURACY AND ISOPERIMETRIC RATIO OF IGP AND SGP FOR CLUSTERING COLON CANCER / NORMAL SAMPLES BASED ON SELECTIVE GENES THROUGH TWO-WAY ORDERING.

	SGP		*IGP*	
m genes	*Q-accuracy*	*Iso. ratio*	*Q-accuracy*	*Iso. ratio*
2000	0.5806	0.9892	0.6290	0.5156
800	0.5968	0.9563	0.7258	0.4984
400	0.7258	0.9132	0.7419	0.4897
200	0.8065	0.8669	0.8226	0.4852

[1] http://sdmc.lit.org.sg/GEDatasets/Datasets.html

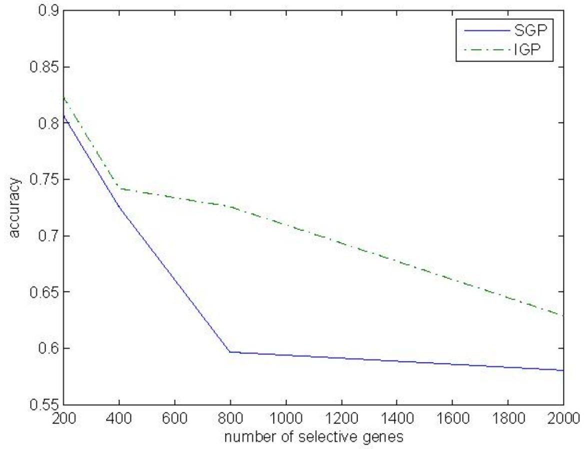

Fig. 1. Comparison of clustering accuracy between IGP (dot line) and SGP (dark line) on colon cancer / normal samples.

B. Analysis of Leukemia Subtypes

In the second expriment, we apply IGP to the leukemia subtypes dataset of Golub et al. [3]. We use the training dataset which consists of 38 bone marrow samples over 7129 probes from 6817 human genes. The target classes here are the two phenotypes of the cancer: 27 from acute lymphoblastic leukemia (ALL) and 11 from acute myeloid leukemia (AML).

Since the intensity readings of this expression data have many negative values, it is necessary to transfer negative values to non-negative values in order to apply two-way ordering of gene expression in a bipartite graph. Specifically, we increase all leukemia data by 1000 and normalize all values in the range between 0 to 5000: the value is set to 0 if below zero, and set to 5000 if above 5000. So, we have,

$$b^*_{ij} = b_{ij} + 1000; 0 \leq b^*_{ij} \leq 5000. \qquad (22)$$

The purpose of this kind of pre-processing is mainly to remove the outliers. Note that shifting by a constant will not change Pearson Correlation and other linear properties. After pre-processing, we should check if elements in a row of b^*_{ij} are all zeros. If so, the row will be deleted from b^*_{ij}. This operation guarantees no degenerated cases when calculating SVD in two-way ordering. The final number of genes is 7122 after the deletion.

We compare IGP with the SGP model on leukemia data and report our results in Table II. As shown in the first row of Table II, Q-accuracy increases from 0.5263 (SGP) to 0.6842 (IGP) when all 7122 genes are used. After applying unsupervised feature selection via two-way ordering, the Q-accuracy increases significantly on selected number of genes, $m = 3000, 2000, 1000$. In addition, the isoperimetric ratio of IGP is also only about half of that of SGP under all selective number of genes. Notice that IGP may create lower Q-accuracy when the selected number of genes continuously decreases to $m = 400, 200$ even though its Q-accuracy is still higher than that of SGP. Figure 2 shows that IGP constantly outperforms SGP in every level of our comparison. It also shows that the number of features selected by two-way ordering should be reasonable. If the number of features is too large (more noise or redundant features) or too small (some important features are lost), we may get sub-optimal clustering results.

In summary, our results indicate that the leukemia phenotype structure appears to be reasonably well-separated because most of the time the clusters obtained by the IGP model agrees with the pre-defined class labels provided by biological experts. However, comparing Table II to Table I, it is clear that the highest Q-accuracy of clustering for leukemia data is lower than that for colon data: 0.7895 vs. 0.8226. This results suggest that the clusters obtained from IGP with two-way ordering may contain some unknown or unavailable classes. Indeed, [3] argues that ALL class should be split into T-lineage ALL and B-lineage ALL classes. Thus, there should be 3 clusters, not 2, in this data set. The additional implicit cluster may explain why the quality of clustering of leukemia subtypes data is not as good as that of colon cancer data, which warrants further investigation in the future.

TABLE II

The Comparison of Q-accuracy and isoperimetric ratio of IGP and SGP for clustering ALL/AML leukemia subtypes based on selective genes through two-way ordering.

	SGP		*IGP*	
m genes	*Q-accuracy*	*Iso. ratio*	*Q-accuracy*	*Iso. ratio*
7122	0.5263	1.0165	0.6842	0.5314
3000	0.5263	0.9995	0.7895	0.5144
2000	0.5526	0.9905	0.7895	0.5106
1000	0.6053	0.9730	0.7895	0.5023
400	0.7105	0.9423	0.7368	0.5040
200	0.7105	0.9027	0.7368	0.4696

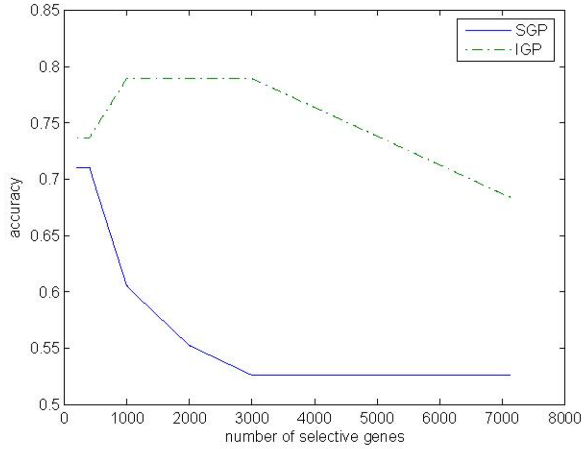

Fig. 2. Comparison of clustering accuracy between IGP (dot line) and SGP (dark line) on ALL/AML leukemia subtypes samples.

V. Conclusions

In this paper, we propose to apply a novel graph partitioning model - Isoperimetric Graph Partitioning (IGP), to group biological samples from gene expression data. In addition, we incorporate unsupervised gene selection into IGP through two-way ordering of gene expression data, so that we can eliminate irrelevant or redundant genes in the data and obtain an improved clustering results. We evaluate our approach on several well-known problems involving gene expression profiles of colon cancer and leukemia subtypes. Our experimental results show that IGP constantly outperforms SGP and produces a result with higher accuracies and lower isoperimetric ratios. Furthermore, the accuracies of clustering are improved significantly when IGP is integrated with the unsupervised feature selection.

References

[1] Z. Szallasi and R. Somogyi, "Genetic Network Analysis- the Millennium Opening Version," in *Proc. Pacific Symposium of Biocomputing Tutorial*, 2001.

[2] M. B. Eisen and P. T. Spellman and P. O. Brown and D. Botstein, "Cluster Analysis and Display of Genome-wide Expression Patterns," in *Proc. Nat'l Acad Sci USA*, vol. 95, pp. 14863-14868, 1998.

[3] T. R. Golub and D. K. Slonim and P. Tamayo and et al,"Molecular Classification of Cancer: Class Discovery and Class Predication by Gene Expression Monitoring," *Science* , vol. 286, pp. 531-537, 1999.

[4] U. Alon and et al, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," in *Proc. Nat'l Acad Sci USA*, vol. 96, pp. 6745-6750, 1999.

[5] R. Sharan and R. Shamir, "CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis," in *Proc. ISMB 2000*,pp. 307-316 , 2000.

[6] E. P. Xing and R. M. Karp, "CLIFF: Clustering of High-dimensional Microaaray Data via Iterative Feature Filtering Using Normalized Cuts," *Bioinformatics*,vol. 17, pp. S306-S315 , 2001.

[7] Chris H. Q. Ding, "Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering," in *Proc. RECOME 2002*,pp. 127-136 , 2002.

[8] Chris H. Q. Ding, "Unsupervised Feature Selection via Two-way Ordering in Gene Expression Analysis," *Bioinformatics*,pp. 1259-1266 , 2003.

[9] J. Cheeger, "A Lower Bound for the Smallest Eigenvalue of the Laplacian," in *Problems in Analysis*,Princeton, N.J.: Princeton Univ. Press, R.C.Gunning, ed. pp. 195-199 , 1970.

[10] B. Mohar, "Isoperimetric Numbers of Graphs," *Journal of Combinatorial Theory, Series B*,vol. 47, pp. 274-291 , 1989.

[11] Leo Grady and Eric L. Schwartz, "Isoperimetric Graph Partitioning for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol. 28, pp. 469-475 , 2006.

[12] G. H. John and R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," *Int'l Conf. Machine Learning*,pp. 121-129 , 1994.

[13] C. Ding and X. He and H. Zha and M. Gu and H. Simon, "A Minmax Cut Algorithm for Graph Partitioning and Data Clustering," in *Proc. IEEE Int'l Conf. Data Mining*, pp. 107-114, 2001.

[14] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888-905, Aug 2000.

[15] S. Sarkar and P. Soundararajan, "Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata," *IEEE Trans. Pattern Analysis and Machine Intelligence* , vol. 11, pp. 504-525, May 2000.

[16] G. Golub and C. Van Loan, "Matrix Computations," Johns Hopkins Univ. Press,third ed. 1996.

[17] Y. F. Hu and R. J. Blak, "Numerical Experiences with Partitioning of Unstructured Meshes," *Parallel Computing*, vol. 20, pp. 815-829, 1994.

[18] M.R. Garey and D. S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness," W. H. Freeman & Company, 1979.

[19] H.D. Simon, "Partitioning of Unstructured Problems for Parallel Processing," *Computing Systems in Engineering*, vol. 2, pp. 135-148, 1991.

[20] C. J. Alpert and A. B. Kahng, "Recent Directions in Netlist Partitioning: A survey," *Integration, the VLSI Journal*, vol. 19, no. 12, pp. 1-81, 1995.

[21] H. Zha, X.He, C. H.Q. Ding, H. Simon and M. Gu, "Bipartite Graph Partitioning and Data Clustering," *Proc. the Tenth International Conference on Information and Knowledge Management*, 2001.

[22] Inderjit Dhillon, "Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *Proc. the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[23] Manjeet Rege, Ming Dong and Farshad Fotouhi, "Co-clustering Documents and Words Using Bipartite Isoperimetric Graph Partitioning ," *Proc. the Sixth International Conference on Data Mining*, vol. 20, pp. 532-541, 2006.

[24] J.R. Gilbert, G.L.Miller and S.H.Teng, "Geometric Mesh Partitioing: Implementation and Experiments," *SIAM Journal on Scientific Computing*, vol. 19,no. 6, pp. 2091-2110, 1998.

[25] B. Hendrickson and R. Leland, "The Chaco User's Guide," *Tech. Rep. SAND95-2344*,Snadia National Laboratories,Albuquerque, NM, 1995.

[26] W. E. Donath and A. J. Hoffman, "Lower Bounds for the Partitioning of Graphs," *IBM Journal of Research and Development*, vol. 17, pp. 420-425, 1973.

[27] M.Fiedler, "Eigenvectors of Acyclic Matrices," *Cxechoslovak Mathematics Journal*, vol. 25, pp. 607-618, 1975.

[28] M.Fiedler, "A Property of Eigenvectors of Nonnegative Symmetric Matrics and Its Application to Graph Theory," *Czechoslovak Mathematics Journal*, vol. 25, pp. 619-633, 1975.

[29] M.Fiedler, "Special Metrices and Their Applications in Numerical Mathematics," Martinus Nijhoff Publishers, 1986.

[30] F.R.K.Chung, "Spectral Graph Theory," *American Mathematical Society*, 1997.

[31] Leo Grady and Eric L. Schwartz, "Isoperimetric Partitioning: a New Algorithm for Graph Partitioning," *SIAM Journal on Scientific Computing*, vol. 27, no. 6, pp. 1844-1866, June 2006.

[32] Leo Grady and Eric L. Schwartz, "Isoperimetric Graph Partitioning for Data Clustering and Image Segmentation, " Technical Report CNS-TR-03-015, 2003.