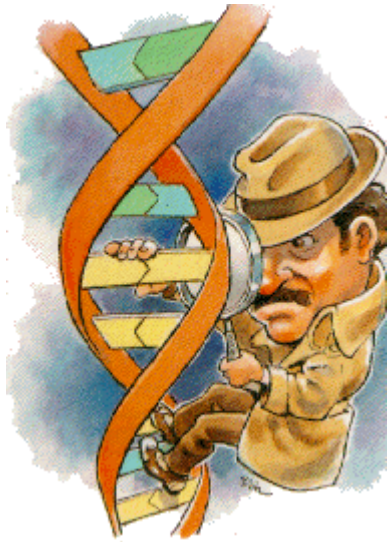
Classification for Gene Function Determination



Outline

- Background Motivation
- Approaches
- Results
- Conclusions

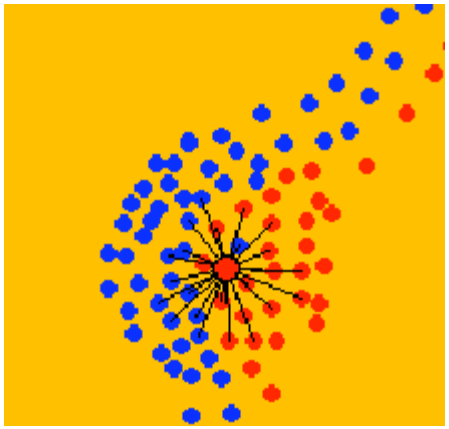
Background and Motivation

- 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

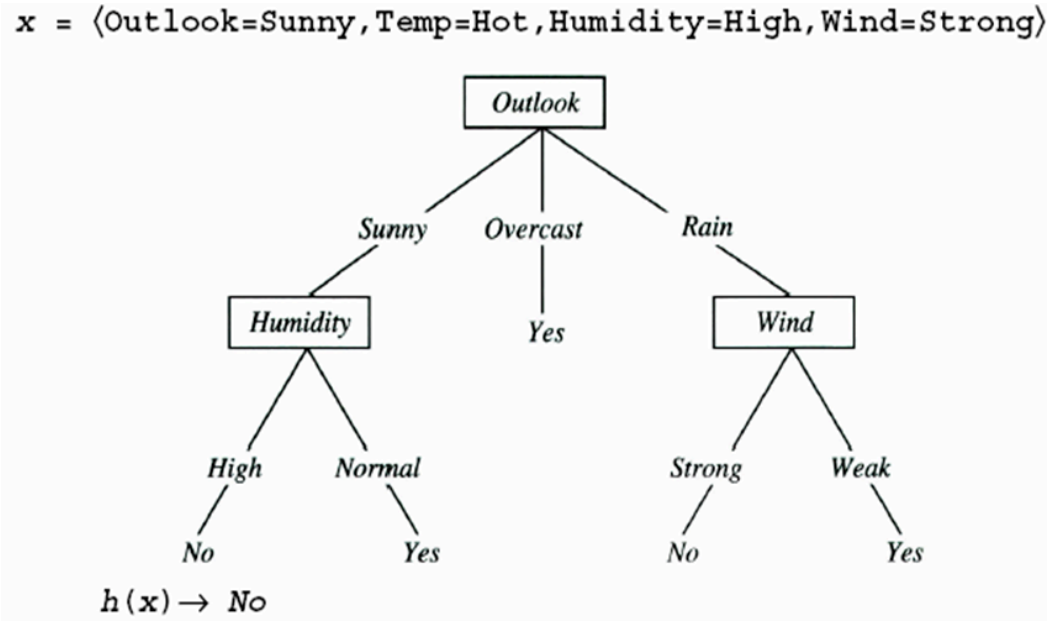
- K-Nearest Neighbor
- Decision Trees
- Boosting
- Support Vector Machines
- Boosting Adaptive Decision Trees
- Others

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



- 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

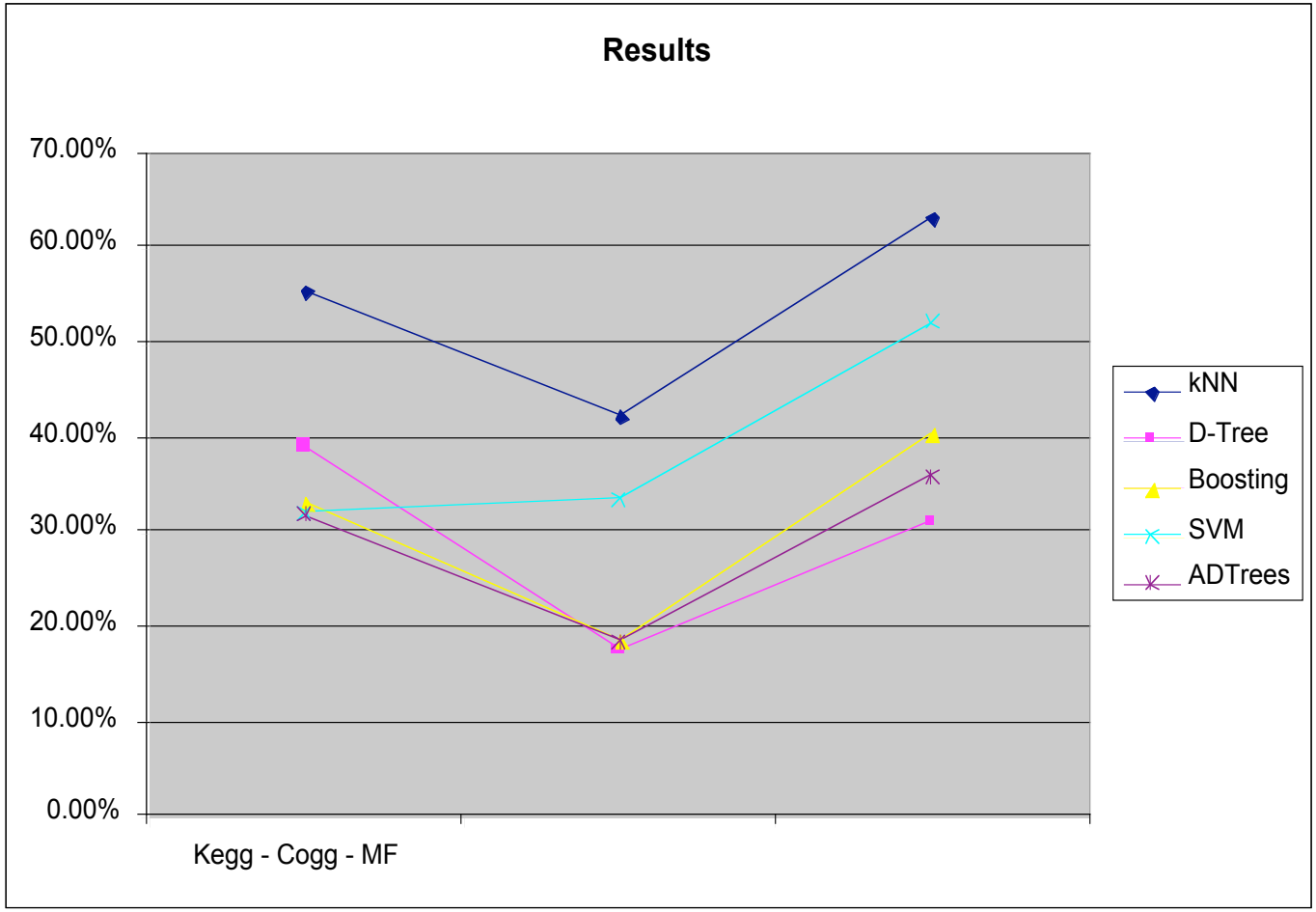
- Direct kNN outperforms SVM (Kuramochi and Karypis)
- SVM ourperforms Decision Trees, Parzen Windows, and Fisher's Linear Discriminant (Brown et. al)
- Boosting Works Pretty Well ☺ (Dettling and Buhlmann)

My Results

(accuracy)

	KEGG	COG	Multi-Function
K-NN	55.4%	42.2%	63.23%
Decision Trees	38.9%	17.4%	31.1%
Boosting	32.9%	18.3%	40.4%
SVM	32.03%	33.45%	52.02%
Alternating Boosting Decision Trees	31.7%	18.5%	35.9%

My Results


Results
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
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