

# Semi-supervised Method for Gene Expression Data Classification with Gaussian Fields and Harmonic Functions

Yun-Chao Gong  
Software Institute  
Nanjing University, China  
gyc05@software.nju.edu.cn

Chuan-Liang Chen  
Department of Computer Science  
Beijing Normal University, China  
c.l.chen86@gmail.com

## Abstract

*In real world applications, there are great many of DNA expressed microarray data, many supervised classification algorithms such as decision tree, KNN and SVM in the machine learning field have been introduced for microarray data classification. However, in real worlds, the labeled examples, especially gene expression data examples are often very difficult and expensive to obtain. The traditional supervised methods can not work well when lack of training examples. So in this paper, we propose to use the semi-supervised learning algorithms which learning with both labeled and unlabeled data to do classification for microarray data. We perform experiments on four public microarray data sets and the results showed the semi-supervised method holds a much higher classification accuracy than the supervised methods and is much more stable when the labeled examples are very few.*

## 1. Introduction

“In recent years, the rapid development of DNA microarray technology has made it possible for scientists to monitor the expression level of thousands of genes with a single experiment” [10]. Many classification and clustering algorithms in the machine learning field have been used on gene expression data sets. A major task for microarray data classification is to first use the existing history data to build a classifier and then use the classifier to classify new coming data. Typical methods for microarray classification are all supervised [6], such as the C4.5 decision tree algorithm [7, 8], Support Vector machine (SVM) [5], the k-nearest neighbor classifier (KNN), and the and ensemble methods, such as Bagging and boosting [11]. In supervised setting algorithms, there is one important problem: when the train-

ing examples are large and sufficient, the classifiers can perform well, but when the training examples are few and insufficient for training, the classifiers performs bad and very unstable.

To deal with these kinds of problems of lacking the training data, in recent years, scientists developed the semi-supervised methods which use both labeled and unlabeled data to train. There are many semi-supervised methods been proposed which use both the labeled and unlabeled data into the learning form. Research of semi-supervised learning have been surveyed in a recent survey [12]. Graph-Based semi-supervised learning methods which define a undirected graph in which the nodes of the graph are the data instances in the data set and the edges between the instances reflect the similarity of two data examples. As indicated in [12], “many graph-based methods can be viewed as estimating a classification function  $f$  on the graph”. In [2], the semi-supervised learning method is formulated as a graph min-cut problem, however the method can only give the hard class labels [12]. [13] address the above problem by using harmonic functions to characterize the Gaussian random field. In this paper, we proposed to use the graph-based methods to do classification for image.

In this paper, we propose to use the classical graph-based semi-supervised method: *Gaussian Fields Approach* [13] to do classification for microarray gene expression data. The rest of this paper is organized as follows: section 2 introduces the algorithm, section 3 shows the setup of our experiments and reports the results. Finally conclusions are made in section 4.

## 2. Graph-based semi-supervised learning

In this paper, we proposed to use the graph-based semi-supervised classification algorithm proposed in [13] which use the Gaussian random field to do semi-

supervised classification. In which the mean of the field is characterized in terms of harmonic functions. The method has two main steps: **construct the graph** and **classification**. We now describe the detail of the two steps of the algorithm:

Then we will now first introduce some important notations:  $\mathcal{X} = \{x_1, \dots, x_l, x_{l+1}, \dots, x_n\}$  represents a set of  $n$  microarray data objects. The first  $l$  points  $x_i \in X$  ( $i \leq l$ ) are labeled and the remaining points  $x_u \in X$  ( $l + 1 \leq u \leq n$ ) are unlabeled. And  $y$  is the class label set.

### 1. Graph Construction:

In the Graph-Based semi-supervised learning algorithms described in [13], the method first define a undirected graph  $W$  on the whole data set. In the graph  $W$ , the nodes are data instances in the graph and the edges are the strength of two data instances in the graph. Then the method in [13] construct a  $k$  nearest neighbors graph with the Gaussian function of Euclidean distance to weight the edges:

$$W_{ij} = \begin{cases} \exp\left(-\frac{(x_i - x_j)^T(x_i + x_j)}{\sigma^2}\right) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

where the  $i \sim j$  denotes that node  $i$  and  $j$  has an edge between them. Then we can denote the graph as the following:

$$W = \begin{pmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{pmatrix}$$

where  $W_{ll}$  denotes the weight of the edges between two labeled microarray data instances and  $W_{uu}$  denotes the weight of the edges between two unlabeled microarray data instances in the graph.  $W_{lu}$  denotes the weights of the edges from the labeled microarray points to the unlabeled microarray points in the graph and  $W_{ul}$  denotes the weights of the edges from the unlabeled microarray points to the microarray labeled points in the graph. And all the weights in  $W$  are weighted by  $w_{ij}$ .

### 2. Graph-based Semi-supervised Classification:

With the graph constructed above, the Graph-Based methods can be viewed as estimating a function  $f$  on the graph  $W$  [13].  $f$  is a real-value class assign matrix which assign the class labels. In graph-based semi-supervised learning, the  $f$  should satisfy two things: (1) the value of  $f$  should be close with the class labels of the labeled data samples [12], then the regularizer in the

graph can be expressed as the following equation:

$$\sum_{i \in L} (f_i - y_i)^2 \quad (1)$$

and then the  $f$  should satisfy the second condition: (2) the  $f$  should be smooth enough on the whole graph [12]. Then the smooth enough graph with the regularizer can be denote as the following:

$$\frac{1}{2} \sum_{ij} w_{ij} (f_i - f_j)^2 \quad (2)$$

Then the classification problem of the graph-based semi-supervised learning can be viewed as the combination of the above two regularizer:

$$\begin{aligned} \sum_{i \in L} (f_i - y_i)^2 + \frac{1}{2} \sum_{ij} w_{ij} (f_i - f_j)^2 \\ = (f - y)^T (f - y) + \frac{1}{2} f^T \Delta f \end{aligned} \quad (3)$$

in which  $\Delta$  is called the graph *Laplacian* [13]:

$$\Delta = D - W \quad (4)$$

Where  $D = \text{diag}(d_i)$ ,  $d_i = \sum_{ij} w_{ij}$  [13]. Then [13] express the function  $f$  as:

$$f = \begin{pmatrix} f_l \\ f_u \end{pmatrix} \quad (5)$$

and then the  $f$  are expressed as:

$$f = P f \quad (6)$$

Where the  $P = D^{-1}W$ ,  $f_u$  defines the label of the unlabeled microarray data examples and  $f_l$  defines the values of the labeled examples. Then [13] get the solution as:

$$f_u = (I - P_{uu})^{-1} P_{ul} f_l \quad (7)$$

Where  $I$  is the identity matrix. Then the classification results are get from the  $f_u$ .

## 3. Experiments

In this section, we will report our results on four public microarray data sets and give some discussion.

### 3.1. Microarray Data Sets

The proposed method has been tested on four public available microarray datasets. The details of the data sets used in our experiments are summarized as below:

**The Leukemia dataset** [3]: contains two types of acute leukemia: 47 acute lymphoblastic leukemia and 25 acute myeloid leukemia. And the gene expression comes from 6,817 human genes.

**The Colon dataset** [1]: contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors and 22 normal.

**The Lung Cancer dataset** [4]: There are 181 tissue samples (31 MPM and 150 ADCA) in the data set. Each data sample is described by 12533 genes.

**The Ovarian dataset** [9]: includes 91 controls (Normal) and 162 ovarian cancers with around 15155 genes.

### 3.2. Experiments setup

In this paper, we use the *Gaussian Field Approach* to do classification for microarray data. We also compare the *Gaussian Field Approach* (GF) with some other methods widely used for microarray data classification: the KNN algorithm, the J48 Decision Tree algorithm and the Random Forest (RF) algorithm. In all the experiments, the entire process is repeated over 20 times with random data partitions.

In the first experiment, we consider the case when there exists very few labeled data examples for train. We split 5% to 90% data per class to train and the rest are used to test. We perform the algorithms on the original data set to compare the results.

In the second experiment, we consider the case when there are very few labeled training examples and we do dimensionality reduction for the data, but traditional supervised dimensionality reduction methods such as LDA and information gain ratio cannot work under this condition, but sometimes for fast computing, we also need to reduce the dimension of the microarray data, so we use the unsupervised method PCA to do dimensionality reduction for the data set. Then we perform the algorithms on the preprocessed data set.

In the third experiment, because some former research [6] have shown that supervised feature selection can effectively enhance the performance of the classification accuracy for microarray data. So, we use the information gain ratio for gene selection. For our experiments, we set the number of genes selected as 50.

In the experiments, we use the *Gaussian Fields Approach* (GF) [13] as the base classifier of our algorithm. For each data set, an initial undirected edge graph  $W$  was constructed as the base line by making a symmetrical connection between each point and its  $k$ -nearest neighbors as measured by Euclidean separation in the input space, with  $k$  set either to 2–4. Weights were then set for each edge according to the function  $w_{ij} = \exp(-s_{ij}^2/\sigma^2)$  of edge length  $s_{ij}$ , with  $\sigma$  set either to

10000. Now we report average accuracy of the following methods on unlabeled data: *Gaussian fields approach* (GF),  $k$ -Nearest Neighbor classifier (KNN), J48 Decision Tree (J48) classifier and the Random Forests classifier (RF). In KNN, the number of nearest neighbors  $k$  was set to 2.

### 3.3. Original datasets results

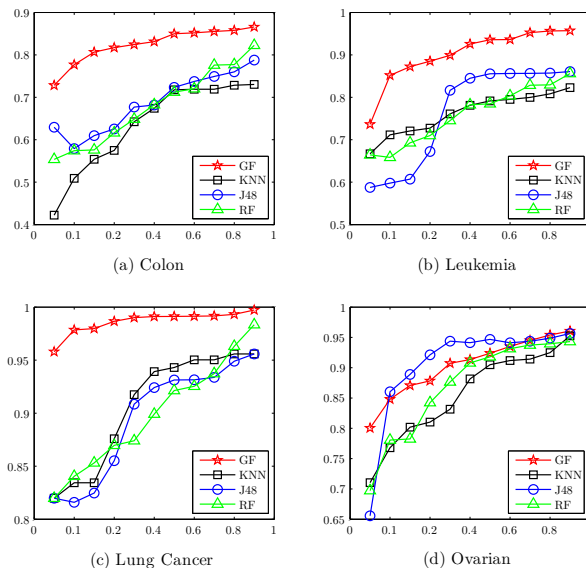
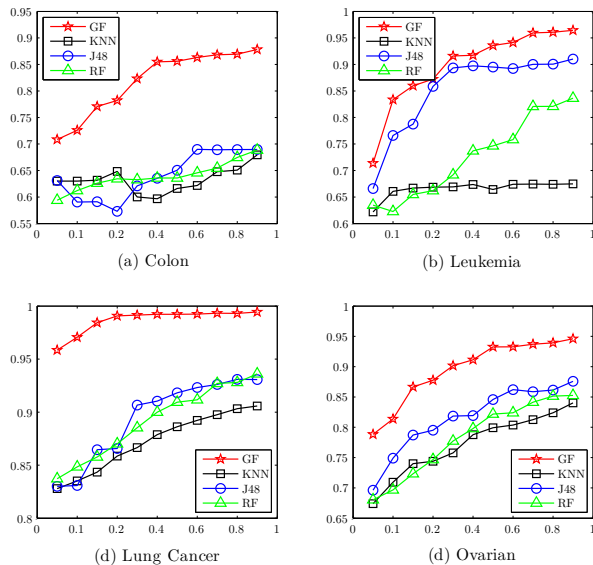


Figure 1. Result on original dataset.

In this experiment, we use the original data set to test all the algorithms. On all data sets tested in our experiments, the results are summarized in Fig. 1. In all the figures, we can find that the *Gaussian Fields Approach* leads to performance that is much better than the other supervised algorithms. When the labeled data examples are few, the supervised methods perform bad and keep a low classification accuracy, but the *Gaussian Fields Approach* keeps a high classification accuracy when the labeled examples are few. Even if the training data are sufficient, the *Gaussian Fields Approach* also outperforms the supervised methods.

### 3.4. PCA preprocessed datasets results

In this experiment, we use the PCA preprocessed data set to test all the algorithms. On all data sets tested in our experiments, the results are summarized in Fig. 2. In all the figures, we can see that the *Gaussian Fields Approach* still leads to performance that is much better than the other supervised algorithms. With PCA pre-processing, the results of the algorithms have not been



**Figure 2. Classification Result on PCA preprocessed data set.**

enhanced compared with the original data set, but it is clear that in this case, the *Gaussian Fields Approach* is still much better than other methods.

### 3.5. Results with gene selection

In this experiment, we firstly use the supervised feature selection method: Info Gain Ratio for gene selection with the number of selected genes set to 50. Then we perform the *Gaussian Fields Approach* and other methods on the preprocessed data sets. The parameter in GF was optimized to the best. We split 90% data to train the classifier. The whole process is repeated 50 times with random data partitions. The results have been summarized in Table 1. From the results we can see clearly that the *Gaussian Fields Approach* is also very effective in the supervised case. The method outperforms the other methods and in all cases. The results have indicate that in supervised case, the *Gaussian Fields Approach* is also better than traditional supervised methods and holds higher classification accuracy.

## 4. Conclusions

In this paper, we proposed to use the semi-supervised learning algorithm *Gaussian Fields Approach* to do classification for microarray data to deal with the problem of lacking training examples. Experiments on pub-

Data set	GF	KNN	J48	RF
Colon	<b>92.01</b>	83.52	89.58	88.21
Lung Cancer	<b>99.17</b>	98.90	96.11	99.09
Leukemia	<b>98.29</b>	94.66	85.79	95.70
Ovarian	<b>99.08</b>	98.73	97.42	98.79

**Table 1. Results with gene selection (%)**

lic microarray data sets have demonstrated the effectiveness of the method. Using the *Gaussian Fields Approach* to do classification for microarray data can not only help to archive very high classification accuracy but also is much more stable than the supervised methods especially when the labeled examples are very few. The *Gaussian Fields Approach* also outperforms the supervised methods when the training data are sufficient.

## References

- [1] U. ALON and N. B. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*.
- [2] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. *ICML*, 2001.
- [3] T. R. Golub, D. K. Slonim, and P. T. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 1999.
- [4] G. Gordon, R. Jensen, and L. H. et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma. *Cancer Research*.
- [5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002.
- [6] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard. A comparative study of classification methods for microarray data analysis. *AusDM*, 2003.
- [7] J. Li and H. Liu. Ensembles of cascading trees. *ICDM*, 2003.
- [8] J. Li, H. Liu, S.-K. Ng, and L. Wong. Discovery of significant rules for classifying cancer diagnosis data. *ECCB*, 2003.
- [9] E. PetricoinIII and A. A. et al. Use of proteomic patterns in serum to identify ovarian cancer. *The lancet* 359.
- [10] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 1995.
- [11] A. C. Tan and D. Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2003.
- [12] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.
- [13] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. *ICML*, 2003.