Evaluating the Effectiveness of a Combination of Multiple Classifiers

Patrick Vacek

7 May 2008

MATH 3210

Overview

- Concept of Voting
- The Spam Dataset
- Algorithms
- Results

Voting

- A combination of multiple classifiers
- Individual strengths and weaknesses
- Simple majority

Source: Dunham 119-120

The Spam Dataset

- Definition
- Source
- Attributes

Source: Hopkins

Naïve Bayesian Classifier

- Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Classification technique

Source: Bayes' theorem; Dunham 52-53, 86-87

C4.5

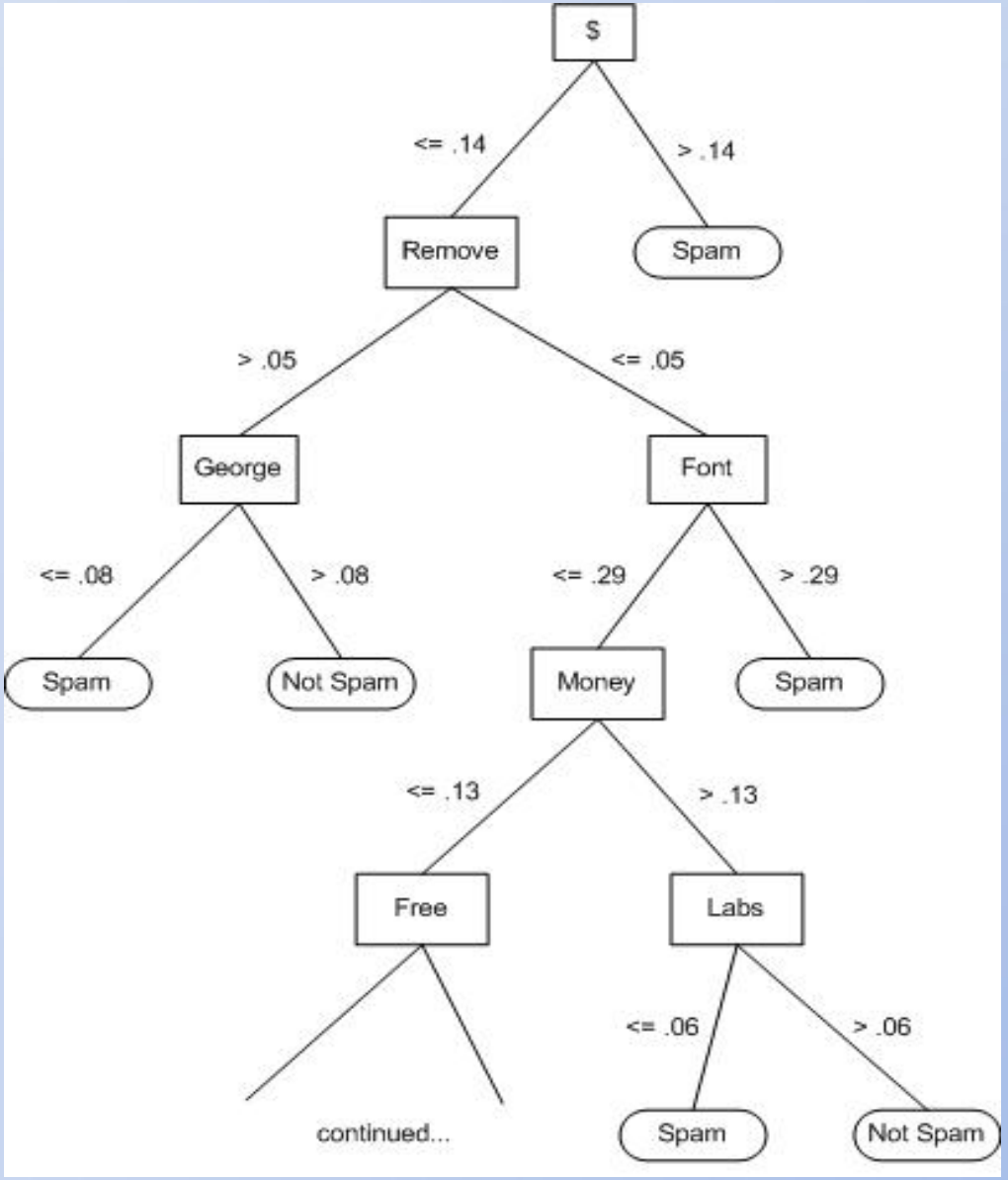
- Decision trees
- Information entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

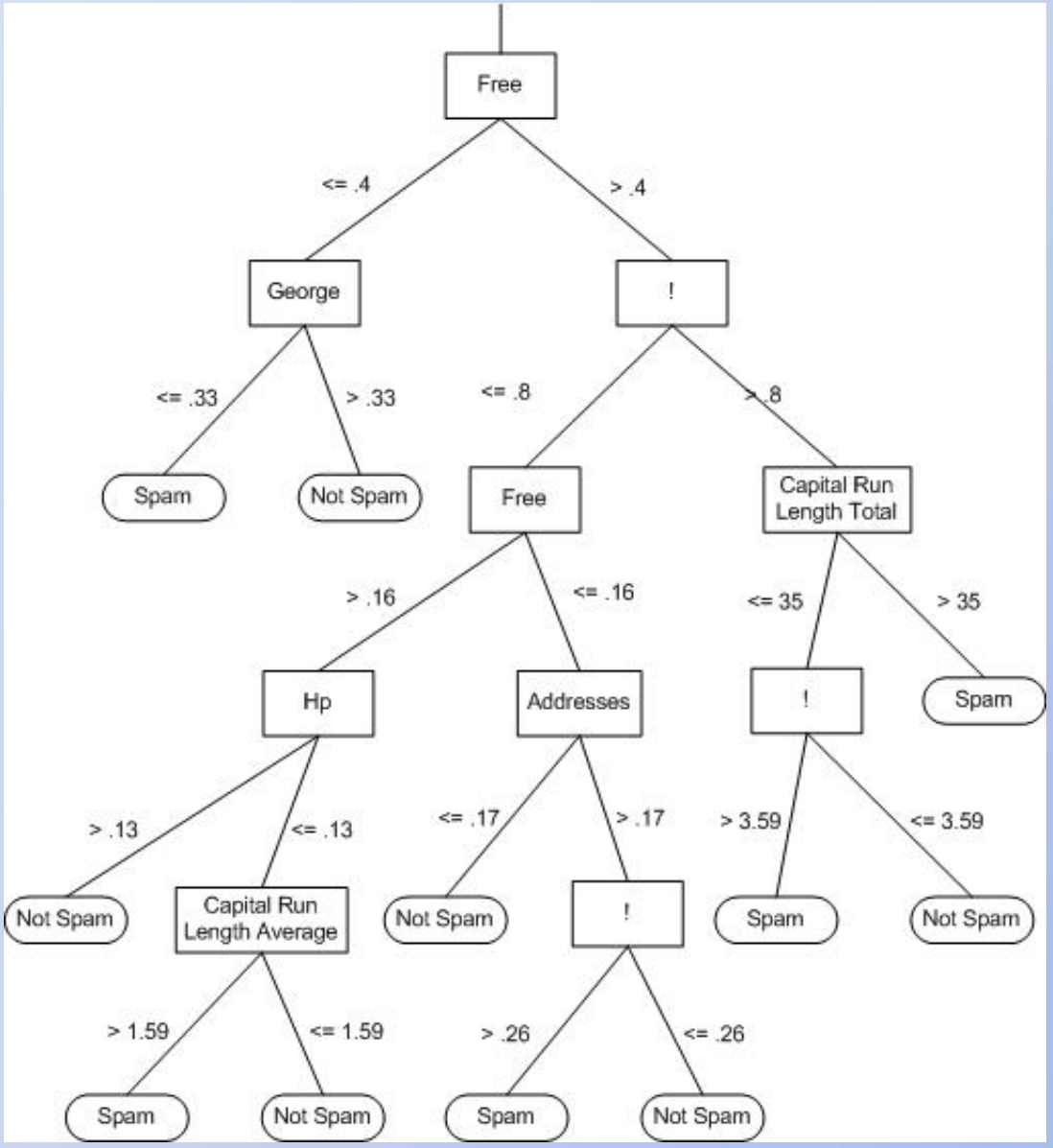
- Information gain
- Splitting attributes

Source: Information entropy, Dunham 92-93, 97-98, 100-101; Han 286

C4.5 Tree (Part One)



C4.5 Tree (Part Two)



K-Nearest Neighbors

- Distance
- Classification

Source: Han 314-315

Multilayer Perceptron

- Artificial neural network
- Input, output, hidden layers
- Supervised learning
- Backpropagation
- Perceptrons

Source: Dunham 103-109, 112-113

Genetic Algorithm

- Natural evolution
- Fitness
- Reproduction
- Generating rules

Source: Han 316

Results

	Naïve Bayes	C4.5	KNN	Perceptron	Genetic	Voting
Errors:	562	333	328	518	513	222
Error %:	0.24435	0.14478	0.14261	0.22522	0.22304	0.09652
False Positives:	532	225	159	488	318	201
False Positive %:	0.38191	0.16152	0.11414	0.35032	0.22828	0.14429
False Negatives:	30	108	169	30	195	21
False Negative %:	0.03308	0.11907	0.18633	0.03308	0.21499	0.02315

Conclusion

- Concept of Voting
- The Spam Dataset
- Choice of Algorithms
- Evaluation of Results

References

- Aleshunas, John, and Ross Quinlan (2007). *C4.5 Decision Tree Induction*. Retrieved May 1, 2008, from http://mercury.webster.edu/aleshunas/MATH%203210/MATH %203210% 20Source%20Code%20and%20Executables.html.
- Bayes' theorem (2008). In *Wikipedia, The Free Encyclopedia*. Retrieved May 6, 2008, from http://en.wikipedia.org/w/index.php?title=Bayes%27_theorem&oldid=209271193.
- Dunham, Margaret H. (2003). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.
- Frank, Eibe, et al. (2007). *Weka 3: Data Mining Software in Java*. The University of Waikato. Retrieved April 26, 2008, from http://www.cs.waikato.ac.nz/ml/weka/.
- *GATree* (2006). AHEADRM.com. Retrieved April 22, 2008, from http://www.gatree.com/ index.html.
- Han, Jiawei, and Micheline Kamber (2001). *Data mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.
- Hopkins, Mark, et al. (1999). *SPAM E-mail database*. Hewlett-Packard Labs. Retrieved April 22, 2008, from ftp://ftp.ics.uci.edu/pub/machine-learning-databases/ spambase/.
- Information entropy (2008). In *Wikipedia, The Free Encyclopedia*. Retrieved May 6, 2008, from http://en.wikipedia.org/w/index.php? title=Information_entropy&oldid=208004419.