Microarrays for medical prognosis

Microarrays measure expression of tens of thousands of genes from an individual in a single experiment. Typical microarray data:

<table>
<thead>
<tr>
<th>gene 1</th>
<th>gene 2</th>
<th>...</th>
<th>gene N</th>
<th>Pathology (class label)</th>
</tr>
</thead>
<tbody>
<tr>
<td>g11</td>
<td>g12</td>
<td>...</td>
<td>g1N</td>
<td>y1</td>
</tr>
<tr>
<td>g21</td>
<td>g22</td>
<td>...</td>
<td>g2N</td>
<td>y2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>gM1</td>
<td>gM2</td>
<td>...</td>
<td>gMN</td>
<td>yM</td>
</tr>
</tbody>
</table>

In our case, samples are observations, genes are features. Given the high cost of this technology, only some tens of experiments may be led so N>>M, which is statistically a hard context for knowledge inference.

Prognosis: Feature Selection and Classification

Double aim:
1. Identify a small subset of genes as pathology risks factors to launch a deeper medical research project and to evaluate treatment development feasibility.
2. Train a classifier based on these genes to be able to make a prognosis about the given pathology.

Semi-Supervised Feature Selection (SSFS)

Motivations

1. Expertise incorporation: The SSFS algorithm may be used to incorporate prior knowledge from field experts. For example, in a microarray experiment, a biologist may already know several genes to be relevant, and others to be irrelevant. Those will get respectively a high or low prior during feature selection.
2. This algorithm may also be useful to increase stability in feature selection. Stability is a desired property since the choice of the relevant dimensions in a problem should not be influenced too much by varying the data sampling.
3. Transfer Learning: To allow the partial transfer of knowledge across several related datasets (e.g., breast cancer datasets). In an SSFS procedure, assign a higher prior to some features already selected on a different dataset.

SSFS Problem

A constrained zero-norm minimization approximation problem is described in [3]:

\[
\min_{x \in \mathbb{R}^n} \sum_{j=1}^{N} \left( \min_{\mu_j} \|x \|^2 \right) \geq 1
\]

Subject to: \( y_i(x + \beta) \geq 1 \)

This problem is elegantly solved by an iterative algorithm called LFAROM. Weight values corresponding to less useful features are rapidly vanishing. The selection process can stop at convergence, or when sufficient sparsity is obtained.

We propose to add a prior relevance \( \beta \) to each feature. The more relevant the feature, the higher the \( \beta \), the higher the prior.

\[
\min_{x \in \mathbb{R}^n} \sum_{j=1}^{N} \beta_j \|x \|^2 \geq 1
\]

Subject to: \( y_i(x + \beta) \geq 1 \)

Franke and Wolfe's constrained gradient descent technique may be applied:

1. At step \( k = 0 \), initialize \( \beta \) at \( \beta \geq 0 \), \( \forall \gamma \in [1,n] \) according to prior information. Set \( w_{0,j} = \beta_j \), \( \forall \gamma \in [1,n] \).
2. Solve \( \min_{\mu_j} \|x \|^2 \) subject to \( y_i(x + \mu_j) \geq 1 \)
3. Let \( W \) be the solution of step 2, set \( w_{k,j} = w_{k-1,j} + w + \beta \)
4. Go to step 2 until convergence

Note: The * is the component wise product.

Experimental Design

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Methods 3, 4, 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%</td>
<td>20%</td>
<td>-L2AROM (All), 20%</td>
</tr>
<tr>
<td>SSFS, All ( \beta ) L (L2AROM)</td>
<td>1 Selection of 100 features</td>
<td></td>
</tr>
<tr>
<td>- L2AROM (All), 20%</td>
<td>20% of 80% from 80%</td>
<td></td>
</tr>
<tr>
<td>- 20 Selections of 100 features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 20 Samples of 80% from 100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 20 Selections of 100 features</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance

Since datasets are unbalanced, Balanced Classification Rate (BCR) is used instead of simple classification rate.

Stability

The stability is, in this case, a measure of robustness of the selected dimensions while varying the sampling. Kuncieva Index [7]:

\[
\text{Stab} = \frac{2}{K - 2} \left( \frac{1}{K} \sum_{i=1}^{K} \left( \frac{1}{K} \sum_{j=1}^{K} \frac{1}{K} \sum_{i=1}^{K} |y_i - y_{ij}| \right)^2 \right)
\]

where \( K \) is the number of selection rounds (here, \( K = 20 \)), \( y_i \) is a signature (set of selected dimensions), \( I \) is the size of the intersection of two signatures, \( s \) is the size of the signatures (here, \( s = 100 \)) and \( \gamma \) is the number of available features.

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>L2AROM (80%)</th>
<th>SSFS</th>
<th>L2AROM</th>
<th>RFE</th>
<th>t-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alon Stab</td>
<td>55.07</td>
<td>50.56</td>
<td>65.00</td>
<td>77.08</td>
<td>78.22</td>
</tr>
<tr>
<td>Golub Stab</td>
<td>50.99</td>
<td>43.21</td>
<td>61.50</td>
<td>78.02</td>
<td>78.39</td>
</tr>
<tr>
<td>Prostate Stab</td>
<td>58.70</td>
<td>87.26</td>
<td>52.52</td>
<td>77.74</td>
<td>79.73</td>
</tr>
<tr>
<td>Alon BCR</td>
<td>88.01</td>
<td>87.74</td>
<td>71.96</td>
<td>74.11</td>
<td>81.29</td>
</tr>
<tr>
<td>Golub BCR</td>
<td>95.65</td>
<td>91.05</td>
<td>95.5</td>
<td>78.56</td>
<td>98.18</td>
</tr>
<tr>
<td>Prostate BCR</td>
<td>93.64</td>
<td>90.66</td>
<td>97.84</td>
<td>78.23</td>
<td>91.95</td>
</tr>
</tbody>
</table>

Conclusions

1. SSFS largely succeeds at improving feature selection stability.
2. SSFS is more stable than RFE, which is known to be stable [6].
3. SSFS provides a good tradeoff between stability and BCR.
4. SSFS allows to include prior knowledge on important dimensions while letting the feature selection procedure depart from it.
5. Further work will investigate the effect of a finer tuning of the \( \beta \) values. For example, higher values still increase stability at the cost of a significant BCR reduction (results not shown).
6. Here, prior knowledge was “simulated” by using features selected on 20% of the data. Real knowledge incorporation will be tested.

Data sets

Colon Cancer [2]
Samples (patients): 62 (40 tumor vs. 22 normal tissues)
Features (genes): 2000 (after quality filtering)
Classes: 2
Proportion: 64.5% tumor and 35.5% normal tissues

Leukemia [3]
Samples (patients): 72 (25 AML vs. 47 ALL tissues)
Features (genes): 7129
Classes: 2
Proportion: 43.7% AML and 56.3% ALL tissues

Prostate Cancer [4]
Samples (patients): 102 (52 vs.50)
Features (genes): 6033
Classes: 2
Proportion: 51% and 49%

Some References