Improved Gene Selection for Classification of Microarrays

Jochen Jäger
Rimli Sengupta
Walter L. Ruzzo

MPI Berlin
IIT Kharagpur
University of Washington

WU CSE Computational Biology Group
Overview

- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- Results
- Summary and Conclusions
Gene Expression: The “Central Dogma”

DNA $\rightarrow$ RNA $\rightarrow$ Protein
Gene Expression

- Proteins do most of the work
- They’re dynamically created/destroyed
- So are their mRNA blueprints
- Different mRNAs expressed at different times/places
- Knowing mRNA “expression levels” tells a lot about the state of the cell
Expression Microarrays

• Thousands to hundreds of thousands of spots per square inch
• Each holds millions of copies of a DNA sequence from one gene

• Take mRNA from cells, put it on array
• See where it sticks – mRNA from gene x should stick to spot x
An Expression Array Experiment

cells

mRNA

array

uv
Overview

- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- Results
- Summary and Conclusions
An Example Application

- 72 leukemia patients
  - 47 ALL
  - 25 AML
- 1 chip per patient
- 7132 human genes per chip

Key Issue: What’s Different?

- What genes are behaving differently between ALL & AML (or other disease/normal states)?
- Potential uses:
  - Diagnosis
  - Prognosis
  - Insight into underlying biology/biologies
  - Treatment
A Classification Problem

• Given an array from a new patient: is it ALL or AML?

• Many possible approaches: LDA, logistic regression, NN, SVM, ...

• Problems:
  – Noise
  – Dimensionality
Feature Selection

• Base the classification on only a subset of the genes
  – Reduce dimensionality – for convenience
  – Drop noisy/irrelevant genes – for accuracy

• Perhaps a very small subset
  – For cost
  – For workload
  – For biological insight
Simple Feature Selection

• Rank genes based on their individual predictive ability, e.g. by t-test or other statistic

• Keep only the top k genes
  + simple, easy, commonly used
  – often highly correlated, so little extra info
An Example

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Adenoma</th>
<th>Normal</th>
<th>t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>M18000</td>
<td>705.41</td>
<td>1227.27</td>
<td>959.35</td>
</tr>
<tr>
<td>X62691</td>
<td>387.91</td>
<td>577.57</td>
<td>578.45</td>
</tr>
<tr>
<td>M82962</td>
<td>91.85</td>
<td>16.27</td>
<td>12.61</td>
</tr>
<tr>
<td>U37426</td>
<td>0.47</td>
<td>7.05</td>
<td>6.30</td>
</tr>
<tr>
<td>HG2564</td>
<td>2.33</td>
<td>0.54</td>
<td>1.58</td>
</tr>
<tr>
<td>Z50853</td>
<td>35.43</td>
<td>26.03</td>
<td>51.49</td>
</tr>
<tr>
<td>M32373</td>
<td>-48.02</td>
<td>-28.20</td>
<td>-64.62</td>
</tr>
</tbody>
</table>
An Example (cont.)

<table>
<thead>
<tr>
<th></th>
<th>M18000</th>
<th>X62691</th>
<th>M82962</th>
<th>U37426</th>
<th>HG2564</th>
<th>Z50853</th>
<th>M32373</th>
</tr>
</thead>
<tbody>
<tr>
<td>M18000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X62691</td>
<td>0.961</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M82962</td>
<td>-0.944</td>
<td>-0.971</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U37426</td>
<td>0.973</td>
<td>0.975</td>
<td>-0.983</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG2564</td>
<td>0.592</td>
<td>0.653</td>
<td>-0.553</td>
<td>0.529</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z50853</td>
<td>0.514</td>
<td>0.616</td>
<td>-0.633</td>
<td>0.597</td>
<td>0.614</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>M32373</td>
<td>-0.509</td>
<td>-0.590</td>
<td>0.602</td>
<td>-0.580</td>
<td>-0.619</td>
<td>-0.874</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Example

![Graph with two lines: M18000 and X62691. The lines show different trends across categories A1 to A4 and N1 to N4.](image-url)
Problem with the simple 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
Overview

• Gene Expression Microarrays
• Classification and Feature Selection
• One Problem & Three Approaches
• Results
• Summary and Conclusions
Three Approaches

A: A greedy algorithm picks low p-values and not too high correlation

B: Cluster genes; pick representatives from each cluster

C: Like B, but “mask out” (omit) clusters having poor p-values

Goal of all 3: broader representation of informative genes & pathways
A: “Correlation”

- First gene picked is the one with best $p$-value.
- $k$th gene picked is the one with best $p$-value among genes having correlation less than threshold $t$ to previous $k-1$. 
B: “Clustering”

• Cluster genes into $g$ groups
• From each cluster, select one or more genes, choosing those with lowest $p$-values
• Take more from clusters with broad dispersion, fewer from tight clusters (which are likely to be highly correlated)
C: “Masked out Clustering”

- Just like B, but don’t take any genes from clusters whose average p-value is poor (> 0.2).
Clustering Algorithms

- K-means
- “Fuzzy” k-means
Hard clustering – k-means

1. Randomly assign cluster to each point
2. Find centroids
3. Reassign points to nearest center
4. Iterate until convergence

Diagram showing the process of hard clustering with k-means.
Soft - Fuzzy Clustering

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
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
Fuzzy softness 1.05

18 tumors, 18 normals, 5 fuzzy clusters, m = 1.05
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
Which genes to pick?

• Choices:
  – Genes closest to center
  – Genes farthest away
  – Sample according to probability function
  – Genes with best discriminative power
Overview

• Gene Expression Microarrays
• Classification and Feature Selection
• One Problem & Three Approaches
• Results
• Summary and Conclusions
Experimental setup

• Datasets:
  – Golub, et al.: Leukemia (47 ALL, 25 AML)
  – Alon, et al.: Colon (40 tumor and 22 normal colon adenocarcinoma tissue samples)
  – Notterman, et al.: 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
Comparison Evaluation

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

apply same feature extraction to left out sample

classify held-out sample

microarray data: n examples with g expression levels each

train data

extract features

train learner
Support Vector Machines

• Find separating hyperplane with maximal distance to closest training example

  – avoids overfitting
  – can handle higher order interactions and noise using kernel functions and soft margin

SVM with linear kernel, decision boundary (black) plus Support Vectors (red)

SVM with RBF kernel, width $\gamma$, decision boundary (black) plus Support Vectors (red)
Results: Alon, Fuzzy, t-test
Alon, Fuzzy, Other Stats

Alon Fisher

Alon Wilcoxon

Alon t-test

Alon Golub

Alon TNoM
ROC Scores: Alon, t-test
More ROC Scores

Golub t-test

Number of features

ROC score

Normal
Clustering
Masked out Clustering
Correlation
More ROC Scores

Golub TNoM

- Normal
- Clustering
- Masked out Clustering
- Correlation

ROC score vs Number features
More ROC Scores

![Golub Wilcoxon ROC Scores Graph]

- **Normal**
- **Clustering**
- **Masked out Clustering**
- **Correlation**

Number of features on the x-axis, ROC score on the y-axis.
More ROC Scores

![Graph showing ROC scores for different methods. The x-axis represents the number of features, while the y-axis represents the ROC score. The graph compares Normal, Clustering, Masked out Clustering, and Correlation methods.]
More ROC Scores

Alon Wilcoxon

ROC score

Number of features

- Normal
- Clustering
- Masked out Clustering
- Correlation
Overview

• Gene Expression Microarrays
• Classification and Feature Selection
• One Problem & Three Approaches
• Results
• Summary and Conclusions
Summary I: Problem

- Sample classification is an important application of microarrays
  - For better diagnostics, prognostics, etc.
- Finding small feature sets with high classification accuracy is important
  - For cost, for biological insight
- “Standard” method (top k genes by your favorite statistical test) is not bad
  - But very often picks highly correlated subset
Summary II: Our Idea

• Explicitly pick subsets to emphasize diversity (reduced correlation) while retaining good individual statistics, hopefully will improve joint accuracy

• Three methods:
  – Greedy selection
  – Selection from clusters
  – Selection from clusters with masking
Summary III: Results

- It works
- Details vary a bit depending on data set and test statistic, but all 3 methods generally better than “standard”
- Improvement most significant for small feature set sizes
- Improvement greater for parametric tests than non-parametric tests
More Information

• Appeared in Pacific Symposium on Biocomputing, 2003
• Preprint, supplementary data
  – http://www.molgen.mpg.de/~jaeger/psb
Acknowledgements

- My coauthors
- Bill Noble
- Ranier Spang

- NIH
- NSF