Un-Normalized Graph P-Laplacian Semi-Supervised Learning Method Applied to Cancer Classification Problem

Loc Hoang Tran
University of Minnesota/Computer Science Department, Minneapolis, USA
Email: tran0398@umn.edu

Linh Hoang Tran
Portland State University/ECE Department, Portland, USA
Email: linht@pdx.edu

***Abstract*—A successful classification of different tumor types is essential for successful treatment of cancer. However, most prior cancer classification methods are clinical-based and have inadequate diagnostic ability. Cancer classification using gene expression data is very important in cancer diagnosis and drug discovery. The introduction of DNA microarray techniques has made simultaneous monitoring of thousands of gene expression probable. With this abundance of gene expression data nowadays, the researchers have the opportunity to do cancer classification using gene expression data. In recent years, a lot of machine learning methods have been proposed to do cancer classification using gene expression data such as clustering-based methods, k-nearest neighbor method, artificial neural network method, and support vector machine method, to name a few. In this paper, we present the un-normalized graph p-Laplacian semi-supervised learning methods. These methods will be applied to the patient-patient network constructed from the gene expression data to predict the tumor types of all patients in the network. These methods are based on the assumption that the labels of two adjacent patients in the network are likely to be the same. The experiments show that that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art network-based method (the un-normalized graph Laplacian based semi-supervised learning method) but often lead to better classification accuracy performance measures.**

***Index Terms*—graph p-Laplacian, genes, cancer classification, semi-supervised learning**

I. Introduction

Cancer contains cells of the body which multiply in an uncontrolled fashion. Cancer cells may develop the capability to leave their tissue of origin and stay alive in other tissue types, causing metastases. The diagnosis and classification of cancer is an essential step for the treatment of the disease. Conventional methods used in the clinics are based on clinical, pathological, and molecular factors. The clinical factors include the age of the patient and the stage of tumor describing the extension of the tumor locally or at the distance from the major position. Pathological factors include the size of the tumor, the lymph node status, and the grading reporting the morphology and proliferative size of the main tumor. Molecular markers are determined regularly by immunohistochemistry methods. Regrettably, the conventional methods cannot precisely define prognosis and predict response to treatment. Moreover, the human expertise is required. In the other words, the well-trained and experienced clinicians and pathologists are required. This is not a simple demand. There also exist the increasing number of tests that the clinicians are demanded to perform. Hence the automated system could provide the clinics with the second opinion or tools for training pathologists.

The introduction of high throughput bimolecular measurements, for e.g. gene expression data, allows the close look at the molecular mechanisms of the diseases. Cancer classification using gene expression data is identified to contain the key for addressing the essential problems involving in cancer diagnosis and drug discovery. The introduction of DNA microarray techniques has made simultaneous monitoring of thousands of gene expression probable. With this abundance of gene expression data nowadays, the researchers have the opportunity to do cancer classification using gene expression data. In recent years, a lot of machine learning methods have been proposed to do cancer classification using gene expression data such as clustering-based methods [1], [2], k-nearest neighbor method [3], artificial neural network method [4], and support vector machine method [5], to name a few. However, there still exist a lot of issues needed to be identified and understood. The first issue is that the number of samples (i.e. patients) is small but the number of genes is large. The second issue is the presence of noise in the gene expression datasets. Finally, the third issue is most of genes in the gene expression datasets are not cancer related.

Manuscript received November 14, 2013; revised February 1, 2014.

doi: 10.12720/joace.3.1.23-28

First studies with gene expression data used clustering-based algorithms (for e.g. k-mean algorithm) to cluster patients into groups that shared common biological features. This is an un-supervised learning method of analyzing the data since no prior knowledge about the patients is used. No class labels assigning the patients to similar group are used. However, the classes are inferred from the results of the clustering algorithms. Normally, these un-supervised learning methods achieve the lowest accuracy performance measures. The other learning methods such as k-nearest neighbor method and support vector machine method are all supervised learning method. In these supervised machine learning methods, a graph (i.e. kernel) which is the natural model of relationship between patients are employed. In this model, the nodes represent patients. The k-nearest neighbor method labels the patient with the class label that occurs frequently in the patient's adjacent nodes in the patient-patient network. Hence k-nearest neighbor method does not utilize the full topology of the patient-patient network. However, the Support Vector Machine method utilizes the full topology of the patient-patient network. The Support Vector Machine method achieves the highest accuracy performance measures in most of classification tasks and is considered the current state of the art machine learning method. However, the time and space complexity of the Support Vector Machine method are high. Since the graph (i.e. kernel) used in Support Vector Machine method is fully-connected, SVM does not rely on the assumption that the labels of two adjacent patients in graph are likely to be the same. The Artificial Neural Networks method, the un-normalized, symmetric normalized and random walk graph Laplacian based semi-supervised learning methods are all based on this assumption.

In the last decade, the symmetric normalized graph Laplacian [6], [9], random walk graph Laplacian [7], [9], and the un-normalized graph Laplacian [8], [9] based semi-supervised learning methods have successfully been applied to some specific classification tasks such as digit recognition, text classification, and protein function prediction. However, to the best of my knowledge, the un-normalized, symmetric normalized, and random walk graph Laplacian based semi-supervised learning methods have not yet been applied to cancer classification problems. In this paper, the un-normalized graph Laplacian based semi-supervised learning method will be served as the baseline method (i.e. the un-normalized graph 2-Laplacian operator) for cancer classification problem.

In [10], [11], the symmetric normalized graph p-Laplacian based semi-supervised learning method has been developed but has not been applied to any practical applications. To the best of my knowledge, the un-normalized graph p-Laplacian based semi-supervised learning method has been developed in [12] and has been applied successfully to protein function prediction problem; however, it has not been applied to the cancer classification problem. In this paper, we will apply the un-normalized graph p-Laplacian based semi-supervised learning method to cancer classification problem. This method is worth investigated because of its difficult nature and its close connection to partial differential equation on graph field. Specifically, in this paper, the un-normalized graph p-Laplacian based semi-supervised learning method will be developed based on the un-normalized graph p-Laplacian operator definition such as the curvature operator of graph (i.e. the un-normalized graph 1-Laplacian operator). Please note that the un-normalized graph p-Laplacian based semi-supervised learning method is developed based on the assumption that the labels of two patients in the patient-patient network are likely to be the same [8,9]. The main point of the un-normalized graph p-Laplacian based semi-supervised learning method is to let every node of the graph iteratively propagates its label information to its adjacent nodes and the process is repeated until convergence [8], [9], [12].

We will organize the paper as follows: Section 2 will introduce the preliminary notations and definitions used in this paper. Section 3 will introduce the definition of the gradient and divergence operators of graphs. Section 4 will introduce the definition of Laplace operator of graphs and its properties. Section 5 will introduce the definition of the curvature operator of graphs and its properties. Section 6 will introduce the definition of the p-Laplace operator of graphs and its properties. Section 7 will show how to derive the algorithm of the un-normalized graph p-Laplacian based semi-supervised learning method from regularization framework. In section 8, we will compare the accuracy performance measures of the un-normalized graph Laplacian based semi-supervised learning algorithm (i.e. the current state of art network-based method applied to bioinformatics problem) and the un-normalized graph p-Laplacian based semi-supervised learning algorithms. Section 9 will conclude this paper and the future direction of researches of other practical applications in bioinformatics utilizing discrete operator of graph will be discussed.

II. Preliminary Notations and Definitions

Given a graph $G=(V,E,W)$ where V is a set of vertices with $|V| = n$, $E \subseteq V * V$ is a set of edges and W is a $n * n$ similarity matrix with elements $w_{ij} > 0$ $(1 \leq i, j \leq n)$.

Also, please note that $w_{ij} = w_{ji}$.

The degree function $d: V \to R^+$ is

$$d_i = \sum_{j \sim i} w_{ij}, \tag{1}$$

where $j \sim i$ is the set of vertices adjacent with i.

Define

$$D = diag(d_1, d_2, \ldots, d_n)$$

The inner product on the function space R^V is

$$< f, g >_V = \sum_{i \in V} f_i g_i \tag{2}$$

Also define an inner product on the space of functions R^E on the edges

$$< F, G >_E = \sum_{(i,j) \in E} F_{ij} G_{ij} \tag{3}$$

Here let $H(V) = (R^V, <.,.>_V)$ and $H(E) = (R^E, <.,.>_E)$ be the Hilbert space real-valued functions defined on the vertices of the graph G and the Hilbert space of real-valued functions defined in the edges of G respectively.

III. Gradient and Divergence Operators

We define the gradient operator $d: H(V) \to H(E)$ to be

$$(df)_{ij} = \sqrt{w_{ij}}(f_j - f_i), \tag{4}$$

where $f: V \to R$ be a function of $H(V)$.

We define the divergence operator $div: H(E) \to H(V)$ to be

$$< df, F >_{H(E)} = < f, -divF >_{H(V)}, \tag{5}$$

where $f \in H(V), F \in H(E)$

Next, we need to prove that

$$(divF)_j = \sum_{i\sim j} \sqrt{w_{ij}}\,(F_{ji} - F_{ij})$$

Proof:

$$< df, F > = \sum_{(i,j)\in E} df_{ij} F_{ij}$$

$$= \sum_{(i,j)\in E} \sqrt{w_{ij}}\,(f_j - f_i) F_{ij}$$

$$= \sum_{(i,j)\in E} \sqrt{w_{ij}}\, f_j F_{ij} - \sum_{(i,j)\in E} \sqrt{w_{ij}}\, f_i F_{ij}$$

$$= \sum_{k\in V} \sum_{i\sim k} \sqrt{w_{ik}}\, f_k F_{ik} - \sum_{k\in V} \sum_{j\sim k} \sqrt{w_{kj}}\, f_k F_{kj}$$

$$= \sum_{k\in V} f_k \left(\sum_{i\sim k} \sqrt{w_{ik}}\, F_{ik} - \sum_{i\sim k} \sqrt{w_{ki}}\, F_{ki}\right)$$

$$= \sum_{k\in V} f_k \sum_{i\sim k} \sqrt{w_{ik}}\,(F_{ik} - F_{ki})$$

Thus, we have

$$(divF)_j = \sum_{i\sim j} \sqrt{w_{ij}}\,(F_{ji} - F_{ij}) \tag{6}$$

IV. Laplace Operator

We define the Laplace operator $\Delta: H(V) \to H(V)$ to be

$$\Delta f = -\frac{1}{2} div(df) \tag{7}$$

Next, we compute

$$(\Delta f)_j = \frac{1}{2} \sum_{i\sim j} \sqrt{w_{ij}}\,((df)_{ij} - (df)_{ji})$$

$$= \frac{1}{2} \sum_{i\sim j} \sqrt{w_{ij}}\,\left(\sqrt{w_{ij}}(f_j - f_i) - \sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \sum_{i\sim j} w_{ij}\,(f_j - f_i)$$

$$= \sum_{i\sim j} w_{ij} f_j - \sum_{i\sim j} w_{ij} f_i$$

$$= d_j f_j - \sum_{i\sim j} w_{ij} f_i$$

Thus, we have

$$(\Delta f)_j = d_j f_j - \sum_{i\sim j} w_{ij} f_i \tag{8}$$

The graph Laplacian is a linear operator. Furthermore, the graph Laplacian is self-adjoint and positive semi-definite.

Let $S_2(f) = < \Delta f, f >$, we have the following theorem 1

$$D_f S_2 = 2\Delta f \tag{9}$$

The proof of the above theorem can be found from [13,14].

V. Curvature Operator

We define the curvature operator $\kappa: H(V) \to H(V)$ to be

$$\kappa f = -\frac{1}{2} div\left(\frac{df}{||df||}\right) \tag{10}$$

Next, we compute

$$(\kappa f)_j = \frac{1}{2} \sum_{i\sim j} \sqrt{w_{ij}}\,\left(\left(\frac{df}{||df||}\right)_{ij} - \left(\frac{df}{||df||}\right)_{ji}\right)$$

$$= \frac{1}{2} \sum_{i\sim j} \sqrt{w_{ij}}\,\left(\frac{1}{\|d_i f\|}\sqrt{w_{ij}}(f_j - f_i) - \frac{1}{\|d_j f\|}\sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \frac{1}{2} \sum_{i\sim j} w_{ij}\,\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i)$$

Thus, we have

$$(\kappa f)_j = \frac{1}{2} \sum_{i\sim j} w_{ij}\,\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i) \tag{11}$$

From the above formula, we have

$$d_i f = ((df)_{ij} : j\sim i)^T \tag{12}$$

The local variation of *f* at *i* is defined to be

$$\|d_i f\| = \sqrt{\sum_{j\sim i} (df)_{ij}^2} = \sqrt{\sum_{j\sim i} w_{ij} (f_j - f_i)^2} \tag{13}$$

To avoid the zero denominators in (11), the local variation of *f* at *i* is defined to be

$$\|d_i f\| = \sqrt{\sum_{j\sim i} (df)_{ij}^2 + \epsilon}, \tag{14}$$

where $\epsilon = 10^{-10}$.

The graph curvature is a non-linear operator.

Let $S_1(f) = \sum_i \|d_i f\|$, we have the following theorem 2

$$D_f S_1 = \kappa f \tag{15}$$

The proof of the above theorem can be found from [13,14].

VI. p-Laplace Operator

We define the p-Laplace operator $\Delta_p: H(V) \to H(V)$ to be

$$\Delta_p f = -\frac{1}{2} div(\|df\|^{p-2} df) \tag{16}$$

Clearly, $\Delta_1 = \kappa$ and $\Delta_2 = \Delta$. Next, we compute

$$(\Delta_p f)_j = \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}\left(\|df\|^{p-2}df_{ij} - \|df\|^{p-2}df_{ji}\right)$$

$$= \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}\left(\|d_i f\|^{p-2}\sqrt{w_{ij}}(f_j - f_i) - \|d_j f\|^{p-2}\sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \frac{1}{2}\sum_{i\sim j} w_{ij}\left(\|d_i f\|^{p-2} + \|d_j f\|^{p-2}\right)(f_j - f_i)$$

Thus, we have

$$(\Delta_p f)_j = \frac{1}{2}\sum_{i\sim j} w_{ij}\left(\|d_i f\|^{p-2} + \|d_j f\|^{p-2}\right)(f_j - f_i) \quad (17)$$

Let $S_p(f) = \frac{1}{p}\sum_i \|d_i f\|^p$, we have the following theorem 3

$$D_f S_p = p\Delta_p f \quad (18)$$

VII. Discrete Regularization on Graphs and Protein Function Classification Problems

Given a patient-patient network *G=(V,E)*. *V* is the set of all patients in the network and *E* is the set of all possible interactions between these patients. Let *y* denote the initial function in *H(V)*. y_i can be defined as follows

$$y_i = \begin{cases} 1 \text{ if patient } i \text{ belongs to the class ALL} \\ -1 \text{ if patient } i \text{ belongs to the class AML} \\ 0 \text{ otherwise} \end{cases}$$

Our goal is to look for an estimated function *f* in *H(V)* such that *f* is not only smooth on *G* but also close enough to an initial function *y*. Then each patient *i* is classified as $sign(f_i)$. This concept can be formulated as the following optimization problem

$$argmin_{f\in H(V)}\{S_p(f) + \frac{\mu}{2}\|f - y\|^2\} \quad (19)$$

The first term in (19) is the smoothness term. The second term is the fitting term. A positive parameter μ captures the trade-off between these two competing terms.

A. 2-Smoothness

When *p=2*, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\frac{1}{2}\sum_i\|d_i f\|^2 + \frac{\mu}{2}\|f - y\|^2\} \quad (20)$$

By theorem 1, we have

Theorem 4: The solution of (20) satisfies

$$\Delta f + \mu(f - y) = 0 \quad (21)$$

Since Δ is a linear operator, the closed form solution of (21) is

$$f = \mu(\Delta + \mu I)^{-1} y, \quad (22)$$

where *I* is the identity operator and $\Delta = D - W$. (22) is the algorithm proposed by [8].

B. 1-Smoothness

When *p=1*, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\sum_i\|d_i f\| + \frac{\mu}{2}\|f - y\|^2\}, \quad (23)$$

By theorem 2, we have

Theorem 5: The solution of (23) satisfies

$$\kappa f + \mu(f - y) = 0, \quad (24)$$

The curvature κ is a non-linear operator; hence we do not have the closed form solution of equation (24). Thus, we have to construct iterative algorithm to obtain the solution. From (24), we have

$$\frac{1}{2}\sum_{i\sim j} w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i) + \mu(f_j - y_j) = 0 \quad (25)$$

Define the function $m: E \to R$ by

$$m_{ij} = \frac{1}{2} w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right) \quad (26)$$

Then (25)

$$\sum_{i\sim j} m_{ij}(f_j - f_i) + \mu(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum_{i\sim j} m_{ij} + \mu\right) f_j = \sum_{i\sim j} m_{ij} f_i + \mu y_j \quad (27)$$

Define the function $p: E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i\sim j} m_{ij} + \mu} & if\ i \neq j \\ \frac{\mu}{\sum_{i\sim j} m_{ij} + \mu} & if\ i = j \end{cases} \quad (28)$$

Then

$$f_j = \sum_{i\sim j} p_{ij} f_i + p_{jj} y_j \quad (29)$$

Thus we can consider the iteration

$f_j^{(t+1)} = \sum_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j$ for all $j \in V$

to obtain the solution of (23).

C. P-Smoothness

For any number *p*, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\frac{1}{p}\sum_i\|d_i f\|^p + \frac{\mu}{2}\|f - y\|^2\}, \quad (30)$$

By theorem 3, we have

Theorem 6: The solution of (30) satisfies

$$\Delta_p f + \mu(f - y) = 0, \quad (31)$$

The p-Laplace operator is a non-linear operator; hence we do not have the closed form solution of equation (31). Thus, we have to construct iterative algorithm to obtain the solution. From (31), we have

$$\frac{1}{2}\sum_{i\sim j} w_{ij}\left(\|d_i f\|^{p-2} + \|d_j f\|^{p-2}\right)(f_j - f_i) + \mu(f_j - y_j) = 0 \quad (32)$$

Define the function $m: E \to R$ by

$$m_{ij} = \frac{1}{2} w_{ij}\left(\|d_i f\|^{p-2} + \|d_j f\|^{p-2}\right) \quad (33)$$

Then equation (32) which is

$$\sum_{i\sim j} m_{ij}(f_j - f_i) + \mu(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum_{i\sim j} m_{ij} + \mu\right) f_j = \sum_{i\sim j} m_{ij} f_i + \mu y_j \quad (34)$$

Define the function $p: E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i-j} m_{ij} + \mu} & if\ i \neq j \\ \frac{\mu}{\sum_{i-j} m_{ij} + \mu} & if\ i = j \end{cases} \quad (35)$$

Then

$$f_j = \sum_{i\sim j} p_{ij} f_i + p_{jj} y_j \quad (36)$$

Thus we can consider the iteration

$f_j^{(t+1)} = \sum_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j$ for all $j \in V$

to obtain the solution of (30).

VIII. Experiments and Results

A. Datasets

In this paper, we use the leukemia dataset available from [13]. The leukemia dataset contains expression levels of 7129 genes over 72 patients. Labels indicate which of two variants of leukemia is present in the sample (i.e. patient). There are 25 AML patients and 47 ALL patients in the dataset. In the other words, we are given gene expression data ($R^{72*7129}$) matrix and the annotation (i.e. the label) vector (R^{72*1}). We then split 72 patients into 38 training and 34 test patients. We filtered the datasets to include only those genes that are expressed differently between two given classes ALL and AML in the training set of the gene expression data by using signal-to-noise ratio method. This resulted in a dataset containing 100 genes with highest signal-to-noise ratio scores.

We refer to this dataset as *leukemia*. There are three ways to construct the similarity graph from the gene expression data:

- The ε-neighborhood graph: Connect all patients whose pairwise distances are smaller than ε.
- k-nearest neighbor graph: Patient *i* is connected with patient *j* if patient *i* is among the k-nearest neighbor of patient j or patient *j* is among the k-nearest neighbor of patient *i*.
- The fully connected graph: All genes are connected.

In this paper, the similarity function is the Gaussian similarity function

$$s(P(i,:), P(j,:)) = \exp\left(-\frac{d(P(i,:), P(j,:))}{t}\right)$$

In this paper, t is set to 10^9 and the 5-nearest neighbor graph (i.e. $k = 5$) is used to construct the similarity graph from leukemia. Please note that $P(i,:)$ is the sample (i.e. patient) expression profile *i*.

B. Experiments

In this section, we experiment with the above proposed un-normalized graph p-Laplacian methods with *p=1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9* and the current state of the art network-based method (i.e. the un-normalized graph Laplacian based semi-supervised learning method *p=2*) in terms of classification accuracy performance measure. The accuracy performance measure Q is given as follows

$$Q = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

All experiments were implemented in Matlab 6.5 on virtual machine. The parameter μ is set to 1.

The accuracy performance measures of the above proposed methods and the current state of the art method is given in the following Table I.

TABLE I. The Comparison of Accuracies of Proposed Methods with Different p-Values

With/without Signal To Noise (i.e. SNR) Ratio filter		With SNR filter	Without SNR filter (k=3)
Accuracy Performance Measures (%)	p=1	**94.12**	76.47
	p=1.1	**94.12**	82.35
	p=1.2	**94.12**	82.35
	p=1.3	**94.12**	88.24
	p=1.4	**94.12**	**91.18**
	p=1.5	91.18	**91.18**
	p=1.6	91.18	88.24
	p=1.7	91.18	88.24
	p=1.8	91.18	88.24
	p=1.9	91.18	88.24
	p=2.0 (current state of the art network-based method)	91.18	88.24

From the above table, we easily recognized that the un-normalized graph p-Laplacian semi-supervised learning method outperform the current state of art network-based method. The results from the above table shows that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art method (*p=2*) but often lead to better classification accuracies.

IX. Conclusion

We have developed the detailed regularization frameworks for the un-normalized graph p-Laplacian semi-supervised learning methods applying to cancer classification problem. Experiments show that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art method (i.e. *p=2*) but often lead to significant better classification accuracy performance measures.

Moreover, these un-normalized graph p-Laplacian semi-supervised learning methods can not only be used in classification problem but also in ranking problem. In specific, given a set of genes (i.e. the queries) making up a protein complex/pathways or given a set of genes (i.e. the queries) involved in a specific disease (for e.g. leukemia), these methods can also be used to find more potential members of the complex/pathway or more genes involved in the same disease by ranking genes in gene co-expression network (derived from gene expression data) or the protein-protein interaction network or the integrated network of them. The genes with the highest

rank then will be selected and then checked by biologist experts to see if the extended genes in fact belong to the same complex/pathway or are involved in the same disease. These problems are also called complex/pathway membership determination and biomarker discovery in cancer classification.

REFERENCES

[1] C. Perou, S. Jeffrey, M. V. D. Rijn, C. Rees, M. Eisen, D. Ross, A. Pergamenschikov, *et al.*, "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *PNAS*, 1999, vol. 96, no. 16, pp. 9212–9217.

[2] A. Ben-Dor, R. Shamir, and R. Yakhini, *Clustering Gene Expression Patterns Journal of Computational Biology*, vol. 6, pp. 281–297, 1999.

[3] M. Sarkar and T. Y. Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem," in *Proc AMIA Symp*, 2000, pp. 759-763

[4] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, *et al*., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–79, 2001.

[5] T. Furey, N. Christianini, N. Duffy, D. Bednarski, M. Schummer, and D. Hauessler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[6] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, *Learning with Local and Global Consistency Advances in Neural Information Processing Systems (NIPS)*, S. Thrun, L. Saul and B. Schölkopf, Eds., MIT Press, Cambridge, MA, 2004, vol. 16, pp. 321-328.

[7] X. J. Zhu and Z. B. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *Technical Report CMU-CALD-02-107*, Carnegie Mellon University, 2002.

[8] K. Tsuda, H. H. Shin, and B. Schoelkopf, "Fast protein classification with multiple networks Bioinformatics," *ECCB'05*, vol. 21, suppl. 2, pp. ii59-ii65, 2005.

[9] L. Tran, "Application of three graph laplacian based semi-supervised learning methods to protein function prediction problem," *International Journal of Bioinformatics & Biosciences*, 2013.

[10] D. Zhou and B. Schölkopf, "Regularization on discrete spaces pattern recognition," in *Proc. 27th DAGM Symposium*, Springer, Berlin, Germany, 2005, pp. 361-368.

[11] D. Zhou and B. Schölkopf, *Discrete Regularization Book chapter, Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, Eds., MIT Press, Cambridge, MA, 2006, pp. 221-232.

[12] L. Tran, "The un-normalized graph p-Laplacian based semi-supervised learning method and protein function prediction problem," presented at The Fifth International Conference on Knowledge Systems and Engineer, 2013

[13] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

Loc Tran completed Bachelor of Science and Master of Science in Computer Science at University of Minnesota in 2003 and 2012 respectively. Currently, he's a PhD student at University of Technology, Sydney.

Linh Tran completed Bachelor of Science and Master of Science in Electrical and Computer Engineer at Portland State University. Currently, he's a PhD student at Portland State University.

