The Painter’s Feature Selection for Gene Expression Data

Daniele Apiletti, Elena Baralis, Giulia Bruno, Alessandro Fiori

Lyon, 23-26 August 2007
Introduction

- Feature selection
  - identifies a minimum set of relevant features
  - is applied before a learning algorithm
  - reduces computation costs
  - increases the speed up of learning process
  - increases the model interpretability
  - improves the classification accuracy performance
Framework

Microarray labeled data → Filter → Feature Selection

Gene Rank → Model Building → Model

Microarray unlabeled data → Classification → labels
Feature selection

- From Painter’s Algorithm in Computer Graphics to paint shadows
- Overlap score to measure:
  - common expression intervals in different classes
  - gaps between expression intervals among classes
- Bonus to genes with
  - largest gaps
  - few overlapping classes

\[
\text{overlapsco} = \frac{\sum_{i=1}^{n} c_i w_i}{w_{\text{tot}}}
\]
Example

\[ \text{overlapscore} = \frac{(1 \times 6) + (1 \times 4)}{11} = \frac{10}{11} \]

0 ≤ \text{overlapscore} ≤ 1 \implies \text{no overlapping}

\[ \text{overlapscore} = \frac{(1 \times 1) + (2 \times 4) + (1 \times 1)}{6} = \frac{10}{6} \]

1 < \text{overlapscore} < N_{\text{classes}} \implies \text{overlapping}
Filter

- **Reason:**
  - noisy data
  - outliers presence

\[ I_{ij} = \mu_d \pm \sum \sigma_d \]

\[ \mu_d = \frac{1}{d_{tot}} \sum_{i=1}^{n} d_i e_i \]

\[ \sigma_d = \sqrt{\frac{1}{d_{tot}} \sum_{i=1}^{n} d_i \epsilon_i - \mu_d^2} \]
Weighted interval

\[ r = \sigma \]

\[ \mu_d \pm 2\sigma_d \]

gene expression value
Experimental design

- 10-cross validation
- SVM kernel: Crammer and Singer (CS)
  - degree: 1
  - cost: 100
- Method compared:
  - analysis of variance (ANOVA)
  - signal-to-noise ratio in OVO (OVO)
  - signal-to-noise ratio in OVR (OVR)
  - ratio of variables between categories to within categories sum of squares (BW)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Patients</th>
<th>Genes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumors9</td>
<td>60</td>
<td>5727</td>
<td>9</td>
</tr>
<tr>
<td>Brain1</td>
<td>90</td>
<td>5921</td>
<td>5</td>
</tr>
<tr>
<td>Brain2</td>
<td>60</td>
<td>10364</td>
<td>4</td>
</tr>
</tbody>
</table>
Experimental results (1)

- 200 genes selected

<table>
<thead>
<tr>
<th></th>
<th>Painter</th>
<th>OVO</th>
<th>BW</th>
<th>ANOVA</th>
<th>OVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain1</td>
<td>86.7 8.8</td>
<td>88.9 7.4</td>
<td>86.7 9.1</td>
<td>88.89 6.3</td>
<td>87.8 7.0</td>
</tr>
<tr>
<td>Brain2</td>
<td>70.7 20.0</td>
<td>64.7 12.4</td>
<td>64.0 14.5</td>
<td>62.2 22.4</td>
<td>61.8 15.7</td>
</tr>
<tr>
<td>Tumors9</td>
<td>71.0 24.4</td>
<td>65.5 15.1</td>
<td>64.2 16.6</td>
<td>69.9 18.4</td>
<td>66.8 20.2</td>
</tr>
<tr>
<td>Average</td>
<td>76.1 17.6</td>
<td>73.0 11.6</td>
<td>71.6 13.4</td>
<td>73.7 15.7</td>
<td>72.1 14.3</td>
</tr>
</tbody>
</table>
Experimental results (2)

- trend on Brain2 dataset

![Graph showing trend on Brain2 dataset with different methods like Painter, OVO, OVR, BW, and ANOVA.](chart)
Conclusion

- **New method:**
  - multi-class approach
  - based on new criterion of gene relevance
  - self adaptation to the datasets distribution
  - density based filter
  - smoothing outliers effect

- **Results**
  - robustness of the algorithm
  - can be applied to any dataset with continuous valued features

- **Future work:**
  - investigation of features groups with the same discriminating power
  - comparison with more feature selection techniques on other datasets
Thanks for the attention!