Spam Filtering:
A Bayesian Approach

RAJ PATEL
HITESH SHARMA

Introduction

- Electronic mail is an efficient and increasingly popular communication medium, however, it is prone to misuse. One such case of misuse is the blind posting of unsolicited e-mail messages, also known as spam, to very large numbers of recipient.
- Many readers of E-mail spend a non-trivial portion of their time online wading through such unwanted messages.
- Commercial products allow users to handcraft a set of logical rules to filter junk mail. This solution is a time-consuming, tedious and error-prone process.
- A junk mail filtering system should be able to automatically adapt to the changes in the characteristics of junk mail over time.

Previous Work

Filtering Junk E-mail using Bayesian Networks (Sahami Et Al)

- Uses Naive Bayesian classifier to solve the problem of junk Email filtering
- Incorporates domain specific properties through introduction of additional features in Bayesian classifier.

$$P(C = c_k \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_k) P(C = c_k)}{P(\mathbf{X} = \mathbf{x})}.$$

Comparison of naive Bayesian and Memory based approach (University of Athens)

- Uses Naive Bayesian classifier and Memory based approach to solve the problem of junk Email filtering.
- Each training instance is represented as a point in a multidimensional space.
- Classification procedure is a variant of the simple K nearest neighbour algorithm.
- Uses TiMBL software for memory based classification algorithm.

Data-set

- Data-set used was collected from **SpamAssassin**
- http://spamassassin.apache.org/publiccorpus/
- Training Data: 3 spam datasets and 2 ham(non-spam) datasets
- Test Data: 2 ham(non-spam) datasets and 1 spam dataset

Our Approach

- **TOKENIZER**: Stanford Tokenizer - http://nlp.stanford.edu/software/tokenizer.shtml
- **METHOD 1**: Naive Bayes Classification

$$\frac{P(S|E)}{P(H|E)} = \frac{P(S)}{P(H)} \prod_{i=1}^{n} \frac{P(w_i|S)}{P(w_i|H)}$$

- **METHOD 2**: Paul Graham's Method - A token is considered unknown if it occurs less than 5 times in the both the datasets.

$$P(S|w_i) = \frac{AS/S}{AS/S + AH/H}$$

$$P(S|E) = \frac{\prod_{i=1}^{n} P(S|w_i)}{\prod_{i=1}^{n} P(S|w_i) + \prod_{i=1}^{n} P(H|w_i)}$$

- These tokens are assigned a neutral probability 0.5. Tokens that occur only in spam dataset are assigned 0.99 probability and those in ham dataset are assigned 0.1.

Our Approach

- **METHOD 3**: Tim Peter's Modification

$$P(S|E) = \frac{P(S)^{1-n} \prod_{i=1}^{n} P(S|w_i)}{P(S)^{1-n} \prod_{i=1}^{n} P(S|w_i) + P(H)^{1-n} \prod_{i=1}^{n} P(H|w_i)}$$

- **METHOD 4**: Gary Robinson smoothing

$$P^*(S|w_i) = \frac{\alpha + AS}{\alpha + \beta + (AS + AH)}$$

α	Strength given to background information
β	Assumed probability of unknown tokens

Results

Method 1		
	HAM	SPAM
HAM	2682	119
SPAM	396	105
Accuracy: 93.126%, Error: 6.783%		

Method 2		
	HAM	SPAM
HAM	2608	193
SPAM	403	98
Accuracy: 91.187%, Error: 8.8.12%		

Method 3		
	HAM	SPAM
HAM	2540	261
SPAM	456	45
Accuracy: 90.732%, Error: 9.2%		

Method 4		
	HAM	SPAM
HAM	2753	48
SPAM	421	80
Accuracy: 96.123%, Error: 3.87%		

References

- http://robotics.stanford.edu/users/sahami/papers-dir/spam.pdf
- http://cgi.di.uoa.gr/~takis/pkdd00.pdf
- http://arxiv.org/pdf/cs/0006013.pdf
- http://www.cs.ubbcluj.ro/~gabis/DocDiplome/Bayesian/000539771r.pdf
- http://dl.acm.org/citation.cfm?id=636753
- http://www.cs.ubbcluj.ro/~gabis/DocDiplome/Bayesian/000539771r.pdf

THANKS