Supervised Learning for Gene Expression Microarray Data

David Page
University of Wisconsin

Joint Work with:

- Mike Waddell, James Cussens, Jo Hardin
- Frank Zhan, Bart Barlogie, John Shaughnessy
Common Approaches

- Comparing two measurements at a time
  - Person 1, gene G: 1000
  - Person 2, gene G: 3200
  - Greater than 3-fold change: flag this gene

- Comparing one measurement with a population of measurements… is it unlikely that the new measurement was drawn from same distribution?

Approaches (Continued)

- Clustering or Unsupervised Data Mining
  - Hierarchical Clustering, Self-Organizing (Kohonen) Maps (SOMs), K-Means Clustering
  - Cluster patients with similar expression patterns
  - Cluster genes with similar patterns across patients or samples (genes that go up or down together)
Approaches (Continued)

- **Classification or Supervised Data Mining.**
  - Use our knowledge of class values… myeloma vs. normal, positive response vs. no response to treatment, etc., to gain added insight.
  - Find genes that are best predictors of class.
    - Can provide useful tests, e.g. for choosing treatment.
    - If predictor is *comprehensible*, may provide novel insight, e.g., point to a new therapeutic target.

Approaches (Continued)

- **Classification or Supervised Learning.**
  - UC Santa Cruz: Furey et al. 2001 (support vector machines).
  - SNPs and Proteomics are coming.
Outline

- Data and Task
- Supervised Learning Approaches and Results
  - Tree Models and Boosting
  - Support Vector Machines
  - Voting
  - Bayesian Networks
- Conclusions

Data

- Publicly-available from Lambert Lab at http://lambertlab.uams.edu/publicdata.htm
- 105 samples run on Affymetrix HuGenFL
  - 74 Myeloma samples
  - 31 Normal samples
Two Ways to View the Data

- Data points are genes.
  - Represented by expression levels across different samples.
  - Goal: find related genes.
- Data points are samples (e.g., patients).
  - Represented by expression levels of different genes.
  - Goal: find related samples.

<table>
<thead>
<tr>
<th>Person</th>
<th>Gene</th>
<th>A28202_ac</th>
<th>AB00014_at</th>
<th>AB00015_at</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>P</td>
<td>1142.0</td>
<td>A</td>
<td>321.0</td>
<td>P  2567.2</td>
</tr>
<tr>
<td>Person 2</td>
<td>A</td>
<td>-586.3</td>
<td>P</td>
<td>586.1</td>
<td>P  759.0</td>
</tr>
<tr>
<td>Person 3</td>
<td>A</td>
<td>105.2</td>
<td>A</td>
<td>559.3</td>
<td>P  3210.7</td>
</tr>
<tr>
<td>Person 4</td>
<td>P</td>
<td>-42.8</td>
<td>P</td>
<td>692.1</td>
<td>P  812.0</td>
</tr>
</tbody>
</table>
## Data Points are Genes

<table>
<thead>
<tr>
<th>Person</th>
<th>Gene</th>
<th>A28202_ac</th>
<th>AB00014_at</th>
<th>AB00015_at</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>P</td>
<td>1142.0</td>
<td>A</td>
<td>321.0</td>
</tr>
<tr>
<td>Person 2</td>
<td>A</td>
<td>-586.3</td>
<td>P</td>
<td>586.1</td>
</tr>
<tr>
<td>Person 3</td>
<td>A</td>
<td>105.2</td>
<td>A</td>
<td>559.3</td>
</tr>
<tr>
<td>Person 4</td>
<td>P</td>
<td>-42.8</td>
<td>P</td>
<td>692.1</td>
</tr>
</tbody>
</table>

## Data Points are Samples

<table>
<thead>
<tr>
<th>Person</th>
<th>Gene</th>
<th>A28202_ac</th>
<th>AB00014_at</th>
<th>AB00015_at</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>P</td>
<td>1142.0</td>
<td>A</td>
<td>321.0</td>
</tr>
<tr>
<td>Person 2</td>
<td>A</td>
<td>-586.3</td>
<td>P</td>
<td>586.1</td>
</tr>
<tr>
<td>Person 3</td>
<td>A</td>
<td>105.2</td>
<td>A</td>
<td>559.3</td>
</tr>
<tr>
<td>Person 4</td>
<td>P</td>
<td>-42.8</td>
<td>P</td>
<td>692.1</td>
</tr>
</tbody>
</table>
Supervision: Add Classes

<table>
<thead>
<tr>
<th>Person</th>
<th>Gene</th>
<th>A28202_ac</th>
<th>AB00014_at</th>
<th>AB00015_at</th>
<th>...</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>P</td>
<td>1142.0</td>
<td>A</td>
<td>321.0</td>
<td>P</td>
<td>2567.2</td>
</tr>
<tr>
<td>Person 2</td>
<td>A</td>
<td>-586.3</td>
<td>P</td>
<td>586.1</td>
<td>P</td>
<td>759.0</td>
</tr>
<tr>
<td>Person 3</td>
<td>A</td>
<td>105.2</td>
<td>A</td>
<td>559.3</td>
<td>P</td>
<td>3210.7</td>
</tr>
<tr>
<td>Person 4</td>
<td>P</td>
<td>-42.8</td>
<td>P</td>
<td>692.1</td>
<td>P</td>
<td>812.0</td>
</tr>
</tbody>
</table>

The Task

Data Points are:
- Genes
- Patients

Clustering

Supervised Data Mining

Predict the class value for a patient based on the expression levels for his/her genes
Outline

- Data and Task
- **Supervised Data Mining Algorithms**
  - Tree Models and Boosting
  - Support Vector Machines
  - Voting
  - Bayesian Networks
- Conclusions

Decision Trees in One Picture

[Decision tree diagram showing classification of Myeloma and Normal based on AvgDiff of G5]
C5.0 (Quinlan) Result

Decision tree:
AD_X57809_at \leq 20343.4: myeloma (74)
AD_X57809_at > 20343.4: normal (31)

Leave-one-out cross-validation accuracy estimate: 97.1%
X57809: IGL (immunoglobulin lambda locus)

Problem with Result

Easy to predict accurately with genes related to immune function, such as IGL, but this gives us no new insight.

Eliminate these genes prior to training.
Ignoring Genes Associated with Immune function

Decision tree:
AD_X04898_rna1_at <= -1453.4: normal (30)
AD_X04898_rna1_at > -1453.4: myeloma (74/1)

X04898: APOA2 (Apolipoprotein AII)
Leave-one-out accuracy estimate: 98.1%.

Next-Best Tree

AD_M15881_at > 992: normal (28)
AD_M15881_at <= 992:
   AC_D82348_at = A: normal (3)
   AC_D82348_at = P: myeloma (74)

M15881: UMOD (uromodulin…Tamm-Horsfall glycoprotein, uromucoid)
D82348: purH
Leave-one-out accuracy estimate: 93.3%
GeneCards Reveals...

**UROM_HUMAN**: uromodulin precursor (tamm-horsfall urinary glycoprotein) (thp).--gene: umod. [640 amino acids; 69 kd]

function: not known. may play a role in regulating the circulating activity of cytokines as it binds to il-1, il-2 and tnf with high affinity.

subcellular location: attached to the membrane by a gpi-anchor, then cleaved to produce a soluble form which is Secreted in urine.

tissue specificity: synthesized by the kidneys and is the most abundant protein in normal human urine.

Boosting

- After building a tree, give added weight to any data point the tree mislabels.
- Learn a new tree from re-weighted data.
- Repeat 10 times.
- To classify a new data point, let trees vote (weighted by their accuracies on the training data).
Boosting Results

- Leave-one-out accuracy estimate: 99.0%.
- With Absolute Calls only: 96.2%.
- But it is much harder to understand, or gain insight from, a weighted set of trees than from a single tree.

Summary of Accuracies

<table>
<thead>
<tr>
<th></th>
<th>AC Only</th>
<th>AC + AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>90.5</td>
<td>98.1</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>95.2</td>
<td>99.0</td>
</tr>
<tr>
<td>SVMs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes Nets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Data and Task
- Supervised Data Mining Algorithms
  - Tree Models and Boosting
  - Support Vector Machines
  - Voting
  - Bayesian Networks
- Conclusions

Support Vector Machines
Maximizing the Margin between Bounding Planes (Mangasarian & Fung)
SVM Results (Defaults)

- Accuracy using Absolute Call only is better than accuracy using AC + AD.
  - AC: 95.2%
  - AC + AD: 93.3%
- Difficult to interpret results… open research area to extract most important genes from SVM.
- Might be useful for choosing a therapy but not yet for gaining insight into disease.

Summary of Accuracies

<table>
<thead>
<tr>
<th></th>
<th>AC Only</th>
<th>AC + AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>90.5</td>
<td>98.1</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>96.2</td>
<td>99.0</td>
</tr>
<tr>
<td>SVMs</td>
<td>95.2</td>
<td>99.5–100.0</td>
</tr>
<tr>
<td>Vote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes Nets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Voting Approach

- Score genes using information gain.
- Choose top 1% (or other number) scoring genes.
- To classify a new case, let these genes vote (majority or weighted majority vote).
- We use majority vote here.
Voting Results (Absolute Call)

- Using only Absolute Calls, accuracy is 94.0%.
- Appears we can improve accuracy by requiring only 40% of genes to predict myeloma in order to make a myeloma prediction.
- Would be interesting to test this on new Lambert Lab data.

Top Voters (AC Only)

<table>
<thead>
<tr>
<th>SCORE</th>
<th>GENE</th>
<th>MP</th>
<th>MA</th>
<th>NP</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.446713</td>
<td>H1F2</td>
<td>57</td>
<td>17</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>0.446713</td>
<td>NCBP2</td>
<td>57</td>
<td>17</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>0.432706</td>
<td>SM15</td>
<td>56</td>
<td>18</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>0.432706</td>
<td>GCN5L2</td>
<td>56</td>
<td>18</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>0.412549</td>
<td>maj hist comp</td>
<td>12</td>
<td>62</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>0.411956</td>
<td>RNASE6</td>
<td>15</td>
<td>59</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>0.411956</td>
<td>TNFRSF7</td>
<td>15</td>
<td>59</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>0.411956</td>
<td>SDF1</td>
<td>15</td>
<td>59</td>
<td>30</td>
<td>1</td>
</tr>
</tbody>
</table>
All top 1% splits are based on AD.

Leave-one-out results appear to be 100%…double-checking this to be sure.

35 is cutoff point for myeloma vote. No normal gets more than 15 votes, and no myeloma gets fewer than 55.

<table>
<thead>
<tr>
<th>SCORE</th>
<th>GENE</th>
<th>SPLIT</th>
<th>MH</th>
<th>ML</th>
<th>NH</th>
<th>NL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.802422</td>
<td>APOA2</td>
<td>-777</td>
<td>74</td>
<td>0</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>0.735975</td>
<td>HERV K22 pol 637</td>
<td>3</td>
<td>71</td>
<td>31</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.704489</td>
<td>TERT</td>
<td>-1610</td>
<td>70</td>
<td>4</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>0.701219</td>
<td>UMOD</td>
<td>1119.1</td>
<td>0</td>
<td>74</td>
<td>28</td>
<td>3</td>
</tr>
<tr>
<td>0.701219</td>
<td>CDH4</td>
<td>-278</td>
<td>74</td>
<td>0</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>0.664859</td>
<td>ACTR1A</td>
<td>3400.6</td>
<td>3</td>
<td>71</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>0.664859</td>
<td>MASP1</td>
<td>-536.6</td>
<td>71</td>
<td>3</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>0.650059</td>
<td>PTPN21</td>
<td>1256.1</td>
<td>6</td>
<td>68</td>
<td>31</td>
<td>0</td>
</tr>
</tbody>
</table>
Summary of Accuracies

<table>
<thead>
<tr>
<th></th>
<th>AC Only</th>
<th>AC + AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>90.5</td>
<td>98.1</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>96.2</td>
<td>99.0</td>
</tr>
<tr>
<td>SVMs</td>
<td>95.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Vote</td>
<td>94.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Bayes Nets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Outline

- Data and Task
- Supervised Data Mining Algorithms
  - Tree Models and Boosting
  - Support Vector Machines
  - Voting
  - Bayesian Networks
- Conclusions
Bayes Nets for Gene Expression Data

- Friedman et al. 1999 has been followed by much work on this approach.
- Up to now, primarily used to discovery dependencies among genes, not to predict class values.
- Recent experience suggests using Bayes nets to predict class values.

KDD-2001 Cup
The Genomics Challenge

Christos Hatzis, Silico Insights
David Page, University of Wisconsin
Co-chairs

August 26, 2001

Special thanks: DuPont Pharmaceuticals Research
Laboratories for providing data set 1, Chris Kostas from Silico Insights for cleaning and organizing data sets 2 and 3

http://www.cs.wisc.edu/~dpage/kddcup2001/
The model & its performance

<table>
<thead>
<tr>
<th></th>
<th>Activity</th>
<th>10695</th>
<th>91839</th>
<th>16794</th>
<th>79651</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Predicted</strong></td>
<td></td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>neg</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
<td>55</td>
<td>128</td>
<td>356</td>
</tr>
</tbody>
</table>

Accuracy: 0.711
Weighted Accuracy: 0.684

Bayes Nets Result

- Network with 23 genes selected.
- Diagnosis node is parent of 20 others. Others have at most three other parents.
- Leave-one-out accuracy estimate is 97%.
- Software is not capable of handling numerical values at this time.
Summary of Accuracies

<table>
<thead>
<tr>
<th>Method</th>
<th>AC Only</th>
<th>AC + AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>90.5</td>
<td>98.1</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>96.2</td>
<td>99.0</td>
</tr>
<tr>
<td>SVMs</td>
<td>95.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Vote</td>
<td>94.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Bayes Nets</td>
<td>95.2</td>
<td>NA</td>
</tr>
</tbody>
</table>
Further Work

- Interpreting SVMs.
- Analyzing new, larger data sets.
- Other classification tasks: prognosis, treatment selection, MGUS vs. Myeloma.

Conclusions

- Supervised learning produces highly accurate predictions for this task. Noise not a problem.
- Don’t throw out negative average differences!
- So far the ability of SVMs to consider magnitude of differences in expression level has not yielded benefit over voting, which just uses consistency.
- Domain experts like readability of trees, voting, Bayes nets, but trees give worse accuracy.
- Many of the most predictive genes line up with expectations of domain experts.
Using Absolute Calls Only

U78525_at = A: normal (21/1)
U78525_at = P:
  M62505_at = P: normal (5)
  M62505_at = A:
    AF002700_at = M: normal (2)
    AF002700_at = A:
      U97188_at = P: normal (2)
      U97188_at = A:
        HG415-HT415_at = A: myeloma (72)
        HG415-HT415_at = P: normal (3/1)