IBM SpamGuru on the TREC 2005 Spam Track

Richard Segal

Jason Crawford, Jeff Kephart, Barry Leiba, V.T. Rajan, Mark Wegman

November 18, 2005
SpamGuru Architecture
Classifier Aggregation

- Combine results of multiple classifiers to predict spam
- Catch more spam with less false positives
  - Emphasize each algorithms strengths.
  - De-emphasize each algorithms weaknesses.
- Harder to attack
  - Must simultaneously break through multiple algorithms.
- Adapts to changes in classifier effectiveness
Aggregation by Optimized Linear Weights

\[ \text{Score}(x) = b + \prod_i (W_i \times \text{Score}_i(x)) \]

- Find optimal values for \( W_i \) and \( b \) using a Nelder-Mead non-linear optimizer.
- Re-optimize values every 10,000 examples.
Naïve Bayesian

Approximate calculation using a greedy update rule.
Very little additional computational cost.
Big performance boost.

Less Naïve Bayesian
SMTP Path Analysis
Submission Details

- **General**
  - Pre-trained on 20,000 labeled messages.
  - Corpus created from honeypots and user voting records.

- **Text Processing**
  - Decoded MIME but left attachments in place.
  - Word-based tokens. No stop words, no feature selection.
  - Special handling of URLs, e-mail addresses, etc.

- **TREC Required Bug (minor)**
  - Classified all messages longer than 100K as good.
“Full” Dataset

![Graph showing classification rates for spam and ham emails using different methods. The x-axis represents the percentage of ham misclassification on a logit scale, while the y-axis represents the percentage of spam misclassification on a logit scale. The methods compared include Best*, Aggregate of LNB+SMTP, Less Naive Bayesian, and SMTP Path Analysis. The graph illustrates the performance of these methods across different classification rates.]
“SB” Dataset

Graph showing the performance of different methods in classifying spam and ham messages. The x-axis represents the percentage of ham misclassification on a logit scale, while the y-axis represents the percentage of spam misclassification on a logit scale. The methods compared include:

- Best*
- Aggregate of LNB+SMTP
- Less Naive Bayesian
- SMTP Path Analysis

The graph illustrates how each method performs under varying degrees of misclassification, with the Best method generally performing the best across the range of misclassification rates.
“TM” Dataset

![Graph showing performance metrics for different spam classification methods. The x-axis represents % Ham Misclassification (logit scale), and the y-axis represents % Spam Misclassification (logit scale). The graph compares Best*, Aggregate of LNB+SMTP, Less Naive Bayesian, and SMTP Path Analysis.]
“Mr. X” Dataset

![Graph showing comparison of misclassification rates for different methods.]

- Best*
- Aggregate of LNB+SMTP (Corrected)
- Aggregate of LNB+SMTP (Original)
- Less Naive Bayesian
- SMTP Path Analysis

% Spam Misclassification (logit scale) vs % Ham Misclassification (logit scale)
Summary

- Classifier aggregation using non-linear optimization.

- Less-naïve Bayesian performs well.

- SMTP path analysis is not very good in isolation, but combines well with Less-naïve Bayesian.

http://www.research.ibm.com/spam