Gene Selection Using Mixed Integer Programming Problem

S. Shenbaga Ezhil
Associate Professor, Department of Mathematics
AMET University, Chennai, Tamil Nadu, India.

1. INTRODUCTION

The rapid advances in microarray technology enable biologist to measure the expression levels at thousand (or) ten thousand genes simultaneously. The initial information from microarray experiments goes through various data processing steps including image processing, quality control and normalization. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account three variables, namely, Survival Time Recode (STR), Vital Status Record (VSR) and Cause of Death (COD).

Keywords: Microarray technology, Gene expression, Survival Time Recode

2. MATHEMATICAL MODEL OF MIXED INTEGER PROGRAMMING IN GENE SELECTION

Consider an expression matrix $S_{m \times n}$ in which rows represents tissues of patients and columns represent the genes whose expression has been measured under the same experimental conditions. The value of $S$ at rows $i$ and column $j$ is therefore provided by the expression level of gene $j$ observed for patient $i$. Let $X_i \in \mathbb{R}^n$ be the vector of expression values corresponding to row $i$ in the matrix $S$.

That is for each example $x_i$, $i \in M = \{1, 2, ..., m\}$, it is known the value $y_i$ of a categorical variable indicating the class associate $X_i$. In this $X_i$ will attain binary classification problems with only two different values. Hence without loss of
generality we will set \( y_i = -1 \) of example \( x_i \) represents a normal or tumor sample tissue respectively.

\[
\begin{array}{ccc|c}
\text{gene 1} & \text{gene 2} & \text{gene n} & Y \\
\hline
\text{tissue 1} & x_{11} & x_{12} & x_{1n} & -1 & \text{normal} \\
\text{tissue 2} & x_{21} & x_{22} & x_{2n} & 1 & \text{tumor} \\
\text{tissue 3} & x_{31} & x_{32} & x_{3n} & -1 & \text{Normal} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

Expression matrix \( S \) class

**Figure 1:** Gene Expression Matrix and Class Value

Under these assumptions solving cancer microarray data classification problem requires to find a function \( f(x) : \mathbb{R}^n \leftrightarrow \{-1, 1\} \) which optimally describes the relationship between the sample tissues and their class values in order to predict the diagnostic category. The mixed integer optimization model represents a classification method itself, searching for the best relationship between the gene expression patterns and the class values of the sample tissues.

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

\[
z_i^- = \begin{cases} 
0, & \text{if the profile } g_i^- \text{ properly identifies normal tissues} \\
1, & \text{otherwise.} 
\end{cases} \tag{1}
\]

\[
z_i^+ = \begin{cases} 
0, & \text{if the profile } g_i^+ \text{ properly identifies tumor tissues} \\
1, & \text{otherwise.} 
\end{cases} \tag{2}
\]

To indicate whether the expression profiles \( i \) correctly characterize the state specified by the class value of the corresponding tissue. If the gene discriminate function takes the form \( w^T g - b = 0 \), 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.
Gene Selection Using Mixed Integer Programming Problem

\[
\min \sum_{i=1}^{\infty} (z_i^- + z_i^+) \quad (3)
\]

Subject to the condition
\[
w'g_i^+ - b \geq -Q z_i^+, \quad i \in F, \quad (4)
\]
\[
w'g_i^- - b < Q z_i^-, \quad i \in F, \quad (5)
\]

\(z-, z+\) are binaries, \(w, b\) free. Where \(Q\) is a sufficiently large constant scalar, and constraints \((8.4)\) and \((8.5)\) set the values of the binary variables \(z_i^+\) and \(z_i^-\), \(i \in F\).

From the solution of the problem \((3)\), 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 \(i \in F\) the following measure, termed classification score, is computed.

\[
CS_i = \delta^-_i + \delta^+_i, \quad (6)
\]

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

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 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.

3. ALGORITHM

Setting the survivability dependent variable for 40 months threshold

if STR = 40 months and VSR is alive then

the record is pre-classified as “survived”.

else if STR < 60 months and COD is Colon Rectum cancer,
then the record is pre-classified as “not survived”

else

ignore the record

end if

In the above approach, the ignored records corresponding to those patient that have an STR less than 40 months but the cause of death is not due to colon rectum cancer.

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. Note that half H FW on and average select the smallest number of genes. Also it reports the leave-one-out accuracy by C 4.5.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>HFW&lt;sub&gt;C4.5&lt;/sub&gt;</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. The experimental results show that our model does not include records with missing data. Hence in the survival time prediction of cancer is seriously low than respiratory
cancer. The survival time in terms of one year and then classifying using the aforementioned data mining techniques in comparison with previous algorithm. Hence our model can be considered as an appropriate model for prediction of survival of Cancer.

REFERENCES


