Classification for Gene Function Determination
Outline

• Background Motivation
• Approaches
• Results
• Conclusions
Background and Motivation

• Large Amount of Data Largely Not Interpreted
• Data Culled from Numerous Sources
• Data Heterogeneous
• Large Attribute to Number of Example Ratio Makes Automatic Classification More Difficult
• Unclear what the best method for classification will be
Approaches

• K-Nearest Neighbor
• Decision Trees
• Boosting
• Support Vector Machines
• Boosting Adaptive Decision Trees
• Others
K-Nearest Neighbor

• Natural algorithm works for many data sets
• Somewhat slow in classification because each example to be classified must be compared to every training example
• Must tune to correct value of \( k \) and correct weighting of neighbors
Decision Trees

$x = \{\text{Outlook=Sunny, Temp=Hot, Humidity=High, Wind=Strong}\}$

- Good algorithm (C4.5) for constructing
- Very Expressive (perhaps too expressive)
Boosting

• Allows the combination of a group of “weak” experts into a single powerful one

• Needs independent errors among experts and experts that are correct more often than not.

• Hot machine learning topic, < 10 years old
Support Vector Machines

• Finds hyperplane separating examples in n-dimensional space

• Input space can be mapped to higher dimensional feature space to allow more expressive separations

• Using kernel functions never have to represent this space explicitly
Alternating Boosting Decision Trees

- Interesting Combination of Decision Tree and Boosting
- Performs Relatively Well
- Nice implementation
Results from the Literature

- Direct kNN outperforms SVM (Kuramochi and Karypis)
- SVM ourperforms Decision Trees, Parzen Windows, and Fisher’s Linear Discriminant (Brown et. al)
- Boosting Works Pretty Well 😊 (Dettling and Buhlmann)
# My Results

(accuracy)

<table>
<thead>
<tr>
<th>Method</th>
<th>KEGG</th>
<th>COG</th>
<th>Multi-Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td>55.4%</td>
<td>42.2%</td>
<td>63.23%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>38.9%</td>
<td>17.4%</td>
<td>31.1%</td>
</tr>
<tr>
<td>Boosting</td>
<td>32.9%</td>
<td>18.3%</td>
<td>40.4%</td>
</tr>
<tr>
<td>SVM</td>
<td>32.03%</td>
<td>33.45%</td>
<td>52.02%</td>
</tr>
<tr>
<td>Alternating Boosting</td>
<td>31.7%</td>
<td>18.5%</td>
<td>35.9%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
My Results

Results

Kegg - Cogg - MF

kNN
D-Tree
Boosting
SVM
ADTrees
Problems – Data

• Data scattered among numerous databases
• Data is non-heterogeneous
• Examples are relatively sparse
• Examples have numerous attributes
Observations

• It’s a lot of work to prepare real data for classification
• Nature of the data makes classification difficult
Conclusions

• K-NN proved best
• Classification in this domain is difficult and requires novel techniques
• Still a young field and seems likely to become easier to apply automated techniques as more data is classified