Improved gene selection in microarrays by combining clustering and statistical techniques

Jochen Jäger
University of Washington
Department of Computer Science

Advisors:
Larry Ruzzo
Rimli Sengupta
Motivation

- Think of a complicated question:
  - Will it be sunny tomorrow?
  - How can you answer it correctly if you DO NOT know the answer?
  - Ask around or better, make a poll
Majority vote

- Student: I heard it is supposed to be sunny
- Weather.com: partly cloudy with scattered showers
- Yourself: Considering the past few days and looking outside I would guess it will rain
- TV: partly sunny

- Result: 2 (sunny) : 2 (not sunny)
- Better: Use weights
- Idea: remove redundant answers as well
Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work
Biological task

- Find informative genes
- (e.g. genes which can discriminate between cancer and normal)
- Use series of microarrays
- Compare results from different tissues
Microarrays

DNA

select genes

spot genes

extract cDNA

label cDNA

Annealing phase
Outline

• Motivating example
• Biological background
• Problem statement
• Current solution
• Proposed attack
• Results
• Future work
Finding informative genes

- Microarrays from different tissues

![Images of microarrays showing cancerous and normal tissues](image-url)
Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work
Current solution

• Use a test statistic on all genes

\[
t = \frac{\bar{x}_{\text{tumor}} - \bar{x}_{\text{normal}}}{\sqrt{\frac{\sigma^2_{\text{tumor}}}{n_{\text{tumor}}} + \frac{\sigma^2_{\text{normal}}}{n_{\text{normal}}}}}
\]

<table>
<thead>
<tr>
<th>Gene</th>
<th>Tumor 1</th>
<th>Tumor 2</th>
<th>Tumor 3</th>
<th>Normal 1</th>
<th>Normal 2</th>
<th>Normal 3</th>
<th>t-test P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>80</td>
<td>72</td>
<td>85</td>
<td>50</td>
<td>44</td>
<td>15</td>
<td>0.0448836</td>
</tr>
<tr>
<td>B</td>
<td>80</td>
<td>72</td>
<td>85</td>
<td>50</td>
<td>44</td>
<td>51</td>
<td>0.0048027</td>
</tr>
<tr>
<td>C</td>
<td>71</td>
<td>53</td>
<td>62</td>
<td>57</td>
<td>64</td>
<td>70</td>
<td>0.8024078</td>
</tr>
</tbody>
</table>

• Rank them
• Select top k
Problem with current solution

- Each gene independently scored
- Top k ranking genes might be very similar and therefore no additional information gain
- Reason: genes in similar pathways probably all have very similar score
- What happens if several pathways involved in perturbation but one has main influence
- Possible to describe this pathway with fewer genes
Problem of redundancy

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Adenoma 1</th>
<th>Adenoma 2</th>
<th>Adenoma 3</th>
<th>Adenoma 4</th>
<th>Normal 1</th>
<th>Normal 2</th>
<th>Normal 3</th>
<th>Normal 4</th>
<th>t-test P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF001548</td>
<td>54.55</td>
<td>43.93</td>
<td>55.69</td>
<td>28.47</td>
<td>1354.36</td>
<td>1565.42</td>
<td>1459.48</td>
<td>1612.85</td>
<td>0.00012</td>
</tr>
<tr>
<td>M12125</td>
<td>35.9</td>
<td>46.64</td>
<td>35.73</td>
<td>35.27</td>
<td>642.46</td>
<td>577.81</td>
<td>580.5</td>
<td>707.35</td>
<td>0.00028</td>
</tr>
<tr>
<td>X13839</td>
<td>46.16</td>
<td>47.72</td>
<td>26.79</td>
<td>17</td>
<td>652.66</td>
<td>653.14</td>
<td>546.12</td>
<td>720.43</td>
<td>0.0003</td>
</tr>
<tr>
<td>X15882</td>
<td>13.52</td>
<td>15.73</td>
<td>27.32</td>
<td>16.15</td>
<td>209.3</td>
<td>209.64</td>
<td>221.24</td>
<td>267.43</td>
<td>0.0004</td>
</tr>
<tr>
<td>AB002533</td>
<td>659.25</td>
<td>958.82</td>
<td>812.77</td>
<td>786.24</td>
<td>407.91</td>
<td>558.33</td>
<td>529.68</td>
<td>379.84</td>
<td>0.00557</td>
</tr>
<tr>
<td>M93651</td>
<td>40.1</td>
<td>54.77</td>
<td>39.93</td>
<td>40.37</td>
<td>8.74</td>
<td>21.07</td>
<td>14.45</td>
<td>32.94</td>
<td>0.01038</td>
</tr>
</tbody>
</table>

Top 3 genes highly correlated!

<table>
<thead>
<tr>
<th></th>
<th>AF001548</th>
<th>M12125</th>
<th>X13839</th>
<th>X15882</th>
<th>AB002533</th>
<th>M93651</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF001548</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M12125</td>
<td>0.99</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X13839</td>
<td>0.991</td>
<td>0.996</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X15882</td>
<td>0.992</td>
<td>0.995</td>
<td>0.988</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB002533</td>
<td>-0.87</td>
<td>-0.898</td>
<td>-0.891</td>
<td>-0.888</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>M93651</td>
<td>-0.8</td>
<td>-0.802</td>
<td>-0.789</td>
<td>-0.776</td>
<td>0.808</td>
<td>1</td>
</tr>
</tbody>
</table>
Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work
Proposed solution

• Several possible approaches
  - next neighbors
  - correlation
  - euclidean distance

• Approach: instead use clustering

• Advantages using clustering techniques
  - natural embedding
  - many different distance functions possible
  - different shapes, models possible
Hard clustering – k-means

1. Randomly assign cluster to each point
2. Find centroids
3. Reassign points to nearest center
4. Iterate until convergence
instead of hard assignment, probability for each cluster

Very similar to k-means but fuzzy softness factor $m$ (between 1 and infinity) determines how hard the assignment has to be
Fuzzy examples

Nottermans carcinoma dataset:
18 colon adenocarcinoma and 18 normal tissues
data from 7457 genes and ESTs

cluster all 36 tissues
Fuzzy softness 1.3

18 tumors, 18 normals, 5 fuzzy clusters, m = 1.3
Fuzzy softness 1.25

18 tumors, 18 normals, 5 fuzzy clusters, m = 1.25

- ○ tumor
- × normal
Fuzzy softness 1.2

18 tumors, 18 normals, 5 fuzzy clusters, m = 1.2

○ tumor
× normal
Fuzzy softness 1.15

18 tumors, 18 normals, 5 fuzzy clusters, m = 1.15

- tumor
- normal
Selecting genes from clusters

- Two way filter: exclude redundant genes, select informative genes
- Get as many pathways as possible
- Consider cluster size and quality as well as discriminative power
How many genes per cluster?

- **Constraints:**
  - minimum one gene per cluster
  - maximum as many as possible

- Take genes proportionally to cluster quality and size of cluster

- Take more genes from bad clusters

- Smaller quality value indicates tighter cluster

- Quality for k-means: sum of intra cluster distance
- Quality for fuzzy c-means: avg cluster membership probability
Which genes to pick?

- Genes closest to center
- Genes farthest away
- Sample according to probability function
- Genes with best discriminative power
Comparison Evaluation

Repeat for each of the n examples:
leave out one sample

test data

extract features

classify held-out sample

train data

train learner

microarray data: n examples
with m expression levels each
Support Vector machines

- Find separating hyperplane with maximal distance to closest training example

- Advantages:
  - avoids overfitting
  - can handle higher order interactions and noise using kernel functions and soft margin
Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work
Experimental setup

**Datasets:**

- Alons Colon (40 tumor and 22 normal colon adenocarcinoma tissue samples)
- Golubs Leukemia (47 ALL, 25 AML)
- Nottermans Carcinoma and Adenoma (18 adenocarcinoma, 4 adenomas and paired normal tissue)

**Experimental setup:**

- calculate LOOCV using SVM on feature subsets
- do this for feature size 10-100 (in steps of 10) and 1-30 clusters
Results

Alon Fuzzy t-test

Error score

Number of features

Number of clusters

10 20 30 40 50 60 70 80 90 100

0 10 20 30 25 5 15 10 20 12 11 10 9 8 7 6

12

30/38
fuzzy c-means vs k-means
Different test-statistics
Comparing best results

![Graph showing error score vs number of features for Alon Fuzzy t-test](image-url)
How about randomly choosing?
Related work

Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work
Future work

- Problem how to find best parameters (model selection, model based clustering, BIC)
- Combine good solutions
- Incorporate overall cluster discriminative power into quality score
- Use of non integer error score
- ROC analysis
Summary

- Used clustering as a pre-filter for feature selection in order to get rid of redundant data
- Defined a quality measurement for clustering techniques
- Incorporated cluster quality, size and statistical property into feature selection
- Improved LOOCV error for almost all feature sizes and different related tests
Result Notterman

Notterman Carcinoma Fuzzy Fisher

Notterman Carcinoma Fuzzy Golub

Notterman Carcinoma Fuzzy Park

Notterman Carcinoma Fuzzy TNoM

Notterman Carcinoma Fuzzy t-test
Result Golub
Result Alon

---

Alon Fuzzy Fisher

Alon Fuzzy Golub

Alon Fuzzy Park

Alon Fuzzy TNoM

Alon Fuzzy t-test

---

41/38
Result Alon 2