Analysis of a Mixed Inter Programming Model for Microarray Data Classification

S. Shenbaga Ezhil¹ and C. Vijayalakshmi²

¹ Sathyabama University, Chennai – 119, India
² Dept. of Mathematics, VIT University, Chennai, India

¹shenbaga_ezhil@rediffmail.com
²vijusesha2002@yahoo.co.in

Copyright © 2013 S. Shenbaga Ezhil and C. Vijayalakshmi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Gene expression profiling has been successfully used to identity potential cancer diagnosis and therapeutic target in the past few years. Since there are often large number of potential genes, an important issue is to find a small subset of genes that are differentially expressed between different clinical outcomes (or) related subset to cancer patients survival time, and thus can be used to build prognosis and predictors.

Keywords: diagnosis, prognosis, gene, subset, cancer

1 Introduction

The rapid advances in microarray technology enable biologist to measure the expression of level of thousand (or) ten thousand genes simultaneously [1]. The initial information from microarray experiments goes through various data processing steps including image processing, quality control and normalization [2]. Microarray technologies give life sciences researchers the opportunity to simultaneously measure the thousands of gene expression levels under different conditions or coming from different cell lines. Therefore selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification [3, 4].

In this paper, a novel gene selection method based on mixed integer optimization is applied with the aim of identifying a small subset of genes which may be useful for discriminating between normal and tumor tissues.
2. Gene Selection for Microarray Data Sets

In this we consider the data base of colon cancer study. The colon cancer dataset was originally analyzed by Alon et al. (1999). This dataset contains expression levels of 2000 genes with highest minimal intensity across 45 tumor and 20 normal colon tissues. Pre filtering according to ‘t’ test left 859 genes that were significantly differentially expressed (p value ≤ 0.1) for further gene selection.

We divide a good subset into three parts:
(i) Features that are absolutely necessary for classification.
(ii) The features that can be chosen based on properties of data set.
(iii) Features that can be found using a classification algorithm as evaluation measures. We name them as dependent features.

3. The Mixed Integer Programming Model

Let \( \mathcal{F} = \{1, 2, ..., f\} \) denote the set of the indices of the different genes in \( G \) and \( g^-_i \) and \( g^+_i \), \( i \in \mathcal{F} \), the vectors of the expression levels of gene \( i \) for normal and tumor tissues, respectively. For each gene \( i \in \mathcal{F} \), define two binary variables

\[
\begin{align*}
\z_i^- &= \begin{cases} 
0, & \text{if the profile } g^-_i \text{ properly identifies normal tissues} \\
1, & \text{otherwise.}
\end{cases} \\
\z_i^+ &= \begin{cases} 
0, & \text{if the profile } g^+_i \text{ properly identifies tumor tissues} \\
1, & \text{otherwise.}
\end{cases}
\end{align*}
\]

indicating if the expression profiles of \( i \) correctly characterize the state specified by the class value of the corresponding tissues. If the gene discriminate function takes the form \( w'g - b = 0 \), where \( w \) defines the orientation of the hyper plane is the \( s \)-dimensional space \( \mathbb{R}^s \) and \( b \) its offset from the origin, the following mixed integer optimization problem can be formulated.

\[
\begin{align*}
\min & \sum_{i=1}^{\infty} (\z_i^- + \z_i^+) \\
\text{Subject to the condition} & \\
& w'g^-_i - b \geq -Q \z_i^+, \quad i \in \mathcal{F}, \\
& w'g^+_i - b < Q \z_i^-, \quad i \in \mathcal{F}, \\
& \z, \z^+ \text{ are binaries, } w, b \text{ free},
\end{align*}
\]

Where \( Q \) is a sufficiently large constant scalar, and constraints (3) and (4) set the values of the binary variables \( \z_i^- \) and \( \z_i^+, i \in \mathcal{F} \).

From the solution of the problem (A), obtained by a truncated branch-and-bound procedure, it is possible to find a set of genes useful for discriminating between normal and tumor tissues. In particular, for each gene \( i \in \mathcal{F} \) the following measure, termed classification score, is computed.
Analysis of a mixed inter programming model

\[ \text{CS}_i = \delta_i^- + \delta_i^+, \]  

(5)

where \( \delta_i^- \) and \( \delta_i^+ \) represent the Euclidean distances of patterns \( g_i^- \) and \( g_i^+ \) from the separating hyper plane.

Summary of microarray data

<table>
<thead>
<tr>
<th>Title</th>
<th>No. of genes</th>
<th>No. of samples</th>
<th>No. of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>12800</td>
<td>138</td>
<td>2</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>25491</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>Colon Rectum</td>
<td>2500</td>
<td>71</td>
<td>2</td>
</tr>
<tr>
<td>Lung Cancer Ontario</td>
<td>2950</td>
<td>51</td>
<td>3</td>
</tr>
</tbody>
</table>

We then apply C.45 and Naïve Bays respectively on each original data set and each newly obtained data set only containing the selected genes, and obtain the overall classification accuracy by leave-one-out cross validation.

The EOD field is composed of five fields including the EOD (Extent of Death) code. These fields are size of tumor, number of positive nodes, number of nodes and number of primaries unlike [6] we have included three fields

1. Survival Time Recode (STR)
2. Vital Status Record (VSR)
3. Cause of Death (COD)

The STR field ranges from 0 to 120 months in the SEER data base.

4. Results and Discussion

The following table reports the running time for each feature selection algorithm. The number of genes selected by each feature selection algorithm. We can see that half H FW on averages selects the smallest number of genes. Also it reports the leave-one-out accuracy by C 4.5.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>( \text{HFW}_{C4.5} )</th>
<th>Half-HFW</th>
<th>FCBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running Tissue (s) for each feature selection Algorithm</td>
<td>59.83</td>
<td>0.91</td>
<td>1.14</td>
</tr>
<tr>
<td>Number of genes selected by each feature selection Algorithm</td>
<td>11</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Validation accuracy of C4.5 on selected genes for each feature selection method</td>
<td>90.32</td>
<td>85.48</td>
<td>88.71</td>
</tr>
<tr>
<td>Leave one out cross validation accuracy selected genes for each feature selection method</td>
<td>75.81</td>
<td>90.32</td>
<td>77.42</td>
</tr>
</tbody>
</table>
5. Conclusion

Hence the classification of the data sets is consistently based on a very small number of genes, compared with the original number of features describing the sample tissues. Extensive experiments on microarray data have demonstrated the superior performance of our approach. The simulation studies illustrate the problem of statistical redundancy and validate the proposed statistic and gene classification method.

References


Received: January 17, 2013