**Microarrays for medical prognosis**

Microarrays measure expression of tens of thousands 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>sample 1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>sample 2</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>sample M</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

In our case, samples are samples, genes are features.

Given the high cost of this technology, only some tens of experiments may be fed 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 risk factors for further medical research or to evaluate treatment efficacy.
2. Train a classifier based on these genes to design a prognosis kit for the given pathology.

**Partially Supervised Feature Selection**

**Motivations**

1. **Expert knowledge**: The PSFS algorithm may be used to incorporate prior knowledge from field experts. For example, in a microarray experiment, a biologist may know that some genes are likely more relevant. These will get a higher prior during optimization.
2. **This technique increases 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. **Increasing classification performances**: In some cases, this technique also increases classification performances.

**PSFS Problem**

Constrained zero-norm minimization approximation problem [1]:

\[
\min_{w} \sum_{j=1}^{n} \ln(\epsilon + |w_j|) \geq 1
\]

Subject to: \( y_i w_i x_{i+1} + b \geq 1 \)

Elegantly solved by an iterative algorithm called L2AROM. Weight values corresponding to less useful features are rapidly vanishing. The selection process can be stopped at convergence, or when sufficient sparsity is obtained.

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

\[
\min_{w} \sum_{j=1}^{n} \ln(\epsilon + |w_j|) \geq 1
\]

Subject to: \( y_i w_i x_{i+1} + b \geq 1 \)

A constrained gradient descent technique is applied:
1. At step \( k = 0 \), initialize \( \beta_k, j \geq 1, \forall j \in \{1, n\} \) according to prior information. Set \( \omega_{0j} = \beta_j, \forall j \in \{1, n\} \).
2. Solve \( \min_{w} |w_k| \) subject to relaxed margin constraints:

\[
\min_{w} w_k x_{i+1} + b \geq 1
\]

3. Let \( w_{k+1} \) be the solution of step 2, set \( w_{0j} = w_{k+1} \).
4. Go to step 2 until convergence.

Note: \( \beta \) denotes the component-wise product.

**Experimental Design**

Protocol 1: Real Knowledge

- 200 Splits 90%-10%:
  1. Feature Selection on 90% with \( \beta = 10 \) for a priori favored features and \( \beta = 1 \) otherwise
  2. Train SVM on 90% on selected features
  3. Test on 10%
  4. Average BCR and Stability

Protocol 2: Simulated Knowledge 20%

- 10 Splits 80%-20%:
  1. Select 50 features on 20%
  2. Use 10% and repeat 20
  3. Train SVM on 90% on selected features
  4. Test on 10%
  5. Average BCR and Stability

**Comparison of PS-12AROM, 12-AROM [1], RFE [8], Gobol Index [3], and Random Selection**

**Classification Performances**

Unbalanced datasets: Balanced Classification Rate (BCR) instead of Accuracy

\[
BCR = \frac{TP}{TP + TN}
\]

Stab = \( \frac{2}{K(K-1)} \sum_{k=1}^{K} \sum_{j=1}^{K} \left( 1 - \frac{Sg(j)}{Sg(k)} \right)^{\frac{1}{2}} \)

where \( K \) is the number of selection rounds (here, \( K = 200 \) or \( 200 \)). \( Sg \) is a signature (set of selected dimensions). \( I \) is the size of the intersection of two signatures, \( \alpha \) is the size of the signatures and \( \beta \) is the total number of features.

**Real Prior Knowledge**

**Simulated Knowledge**

**Conclusions**

1. PSFS allows to include prior knowledge on a priori important dimensions while letting the feature selection procedure depart from it.
2. PSFS naturally extends AROM methods [1].
3. Partial Supervision increases stability of selected features with respect to sampling variations.
4. Multivariate method: supervision of few dimensions influences the selection of the other ones.

**Data sets**

**Colon Cancer [2]**

- Samples (patients): 62 (40 tumor vs. 22 normal tissues)
- Features (genes): 2000 (after non-specific filtering)
- 2 Classes: 64.5% tumor and 35.5% normal tissues

**Leukemia [3]**

- Samples (patients): 72 (25 AML vs. 47 ALL tissues)
- Features (genes): 7129
- 2 Classes: 34.7% AML and 65.3% ALL tissues

**Prostate Cancer [4]**

- Samples (patients): 102 (52 tumor vs.50 normal tissues)
- Features (genes): 6033
- 2 Classes: 51% tumor and 49% normal tissue

**DLBCL [5]**

- Samples (patients): 77 (58 tissue 1 vs. 19 tissue 2 features): 7129
- 2 Classes: 75% tumor and 25% normal tissue

**Main References**


