Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 826373, 10 pages
http://dx.doi.org/10.1155/2014/826373

Research Article

NIM: A Node Influence Based Method for Cancer Classification

Yiwen Wang, Min Yao, and Jianhua Yang

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

Correspondence should be addressed to Jianhua Yang; jhyang@zju.edu.cn

Received 10 March 2014; Revised 16 June 2014; Accepted 23 June 2014; Published 11 August 2014

Academic Editor: Shengyong Chen

The classification of different cancer types owns great significance in the medical field. However, the great majority of existing cancer classification methods are clinical-based and have relatively weak diagnostic ability. With the rapid development of gene expression technology, it is able to classify different kinds of cancers using DNA microarray. Our main idea is to confront the problem of cancer classification using gene expression data from a graph-based view. Based on a new node influence model we proposed, this paper presents a novel high accuracy method for cancer classification, which is composed of four parts: the first is to calculate the similarity matrix of all samples, the second is to compute the node influence of training samples, the third is to obtain the similarity between every test sample and each class using weighted sum of node influence and similarity matrix, and the last is to classify each test sample based on its similarity between every class. The data sets used in our experiments are breast cancer, central nervous system, colon tumor, prostate cancer, acute lymphoblastic leukemia, and lung cancer. experimental results showed that our node influence based method (NIM) is more efficient and robust than the support vector machine, K-nearest neighbor, C4.5, naive Bayes, and CART.

1. Introduction

Cancer research is one of the major research areas in the medical field. In cancer, cells divide and grow uncontrollably, forming malignant tumors and invading adjacent parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or bloodstream. Many things are deemed to increase the risk of cancer, including tobacco use, dietary factors, certain infections, exposure to radiation, lack of physical activity, obesity, and environmental pollutants. The famous Apple founder Steve Jobs also died of pancreatic cancer. Any method which benefits cancer treatment should receive sufficient attention.

The biggest challenge facing cancer treatment process is a means of developing individualized treatment programs for specific tumor types. Traditional diagnosis of cancer depends on the type of tissue-derived tumor cells, cell morphology, and protein markers, and biological behavior does not adequately reflect the real situation of the tumor; it is sometimes difficult to make a correct diagnosis of forecasts.

In order to gain a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been proposed [1, 2]. The expression level of genes is deemed to contain the keys to addressing fundamental problems relating to the prevention and cure of diseases, biological evolutionary mechanisms, and drug discovery. The recent advent of microarray technology has upheld the simultaneous monitoring of thousands of genes, which motivated the development in cancer classification using gene expression data [3]. From the data mining perspective, measuring the gene sequence to predict tumor is actually a classification problem. Due to the characteristics of gene expression data, there are three challenges for cancer classification.

(1) High Dimension. Each species genome is composed of a nucleotide sequence encoding a protein and nonprotein coding; the former is the traditional sense of the gene, which is a potential gene. Usually, the number of genes is the total of both. The number of genes in the human genome is approximately 30000. The dimension of the data is so high, brought great difficulties to the analysis of the experimental results. For example, used in our experiments the maximum dimension of a data set is up to 24481 in breast cancer [4].

(2) Small Sample Size. Since the acquisition of gene expression experiments in extreme cost data, publicly available data size is very small. Most tumor gene expression data sample

numbers of only tens or hundreds. But the traditional classification methods often require a large set of test samples to obtain the high classification accuracy. This is a huge challenge for classification algorithm. For instance, used in our experiments later cancer data central nervous system [5] has only 60 samples, but with 7129 dimensions.

(3) Nonbalanced Distribution. Usually, the traditional classification methods can achieve outstanding results when using balanced distribution data. However, gene microarray data for cancer classification are nonbalanced distribution. For example, in lung cancer [6] used in our experiments later, the number of samples of the MPM class is 31 and the number of samples of ADCA class is 150, which is nearly 5 times that of the former.

2. Node Influence Model

Let m indicate the number of genes measured. Every cancer sample can be viewed as a point in m-dimensional space. And the set of cancer samples can be viewed as a graph (or network) in m-dimensional space. Our idea is to confront the problem of cancer classification from graph-based view. In graph theory, a graph (or network) is usually presented by an adjacency matrix. If a graph has N vertices, we may associate it with an $N \times N$ matrix A. The adjacency matrix A is defined by

$$A\left(v_i, v_j\right) = \begin{cases} 1, & v_i \text{ and } v_j \text{ connected,} \\ 0, & v_i \text{ and } v_j \text{ not connected.} \end{cases} \quad (1)$$

2.1. Centrality Measures for Node Influences. The centrality of nodes, or the identification of the importance of nodes, is a key issue in network analysis. Degree is the simplest of the node centrality measures by using the local structure around nodes only. In an undirected network, the degree is equal to the number of edges a node has. In a directed network, a node may have a different number of outgoing and incoming edges, and therefore, degree is split into out-degree and in-degree, respectively. The degree centrality of a vertex v_i, for a given graph $G = (V, E)$ with $|V| = N$ vertices and $|E| = M$ edges, is defined as

$$\text{Degree}\left(v_i\right) = \sum_{j=1}^{N} \delta_i^j, \quad (2)$$

$$\delta_i^j = \begin{cases} 1, & v_i \text{ and } v_j \text{ connected,} \\ 0, & v_i \text{ and } v_j \text{ not connected.} \end{cases}$$

Closeness is defined as the inverse of farness, which in turn, is the sum of distances to all other nodes [7]. The intent behind this measure is to identify the nodes which could reach others quickly. The closeness centrality of a vertex v_i, for a given graph $G = (V, E)$ with $|V| = N$ vertices and $|E| = M$ edges, is defined as

$$\text{Closeness}\left(v_i\right) = \frac{1}{\sum_{i \neq j, v_j \in E} d_{ij}}, \quad (3)$$

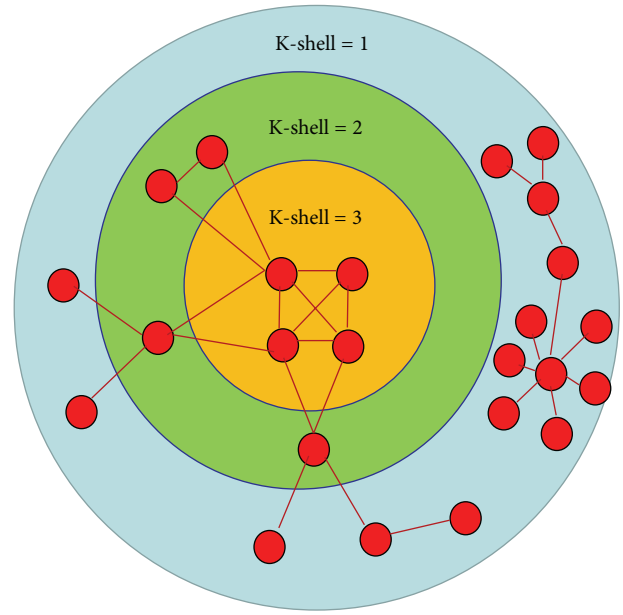

Figure 1: A schematic representation of the K-shell.

where d_{ij} is the distance of shortest path from node v_i to node v_j.

Another famous node centrality is betweenness [7], a measure of how many shortest paths cross through this node, which is believed to determine who has more interpersonal influence on others. High betweenness individuals often do not have the shortest average path to everyone else, but they have the greatest number of shortest paths that necessarily have to go through them. Betweenness centrality of a vertex v_i, for a given graph $G = (V, E)$ with $|V| = N$ vertices and $|E| = M$ edges, is defined as

$$\text{Betweenness}\left(v_i\right) = \sum_{v_s \neq v_i \neq v_s \in V} \frac{\sigma_{st}\left(v_i\right)}{\sigma_{st}}, \quad (4)$$

where σ_{st} is total number of shortest paths from node v_s to node v_t and $\sigma_{st}(v_i)$ is the number of those paths that pass through node v_i.

K-shell [8] is a relatively recent and robust centrality. Nodes are assigned to K shells according to their remaining degree, which is obtained by successive pruning of nodes with degree smaller than the K-shell value of the current layer. Figure 1 is a schematic representation of the K-shell. The outermost circle of Figure 1 is the nodes with K-shell = 1; delete theses nodes and then consider the remaining nodes of degree 2. Then we obtain the second layer nodes with K-shell = 2. Delete the nodes of K-shell = 2; we finally obtain the innermost nodes with K-shell = 3.

2.2. Node Influence Centrality. As can be seen from the four above node centralities in complex networks, degree is the most intuitive and simple, but only considering local information. Both betweenness and closeness use shortest paths between every pair of nodes in the network as primary

factor. K-shell approach is based on the node degree, but it is from a global perspective.

In our opinion, the evaluation of node centrality can start from the influence of a node on another node. Now consider the influence of node v_i on v_j. If v_i can influence v_j, that means that there are some paths which connected the two nodes from the topological view. So the number of paths connecting node v_i and node v_j is able to reflect the influence. From a global perspective, the number of connected paths between all nodes and node v_j must be taken into account. Therefore, we define the influence of one node on another node with length k as follows:

$$\text{Influence}_{i\to j}^{k} = \frac{\sigma_j^k(v_i)}{\sigma_j^k}, \tag{5}$$

when σ_j^k represents the number of connected paths between all nodes and node v_j with length k and $\sigma_j^k(v_i)$ represents the number of connected paths between node v_i and node v_j. We found that, in an undirected network, when k tends to infinity, $\text{Influence}_{i\to j}^{k}$ will fluctuate at the beginning and then stabilize; that is, it will converge to a certain value.

Theorem 1. *When k tends to infinity, $\text{Influence}_{i\to j}^{k}$ will converge to a certain value in an undirected network.*

Proof. To facilitate the proof, we introduce one good nature of adjacency matrix. That is, the kth power of the adjacency matrix elements represents the corresponding number of connected paths between two nodes with length k. Consider

$$\begin{aligned} \sigma_j^k(v_i) &= A^k(v_i, v_j), \\ \sigma_j^k &= \sum_{m=1}^{N} A^k(v_m, v_j). \end{aligned} \tag{6}$$

So (2) can change to

$$\text{Influence}_{i\to j}^{k} = \frac{A^k(v_i, v_j)}{\sum_{m=1}^{N} A^k(v_m, v_j)}. \tag{7}$$

As we consider the undirected network, A is a real symmetric matrix, which can be diagonalized. That is, $A = P\cdot D\cdot P'$, P' is the transpose of P, and $P' = P^{-1}$; P^{-1} is inverse of P. D is a diagonal matrix, whose elements are the eigenvalues of the matrix A; P is the corresponding eigenvector. So,

$$\begin{aligned} A^k &= \left(PDP'\right)^k = \left(PDP^{-1}\right)^k = PD^kP', \\ D^k &= \begin{pmatrix} d_1^k & & & \\ & d_2^k & & \\ & & \cdots & \\ & & & d_N^k \end{pmatrix}, \end{aligned} \tag{8}$$

$A^k(v_i, v_j) = d_1^k P_{v_i1}P_{v_j1} + d_2^k P_{v_i2}P_{v_j2} + \cdots + d_N^k P_{v_iN}P_{v_jN} = \sum_{n=1}^{N} d_n^k P_{v_in}P_{v_jn}$, so we get $\text{Influence}_{i\to j}^{k} = A^k(v_i, v_j)/\sum_{m=1}^{N} A^k(v_m, v_j) = \sum_{n=1}^{N} d_n^k P_{v_in}P_{v_jn}/\sum_{m=1}^{N}\sum_{n=1}^{N} d_n^k P_{mn}P_{v_jn}$; let $d_{\max}$ be the largest absolute value of eigenvalues of matrix A; then

$$\begin{aligned} \text{Influence}_{i\to j}^{k} &= \frac{\sum_{n=1}^{N}\left(d_n^k/d_{\max}^k\right)P_{v_in}P_{v_jn}}{\sum_{m=1}^{N}\sum_{n=1}^{N}\left(d_n^k/d_{\max}^k\right)P_{mn}P_{v_jn}} \\ &= \frac{P_{v_i\max}P_{v_j\max}}{\sum_{m=1}^{N}P_{m\max}P_{v_j\max}} = \frac{P_{v_i\max}}{\sum_{m=1}^{N}P_{m\max}} \\ &\qquad (k \longrightarrow \infty) \end{aligned} \tag{9}$$

if $d_{\max}$ is t-repeated characteristic roots, and $P_{\max}$ is the corresponding eigenvector associated with t-repeated roots. Consider

$$\text{Influence}_{i\to j}^{k} = \frac{\sum_{w=1}^{t}P_{v_i\max_w}}{\sum_{m=1}^{N}\sum_{w=1}^{t}P_{m\max_w}}, \quad (k \longrightarrow \infty). \tag{10}$$

There is a special case that the largest absolute eigenvalues of matrix A are two opposite numbers. But this only happens in bipartite graph [9] and the cancer samples network is not a bipartite graph. □

Theorem 2. *When k tends to infinity, $\text{Influence}_{i\to j}^{k}$ will converge to a certain value independent of the j in an undirected network.*

Proof. From the proof of Theorem 1, we see (9) and (10), so $\text{Influence}_{i\to j}^{k}$ will converge to a certain value independent of the j. □

From Theorem 2, we know that the influence of node v_i on every other node in network with length k is the same when k tends to infinity. This reflects the impact of a single node v_i on the whole network. So we define the node influence centrality as

$$\begin{aligned} &\text{Node Influence}(v_i) \\ &= \frac{A^k(v_i, v_j)}{\sum_{m=1}^{N}A^k(v_m, v_j)}, \quad (1 \le i, j \le N, k \longrightarrow \infty). \end{aligned} \tag{11}$$

2.3. Example for Node Influence. For example, the network shown in Figure 2 is represented by the adjacency matrix as follows:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{12}$$

According to (5), we calculate node 4 to each node's influence. Curves are shown in Figure 3. Curves with different colors represent the influence from node 4 every different node. From Figure 3, we can see that the influence from node 4 on each node flickers at the beginning and finally converges

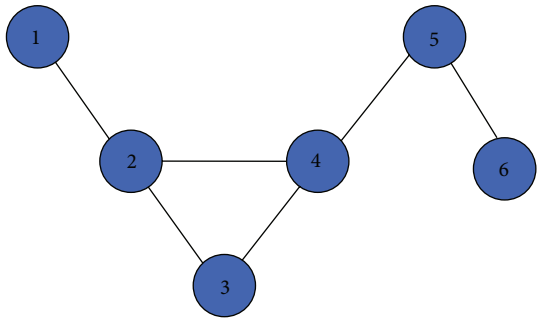

Figure 2: Node influence centrality example network.

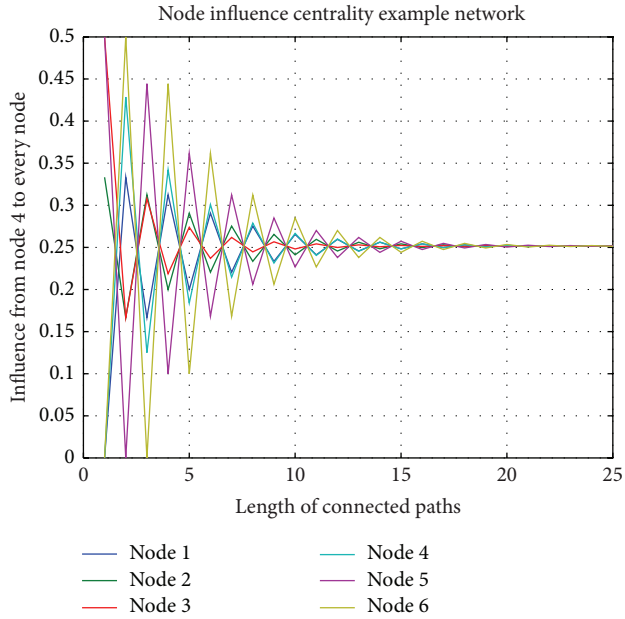

Figure 3: Node 4 to each node's influence.

to about 0.25 (accurate 0.2517). This result is consistent with Theorem 2.

We also calculated the influence of each node on node 4, curves as shown in Figure 4. Curves with different colors represent the influence from each node on node 4. From Figure 4, we can see that the influence from each node on node 4 flickering at the beginning finally converges to different value. It is obvious that the result is consistent with Theorem 1.

3. Methods

3.1. Similarity Matrix. Let m indicate the number of genes measured. Every cancer sample can be viewed as a point in m-dimensional space. Let N indicate the number of samples. The according cancer samples network can be described by an $N \times N$ adjacency matrix. Edges between two nodes represent similarity between two cancer samples. For example, there are two cancer samples X and Y, $X = (x_1, x_2, x_3, \ldots, x_m)$,

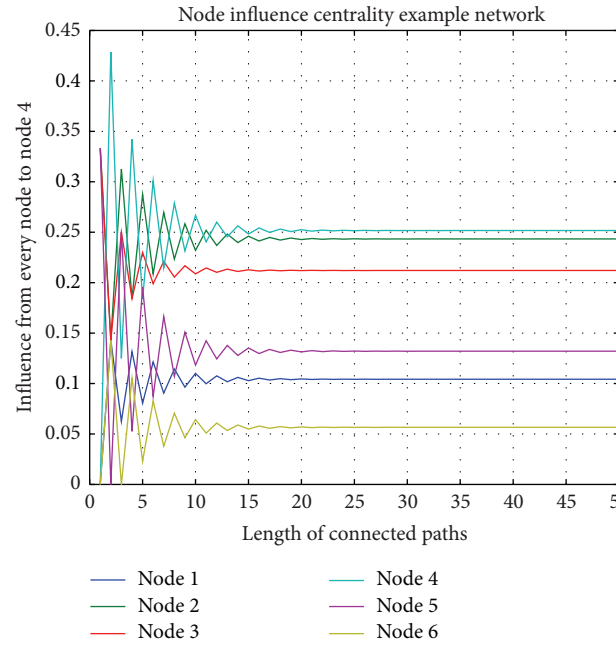

Figure 4: Each node to node 4 influence.

$Y = (y_1, y_2, y_3, \ldots, y_m)$. The weight of edge between X and Y is defined as follows:

$$\text{Similarity}(X, Y) = \exp\left(-\frac{\text{Dist}(X, Y)}{2\delta^2}\right), \tag{13}$$

where the $\text{Dist}(X, Y)$ is the distance metric function for two cancer samples. There are various distance metric functions. And Euclidean distance is a commonly used measure of distance when the prior knowledge is absent. Consider

$$\text{Dist}_{\text{Eu}}(X, Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{14}$$

After using Euclidean distance, (13) becomes

$$\text{Similarity}(X, Y) = \exp\left(-\frac{\sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}}{2\delta^2}\right). \tag{15}$$

For example, the distance matrix of prostate cancer [10] described in Table 1 is shown in Figure 5. The according similarity matrix with $\delta = 3.16$ is shown in Figure 6. Since there are 136 samples in prostate cancer dataset, the according distance matrix and similarity matrix are both 136×136.

3.2. Node Influence Based Method 1 (NIM1). Node influence centrality plays a significant role in our graph-based method for cancer classification. Let X_{train} represent the training set, and let X_{test} represent the test set. All samples are divided into n classes, namely, $C_1, C_2, \ldots, C_n$. Every sample has m dimensions, namely, $a1, a2, \ldots, am$. There are seven main

Table 1: Description for cancer gene data sets.

Dataset	Number of samples	Number of genes	Number of classes	Test method
Breast cancer	97	24481	2	78train-19test
Central nervous system	60	7129	2	LOOCV
Colon tumor	62	2000	2	LOOCV
Prostate cancer	136	12600	2	102train-34test
Acute lymphoblastic leukemia	327	12558	7	215train-112test
Lung cancer	181	12533	2	32train-149test

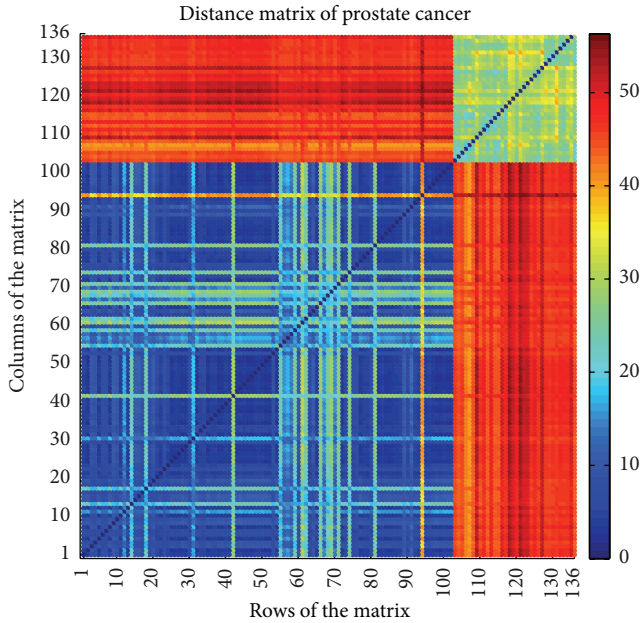

Figure 5: Distance matrix of prostate cancer.

steps in node influence based method 1 (NIM1) for cancer classification.

Step 1. Data preprocessing, mainly normalization, the training set, and testing set are mapped to $[0, 1]$ range in each dimension. Only in this way can we make meaningful comparisons in later steps. Consider

$$x \cdot aj = \frac{\max(aj) - x \cdot aj}{x \cdot aj - \min(aj)}, \quad (x \in \{X_{\text{train}}, X_{\text{test}}\}, 1 \le j \le m). \tag{16}$$

Step 2. Select the appropriate distance metric function based on the actual problem background. If there is no prior

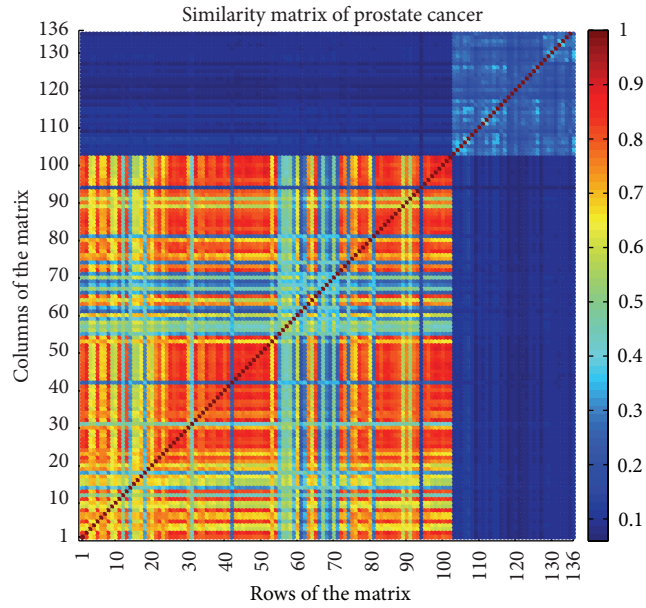

Figure 6: Similarity matrix of prostate cancer.

knowledge, we recommend using the Euclidean distance. Consider

$$\text{Dist}(x1, x2) = \sqrt{\sum_{j=1}^{m} (x1 \cdot aj - x2 \cdot aj)^2} \quad (x1, x2 \in \{X_{\text{train}}, X_{\text{test}}\}). \tag{17}$$

Step 3. Set the only parameter δ; calculate the similarity between every two samples to construct the similarity matrix. Consider

$$\text{Similarity}(x_1, x_2) = \exp\left(-\frac{\text{Dist}(x_1, x_2)}{2\delta^2}\right) \quad (x1, x2 \in \{X_{\text{train}}, X_{\text{test}}\}). \tag{18}$$

Step 4. The training set and test set are treated as a nonnegative weighted undirected network. That is, each sample in the training set or test set is treated as a node in a graph. The similarity obtained in Step 3 for every two samples is treated as the weight of the edge connecting the two corresponding nodes. Then we obtain the adjacency matrix for the whole cancer samples. Consider

$$A(x1, x2) = \text{Simlarity}(x1, x2) \quad (x1, x2 \in \{X_{\text{train}}, X_{\text{test}}\}). \tag{19}$$

Step 5. Calculate the node influence centrality of each training sample node, and treat it as the weight. Consider

$$\text{Node Influence}(x_{\text{train}}) = \frac{A^k(x_{\text{train}}, x_a)}{\sum_{x \in \{X_{\text{train}}, X_{\text{test}}\}} A^k(x, x_a)} \quad (x_{\text{train}} \in X_{\text{train}}, k \longrightarrow \infty). \tag{20}$$

x_a is an arbitrary element in set $\{X_{\text{train}}, X_{\text{test}}\}$. Consider

$$\text{weight}\left(x_{\text{train}}\right) = \text{Node Influence}\left(x_{\text{train}}\right). \tag{21}$$

Step 6. Calculate the similarity between every test sample and each class. Consider

$$\begin{aligned}&\text{Similarity Class}\left(x_{\text{test}}, C_i\right)\\ &= \frac{\sum \text{Similarity}\left(x_{\text{test}}, x_{\text{train}}\right) \cdot \text{weight}\left(x_{\text{train}}\right)}{\sum \text{weight}\left(x_{\text{train}}\right)}\\ &\left(x_{\text{test}} \in X_{\text{test}}, x_{\text{train}} \in X_{\text{test}}, \text{Class}\left(x_{\text{train}}\right) = C_i, 1 \le i \le n\right).\end{aligned} \tag{22}$$

Step 7. Classify each test sample to the class with highest similarity. Consider

$$\text{Class}\left(x_{\text{test}}\right) = \arg\max\left(\text{Similarity Class}\left(x_{\text{test}}, C_i\right)\right). \tag{23}$$

3.3. Node Influence Based Method 2 (NIM2). Similarity Matrix is used twice in seven main steps of NIM1. The first is located in Step 4, in order to obtain the adjacency matrix. The second is in Step 6, in order to calculate the similarity between every test sample and each class. We believe in two steps used in different similarity matrix, resulting in node influence based method 2 (NIM2). Only two main steps of NIM2 are different from NIM1, as shown below.

Step 3. Set the parameter δ_1; calculate the similarity between every two samples to construct the similarity matrix. Consider

$$\begin{aligned}&\text{Similarity}\left(x_1, x_2\right)\\ &= \exp\left(-\frac{\text{Dist}\left(x_1, x_2\right)}{2\delta_1^2}\right) \quad \left(x1, x2 \in \{X_{\text{train}}, X_{\text{test}}\}\right).\end{aligned} \tag{24}$$

Step 6. Set the parameter δ_2; calculate the similarity between every two samples and then obtain the similarity between every test sample and each class. Consider

$$\begin{aligned}&\text{Similarity } 2\left(x_1, x_2\right)\\ &= \exp\left(-\frac{\text{Dist}\left(x_1, x_2\right)}{2\delta_2^2}\right) \quad \left(x1, x2 \in \{X_{\text{train}}, X_{\text{test}}\}\right),\\ &\text{Similarity Class}\left(x_{\text{test}}, C_i\right)\\ &= \frac{\sum \text{Similarity } 2\left(x_{\text{test}}, x_{\text{train}}\right) \cdot \text{weight}\left(x_{\text{train}}\right)}{\sum \text{weight}\left(x_{\text{train}}\right)}\\ &\left(x_{\text{test}} \in X_{\text{test}}, x_{\text{train}} \in X_{\text{test}}, \text{Class}\left(x_{\text{train}}\right) = C_i, 1 \le i \le n\right).\end{aligned} \tag{25}$$

4. Experimental Results and Analysis

4.1. Benchmark Data Sets. We use 6 data sets to validate NIM1 and NIM2. Below are six publicly available gene expression data from DNA microarray that are widely used by researchers for cancer classification experiments. All the data sets are used to predict various kinds of cancers by measuring gene sequences and are outlined in Table 1.

The first data set is breast cancer [4]. The training data contains 78 patient samples, 34 of which are from patients who had developed distance metastases within 5 years. The remaining 44 samples are from patients who remained healthy from the disease after their initial diagnosis for an interval of at least 5 years.

The second data set is central nervous system [5]. Survivors are patients who are alive after treatment while the failures are those who succumbed to their disease. The data set contains 60 patient samples; 21 are survivors and 39 are failures. There are 7129 genes in the dataset.

The third data set is colon tumor [11]. It contains 62 samples gathered from colon-cancer patients. Among them, 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected founded on the confidence in the measured expression levels.

The fourth data set is prostate cancer [10]. The training set contains 52 prostate tumor samples and 50 nontumor prostate samples with around 12600 genes.

The fifth data set is acute lymphoblastic leukemia [12]. The data have been divided into six diagnostic groups and one that contains diagnostic samples that did not fit into any one of the above groups.

The sixth data set is lung cancer [6]. It is about the classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 of them, 16 MPM and 16 ADCA. The remaining 149 samples are used for testing. Each sample is characterized by 12533 genes.

If the dataset has not been divided into training set and testing set, we adopt leave-one-out cross validation (LOOCV) to validate NIM1 and NIM2. LOOCV involves using a single observation from the original sample as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sampling.

4.2. Results. Most proposed cancer classification methods are from the statistical and machine learning area, ranging from the old nearest neighbor analysis to the new support vector machines. There is no single classifier that is superior over the rest. Some of the methods only work well on binary-class problems and are not extensible to multiclass problems, while others are more general and flexible. The methods we choose for comparing are all top 10 algorithms in data mining, mentioned in [13]. They are support vector machine (SVM) [14], k-nearest neighbor (KNN) [15], C4.5 [16], naive Bayes [17], and CART [18]. And we use the popular noncommercial open platform Weka (Waikato Environment for Knowledge Analysis) [19] for the implementation of the algorithms

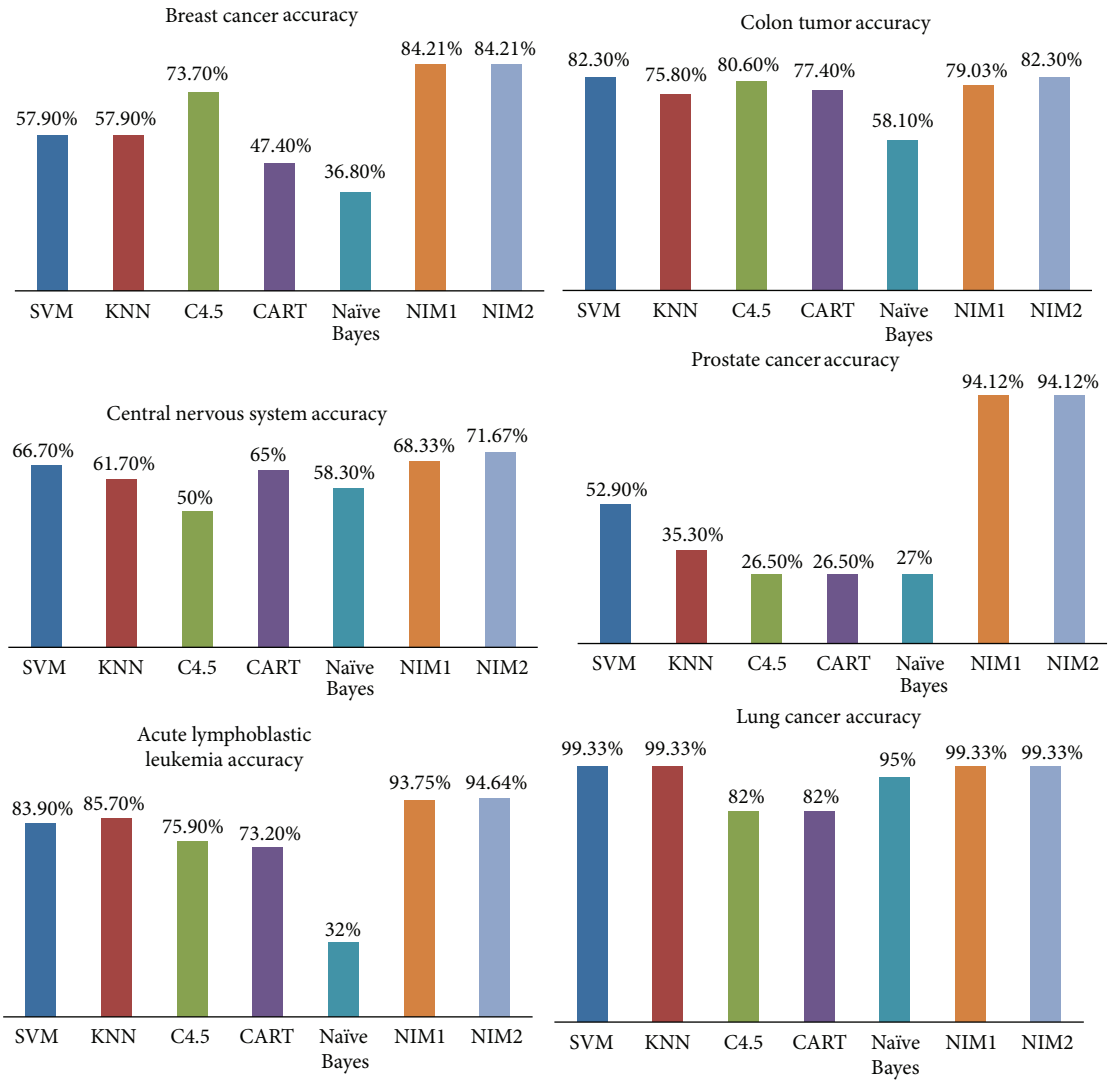

Figure 7: Experimental results for cancer gene datasets.

above. Experimental results on these six data sets using SVM, KNN, C4.5, Naive Bayes, NIM1, and NIM2 are presented in Figure 7.

Due to high dimension, small sample size, and nonbalanced distribution, traditional classification algorithms do not obtain high accuracy in these data sets. From Figure 6, we can see clearly that NIM1 obtain the highest accuracy in 5 of 6 data sets, and especially 94.12% in prostate cancer, compared to poor performance of other algorithms. And in colon tumor in which NIM1 does not get the highest accuracy, the performance of NIM1 differs very little with the highest one.

NIM2 is an improved version of NIM1 and has one more parameter. NIM1 can be viewed as a special case of NIM2 when $\delta_1 = \delta_2$. So the results of NIM2 are at least as good as NIM1. From Figure 7, we can see clearly that NIM2 obtain the highest accuracy in all 6 data sets. Thus, NIM1 and NIM2 are more efficient and robust than traditional classification algorithms in these cancer gene data sets.

Table 2: Parameter setting for δ in NIM1.

Dataset	Minimum of δ	Maximum of δ	Change interval	Number of experiments
Colon tumor	0.501	1.5	0.001	1000
Acute lymphoblastic leukemia	0.01	10	0.01	1000
Lung cancer	0.01	10	0.01	1000

4.3. Parameters Discussion. The traditional classification methods usually tend to have many parameters need to be set before application. And the parameters are closely related to the performance. However, there is little information on how to set parameters, usually based on experience. So we try to propose an algorithm with as few parameters as possible. NIM1 has only one parameter δ, and NIM2 has only two parameters δ_1 and δ_2.

The parameter setting for δ in NIM1 is shown in Table 2, and parameters setting for δ_1 and δ_2 in NIM2 is shown in

Table 3: Parameters setting for δ_1, δ_2 in NIM2.

Dataset	Minimum of δ_1, δ_2	Maximum of δ_1, δ_2	Change interval of δ_1, δ_2	Number of experiments
Colon tumor	0.501, 0.501	1.5, 1.5	0.001, 0.001	1000000
Acute lymphoblastic leukemia	0.501, 0.501	10.5, 10.5	0.01, 0.01	1000000
Lung cancer	0.501, 0.501	10.5, 10.5	0.01, 0.01	1000000

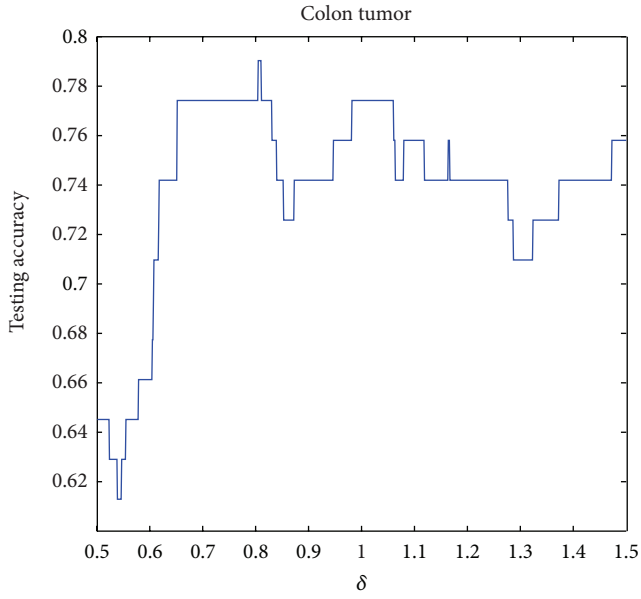

Figure 8: NIM1 results in colon tumor with the variation of δ.

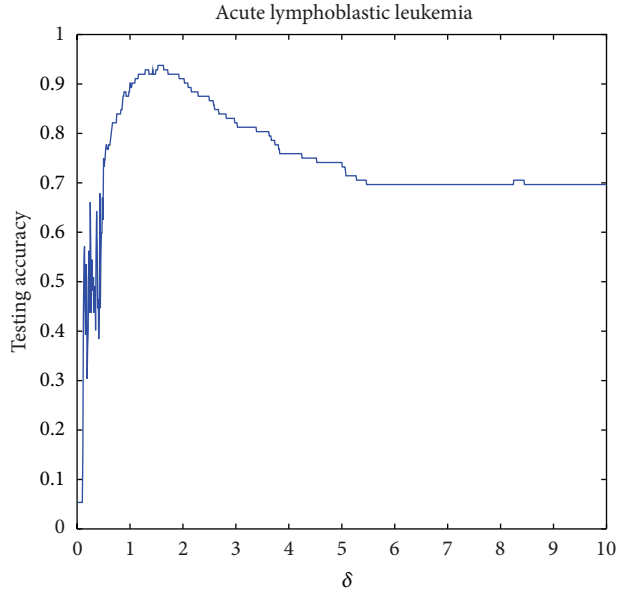

Figure 9: IM1 results in ALL with the variation of δ.

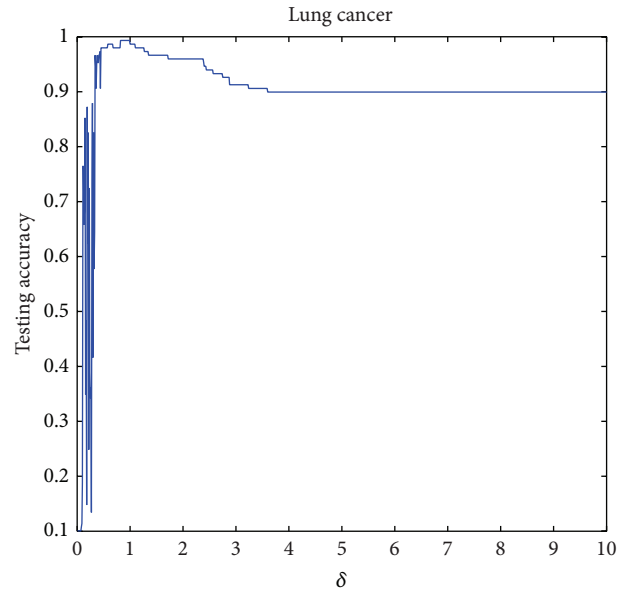

Figure 10: NIM1 results in lung cancer with the variation of δ.

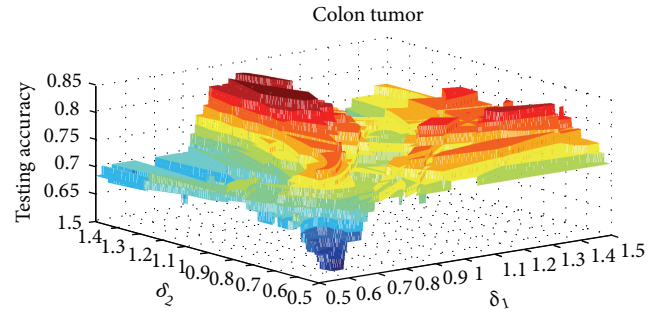

Figure 11: NIM2 results in colon tumor with the variation of δ_1, δ_2.

Table 3. Three data sets are selected for parameter variation experiments; they are colon tumor, acute lymphoblastic leukemia, and lung cancer. Figures 8, 9, and 10 show the results of NIM1 with the variation of δ in the 3 data sets. Figures 11, 12, and 13 show the results of NIM2 with the variation of δ_1, δ_2 in the 3 data sets. From the experimental results shown in Figures 11, 12, and 13, we can see clearly that both of the δ_1 and δ_2 play an important role in the performance of NIM2.

5. Conclusion

Graph is a powerful representation formalism that has been widely employed in machine learning and data mining. In order to gain deep insight into the cancer classification problem, we analyze the problem from graph-based view. Let m indicate the number of genes measured. Every cancer sample can be viewed as a point in m-dimensional space.

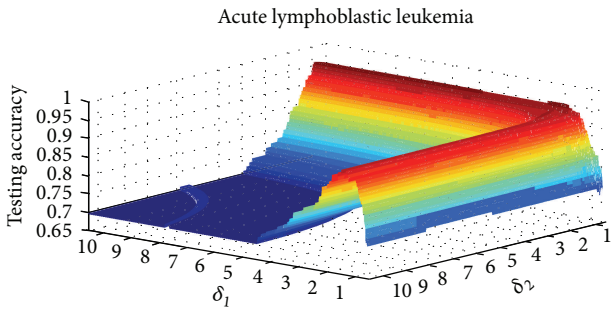

Figure 12: NIM2 results in ALL with the variation of δ_1, δ_2.

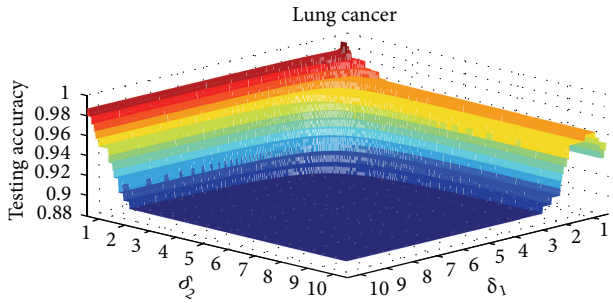

Figure 13: NIM2 results in lung cancer with the variation of δ_1, δ_2.

And the set of cancer samples can be viewed as a graph (or network) in m-dimensional space.

In the method NIM1, after selecting the appropriate distance metric, the graph (or network) of all samples is created by computing the similarity matrix. Then the node influence of training samples is calculated. Treat node influence as weight; the similarity between every test sample and each class is obtained. At last, every test sample is classified according to its similarity between each class.

Furthermore, we also propose NIM2, which is an improved version of NIM1. NIM1 can be viewed as a special case of NIM2 when $\delta_1 = \delta_2$. Both NIM1 and NIM2 can deal with binary and multiclass cancer classification. NIM2 is more time consuming than NIM1 but owns a higher accuracy.

Due to high dimension, small sample size, and nonbalanced distribution, SVM, KNN, C4.5, Naive Bayes, and CART do not obtain high accuracy in these cancer gene data sets. From the experimental results in the 6 cancer gene data sets, it can be seen that NIM1 and NIM2 are more efficient than these traditional algorithms. At the end, we also discuss the parameters in both NIM1 and NIM2. The parameters play an important role in the performance of NIM1 and NIM2.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The research work was the partial achievement of Project 2013CB329504 supported by National Key Basic Research and Development Program (973 program) and STD of Zhejiang (2012C21002).

References

[1] A. Goffeau, "DNA technology: molecular fish on chips," *Nature*, vol. 385, no. 6613, pp. 202–203, 1997.

[2] A. Marshall and J. Hodgson, "DNA chips: all array of possibilities," *Nature Biotechnology*, vol. 16, no. 1, pp. 27–31, 1998.

[3] Z. Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," *Artificial Intelligence in Medicine*, vol. 42, no. 1, pp. 81–93, 2008.

[4] L. J. van't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[5] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[6] G. J. Gordon, R. V. Jensen, L. Hsiao et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[7] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[8] M. Kitsak, L. K. Gallos, S. Havlin et al., "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[9] A. E. Brouwer and W. H. Haemers, *Spectra of Graphs*, Springer, New York, NY, USA, 2012.

[10] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[11] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[12] E. Yeoh, M. E. Ross, S. A. Shurtleff et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[13] X. Wu, V. Kumar, Q. J. Ross et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[15] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, Calif, USA, 1993.

[17] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Belmont, Wadsworth, Ohio, USA, 1984.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *Sigkdd Explorations*, vol. 11, no. 1, pp. 10–18, 2009.