Microarray Data Mining: Puce a ADN Recent Developments

Gregory Piatetsky-Shapiro

KDDnuggets

EGC 2005, Paris
Role of Gene Expression

Gene expression

Gene (mRNA), single strand

Gene (DNA)

Protein

Cell

Nucleus

Chromosome

Graphics courtesy of the National Human Genome Research Institute

© 2005 KDnuggets
Gene Expression

- Cells are different because of differential gene expression.
- About 40% of human genes are expressed at one time.
- Gene is expressed by transcribing DNA exons into single-stranded mRNA.
- mRNA is later translated into a protein.
- Microarrays measure the level of mRNA expression.
Gene Expression Measurement

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured with latest technology
- mRNA is isolated and labeled with fluorescent protein
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser
Affymetrix DNA Microarrays

500,000 locations on each GeneChip® array

Actual size of GeneChip® array

1.28 cm

Millions of DNA strands built up in each location

Actual strand = 25 base pairs

image courtesy of Affymetrix
Some Probes Match

<table>
<thead>
<tr>
<th>Match Probe</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

© 2005 KDnuggets
# Other Probes MisMatch

<table>
<thead>
<tr>
<th>Gene</th>
<th>MisMatch Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

...
Affymetrix Microarray Concept

1. mRNA segments tagged with fluorescent chemical
2. Matches with complementary probes
3. Fluorescence measured with laser

image courtesy of Affymetrix
Affymetrix Microarray Raw Image

enlarged section of raw image

Scanner

raw data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D26528_at</td>
<td>193</td>
</tr>
<tr>
<td>D26561_cds1_at</td>
<td>-70</td>
</tr>
<tr>
<td>D26561_cds2_at</td>
<td>144</td>
</tr>
<tr>
<td>D26561_cds3_at</td>
<td>33</td>
</tr>
<tr>
<td>D26579_at</td>
<td>318</td>
</tr>
<tr>
<td>D26598_at</td>
<td>1764</td>
</tr>
<tr>
<td>D26599_at</td>
<td>1537</td>
</tr>
<tr>
<td>D26600_at</td>
<td>1204</td>
</tr>
<tr>
<td>D28114_at</td>
<td>707</td>
</tr>
</tbody>
</table>
Microarray Potential Applications

- New and better molecular diagnostics
  - Jan 11, 2005: FDA approved Roche Diagnostic AmpliChip, based on Affymetrix technology

- New molecular targets for therapy
  - Few new drugs, large pipeline, ...

- Improved treatment outcome
  - Partially depends on genetic signature

- Fundamental Biological Discovery
  - Finding and refining biological pathways

- Personalized medicine ?!
Microarray Data Analysis Types

- Gene Selection
  - Find genes for therapeutic targets (new drugs)

- Classification (Supervised)
  - Identify disease
  - Predict outcome / select best treatment

- Clustering (Unsupervised)
  - Find new biological classes / refine existing ones

- Discovery of Associations and Pathways
Outline

- DNA and Biology
- Microarray Classification - Best practices
- Gene Set Analysis
- Synthetic microarray data sets
Microarray Data Analysis Challenges

- Biological, Process, and Other variation
- Model needs to be explainable to biologists
- Few records (samples), usually < 100
- Many columns (genes), usually > 1,000
  - This is very likely to result in false positives, “discoveries” due to random noise
Data Mining from Small Data: Example

- Les Américains boivent beaucoup de vin et ont beaucoup de maladie de coeur
- Les Français boivent beaucoup de vin et ont peu de maladie de coeur
- Les Anglais boivent beaucoup de biere et ont beaucoup de maladie de coeur
- Les Allemands boivent beaucoup de biere et ont peu de maladie de coeur
Conclusion

- Vous pouvez boire tout que vous vouliez

- C’est de parler en anglais que cause la maladie de coeur

- 😊
Same Data, Different Results – Who is Right?

- Many examples (e.g. CAMIDA conferences) where researchers analyzed the same data, and found different results and gene sets.
  - Who is right?

- Good methodology is needed
  - for avoiding errors
  - for best results
Capturing Best Practices for Microarray Data Analysis

- Worked with SPSS and S. Ramaswamy (MIT / Whitehead Institute) to capture best practices.
- Implemented as SPSS Microarray CATs (Clementine Application Template).
- Also implemented using Weka and open-source software.

Capturing Best Practice for Microarray Gene Expression Data Analysis, G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, Proceedings of KDD-2003

www.kdnuggets.com/gpspubs/
Best Practices

- Capture the complete process, from raw data to final results
- Gene (feature) selection inside cross-validation
- Randomization testing
- Robust classification algorithms
  - Simple methods give good results
  - Advanced methods can be better
- Wrapper approach for best gene subset selection
- Use bagging to improve accuracy
- Remove/relabel mislabeled or poorly differentiated samples
Observations

- Simplest approaches are most robust
- Advanced approaches can be more accurate
- “Small” increase in diagnostic accuracy (e.g. 90% to 98%) can greatly reduce rate of false positives (5-fold)
Microarray Classification Desired Features

- Robust in presence of false positives
- Stable under cross-validation
- Results understandable by biologists
- Return confidence/probability
- Fast enough
Popular Classification Methods

- Decision Trees/Rules
  - Find smallest gene sets, but not robust – poor performance
- Neural Nets - work well for reduced number of genes
- K-nearest neighbor – good results for small number of genes, but no model
- Naïve Bayes – simple, robust, but ignores gene interactions
- Support Vector Machines (SVM)
  - Good accuracy, does own gene selection, but hard to understand
- Specialized methods, D/S/A (Dudoit), …
Gene Reduction Improves Classification Accuracy - Why?

- Most learning algorithms look for non-linear combinations of features
  - Can easily find spurious combinations given few records and many genes – “false positives problem”

- Classification accuracy improves if we first reduce number of genes by a linear method
  - e.g. T-values of mean difference
Feature selection approach

- Rank genes by measure & select top 100-200

- T-test for Mean Difference =

\[
\frac{(\text{Avg}_1 - \text{Avg}_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}\]

- Signal to Noise (S2N) =

- Other methods ...
Global Feature / Parameter Selection is wrong

Because the information is “leaked” via gene selection. Leads to overly “optimistic” classification results.
Global Feature Selection Bias

- Example: Data w 6000 features, ~ 100 samples
- Used 3 Weka algorithms: SVM, IB1, J48
- Error with Global feature selection was 50% to 150% lower than using X-validation

![Graph showing average % increase in error from using Global Feature Selection (A)]
Microarray Classification Process

- Gene data
  - Train data
  - Class data
  - Test data

1. Data Cleaning & Preparation
2. Feature/Parameter Selection
3. Model Building
4. Evaluation

- Model Building

© 2005 KDnuggets
Evaluation

- Evaluation on one test dataset is not accurate (for small datasets)

- Evaluation, including feature/parameter selection needs to be inside cross-validation (Validation Croisee)
Evaluation inside X-validation

Gene Data

Microarray Classification Process

Final Test 1

Final Model 1

Final Result 1
Evaluation inside X-validation, 2

Gene Data

Train

Final Test 2

Microarray Classification Process

Final Model 2

Final Result 2
Evaluation inside X-validation, N

Gene Data

Final Test
N
Train

Microarray Classification Process

Final Model N

Final Result N
Iterative Wrapper approach to selecting the best gene set

- Model with top 100 genes is not optimal
- Test models using 1, 2, 3, ..., 10, 20, 30, 40, ..., N top genes with cross-validation.

Gene selection:
- Simple: equal number of genes from each class
- advanced: best number from each class
- For “randomized” algorithms (e.g. neural nets), average 10 cross-validation runs
Best Gene Set: one X-validation run

Error Avg for 10 Gene sets per Class

Single Cross-Validation run
Best Gene Set
10 X-validation runs

- Select gene set with lowest combined Error
- Good, but not optimal!

Average, high and low error rate for all classes
Multi-class classification

- Simple: One model for all classes
- Advanced: Separate model for each class
Error rates for each class

![Error rates graph](image)
Example: Pediatric Brain Tumor Data

- 92 samples, 5 classes (MED, EPD, JPA, MGL, RHB) from U. of Chicago Children’s Hospital
- Photomicrographs of tumours (400x)
## Single Neural Net

<table>
<thead>
<tr>
<th>Class</th>
<th>1-Net Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED</td>
<td>2.1%</td>
</tr>
<tr>
<td>MGL</td>
<td>17%</td>
</tr>
<tr>
<td>RHB</td>
<td>24%</td>
</tr>
<tr>
<td>EPD</td>
<td>9%</td>
</tr>
<tr>
<td>JPA</td>
<td>19%</td>
</tr>
<tr>
<td><em>ALL</em></td>
<td>8.3%</td>
</tr>
</tbody>
</table>
Bagging Improves Results

- Bagging or simple voting of N different neural nets improves the accuracy
## Bagging 100 Networks

<table>
<thead>
<tr>
<th>Class</th>
<th>1-Net Error Rate</th>
<th>Bag Error rate</th>
<th>Bag Avg Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED</td>
<td>2.1%</td>
<td>2% (0)*</td>
<td>98%</td>
</tr>
<tr>
<td>MGL</td>
<td>17%</td>
<td>10%</td>
<td>83%</td>
</tr>
<tr>
<td>RHB</td>
<td>24%</td>
<td>11%</td>
<td>76%</td>
</tr>
<tr>
<td>EPD</td>
<td>9%</td>
<td>0</td>
<td>91%</td>
</tr>
<tr>
<td>JPA</td>
<td>19%</td>
<td>0</td>
<td>81%</td>
</tr>
<tr>
<td><em>ALL</em></td>
<td>8.3%</td>
<td>3% (2)*</td>
<td>92%</td>
</tr>
</tbody>
</table>

- **Note:** suspected error on one sample (labeled as MED but consistently classified as RHB)
Cross-validated prediction strength

MED: Strong predictions

misclassified?

poorly differentiated
Outline

- DNA and Biology
- Microarray Classification - Best practices
- Gene Set Analysis
- Synthetic microarray data sets
Gene Sets – Key to Power

- Number of genes is a problem for single gene analysis
- Analyzing gene sets increases statistical power
- Group genes statistically (prototypes, correlations)
- Group genes using biological knowledge (pathways, Gene Ontology, medical text, …)
Gene Set Analysis

- Gene Set Enrichment
  - Mootha et al, Nature Genetics, July 2003

- Question: Genes involved in oxidative phosphorylation (~100) vs. diabetes?

- No genes in OXPHOS group are individually significant (avg. ~20% decrease)
Mean expression of all genes (gray) and of OXPHOS genes (red) for people with DM2 (Type 2 diabetes mellitus) vs Normal.
Outline

- DNA and Biology
- Microarray Classification - Best practices
- Gene Set Analysis
- Synthetic microarray data sets
Synthetic Microarray Data

- Getting high accuracy is very important
- Current data has too few samples; the “true” labels may be unknown.
- What methods are best for given data and what is their estimated accuracy?
- Proposal: generate synthetic but realistic microarray data with KNOWN labels to evaluate algorithms under different conditions
First Step: Analysis of Existing Microarray Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Platform, Software</th>
<th>Tissue</th>
<th>Genes</th>
<th>Samples</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golub train</td>
<td>Affy HuGeneFL; MAS4</td>
<td>Blood</td>
<td>7070</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>Golub test</td>
<td>Affy HuGeneFL; MAS4</td>
<td>Blood</td>
<td>7070</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Mich Lung</td>
<td>Affy HuGeneFL</td>
<td>Lung</td>
<td>7070</td>
<td>96</td>
<td>3</td>
</tr>
<tr>
<td>GCM train</td>
<td>Affy Hu6800 and Hu35KsubA</td>
<td>Many</td>
<td>16004</td>
<td>144</td>
<td>14</td>
</tr>
<tr>
<td>GCM normal</td>
<td>Affy Hu6800 and Hu35KsubA</td>
<td>Many</td>
<td>16004</td>
<td>90</td>
<td>13</td>
</tr>
<tr>
<td>Bremer pp5</td>
<td>Affy HuGeneFL; Probe Profiler</td>
<td>Brain</td>
<td>7070</td>
<td>92</td>
<td>5</td>
</tr>
<tr>
<td>Brain U133</td>
<td>Affy U133; Probe Profiler</td>
<td>Brain</td>
<td>22215</td>
<td>33</td>
<td>4</td>
</tr>
</tbody>
</table>

Different Technologies, Tissues, Normal/Abnormal
Global Distribution of Gene Means

Analyzed means of ALL genes across samples
Excluded Mean < 1 (mismatch, random)

Found: log(means) ~ normal distribution
Gene means are globally logNormal
Global Gene Means (Q-Q plot)

Plot: $\log_{10}(\text{means})$ (blue) vs normal quantiles (red)
Gene Expression Standard Deviation vs Mean

Globally

\[
\log(\text{SD}) \sim a \log(\text{means}) + b
\]
True Rank vs T-value Rank

- Simulated 100 microarray datasets, with 6000 genes and 30 samples (using R)
- First 10 genes up-regulated by UP factor

![Graph showing True Rank vs T-value Rank](image-url)
Future Research Directions

- Why is gene distribution log normal?
- Develop a synthetic microarray data generator
- Which gene selection methods are most robust and under what conditions?
- Evaluate algorithm accuracy for a given test set by creating similar synthetic sets with known answers
Acknowledgements

- Eric Bremer, Children’s Hospital (Chicago) & Northwestern U.
- Tom Khabaza, SPSS
- Sridhar Ramaswamy, MIT/Broad Institute
- Pablo Tamayo, MIT/Broad Institute
SIGKDD Explorations Special Issue on Microarray Data Mining

December 2003. Volume 5, Issue 2

Table of Contents:

Editorial
Usama M. Fayyad

Articles on Microarray Data Mining

Microarray Data Mining: Facing the Challenges
Gregory Piatetsky-Shapiro and Pablo Tamayo

A Novel Approach to Determine Normal Variation in Gene Expression Data
V. Nadimpally and M. Zaki

Gene Ranking Using Bootstrapped P-values
S. N. Mukherjee, P. Skykacek, S. J. Roberts, S. J. Gurr

Guest Editors
Gregory Piatetsky-Shapiro
KDNuggets and U. Mass Lowell
gregory@kdnuggets.com

Pablo Tamayo
MIT / Broad Institute	
tamayo@broad.mit.edu

Editor-in-Chief
Sunita Sarawagi
IIT Bombay	sunita@iitb.ac.in

Associate Editor
Paul Bradley
Apollo Data Technologies	
paul@apollodatatech.com

Associate Editor
Usama Fayyad
DMX Group	fayyad@dmxgroup.com

www.acm.org/sigkdd/explorations/issue5-2.htm
Merci!

Further resources on Data Mining:  www.KDnuggets.com

Microarrays:
www.KDnuggets.com/websites/microarray.html

Data Mining Course (one semester)
www.KDnuggets.com/dmcourse
Questions?