Spam Detection using Clustering, Random Forests, and Active Learning

Dave DeBarr
George Mason University
4400 University Drive
Fairfax, VA 22030
ddebarr@gmu.edu

Harry Wechsler, PhD
George Mason University
4400 University Drive
Fairfax, VA 22030
wechsler@gmu.edu

ABSTRACT
This paper describes work in progress. Our research is focused on efficient construction of effective models for spam detection. Clustering messages allows for efficient labeling of a representative sample of messages for learning a spam detection model using a Random Forest for classification and active learning for refining the classification model. Results are illustrated for the 2007 TREC Public Spam Corpus. The area under the Receiver Operating Characteristic (ROC) curve is competitive with other solutions while requiring much fewer labeled training examples.

1. INTRODUCTION
Undesired, unsolicited email is a nuisance for its recipients; however, it also often presents a security threat. For example, it may contain a link to a phony website intending to capture the user's login credentials (identity theft, phishing), or a link to a website that installs malicious software (malware) on the user's computer. Installed malware can be used to capture user information, send spam, host malware, host phish, or conduct denial of service attacks as part of a "bot" net. While prevention of spam transmission would be ideal, detection allows users and email providers to address the problem today.

2. BACKGROUND
Traditional machine learning techniques involve having a user label examples of both spam and ham (not spam) messages so that a computer algorithm can learn to identify unwanted email. For email systems that process large quantities of messages, it is practically impossible to label all messages processed for some time period. In order to reduce the burden, the messages to be labeled are often selected randomly (passive learning), or they are selected by computer algorithm (active learning). Two common criteria used to select examples for training a pattern recognition model include:

- density based selection: a small set of representative examples are chosen for labeling, based on how well the examples characterize the data
- uncertainty based selection: an initial machine learning model is constructed from a labeled sample of examples, then the algorithm asks for labels for unlabeled examples where it is most uncertain about the class label

CEAS 2009 – Sixth Conference on Email and Anti-Spam, July 16-17, 2009, Mountain View, California USA

3. PROPOSED METHOD
The method proposed in this paper relies on using term frequency and inverse document frequency representation for messages; clustering a sample of messages from the training pool and obtaining labels for cluster medoids; constructing an initial Random Forest for spam detection; requesting labels for additional training examples based on uncertainty; and refining the Random Forest model. Figure 1 illustrates the process.

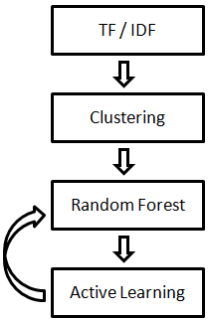

Figure 1. Training Process

3.1 TF / IDF
RFC 822 (Internet) email messages are divided into two sections. The lines above the first blank line are headers, while the lines below the first blank line comprise the message body. As shown in Figure 2, common headers include the "From", "To", "Subject", "Date", "Return-Path", and "Received" lines. Other headers include the "Content-Type" and "X-Mailer" lines.

```
Return-Path: <RickyAmes@aol.com>
Received: from 129.97.78.23 ([211.202.101.74])
        by speedy.uwaterloo.ca (8.12.8/8.12.5)
        with SMTP id l38H7GOI003017;
        Sun, 8 Apr 2007 13:07:21 -0400
Received: from 0.144.152.6 by 211.202.101.74;
        Sun, 08 Apr 2007 19:04:48 +0100
Message-ID: <WYADCKPDFWWTWTXNFVUE@yahoo.com>
From: "Tomas Jacobs" <RickyAmes@aol.com>
Reply-To: "Tomas Jacobs" <RickyAmes@aol.com>
To: the00@speedy.uwaterloo.ca
Subject: Generic Cialis, branded quality@
Date: Sun, 08 Apr 2007 21:00:48 +0300
X-Mailer: Microsoft Outlook Express 6.00.2600.0000
MIME-Version: 1.0
Content-Type: multipart/alternative;
        boundary="--8896484051606557286"
X-Priority: 3
X-MSMail-Priority: Normal
Status: RO
Content-Length: 988
Lines: 24
```

Figure 2. Example of RFC822 Message Headers

Each message was converted to a set of normalized term frequency, inverse document frequency (TF/IDF) features [6]. Tokens were extracted from messages by converting the characters to lower case, separating on symbolic ASCII characters (such as the comma and the period) and white space, and prefixing each token from a header with the name of the header. For example, the tokens for the headers in Figure 2 would include "subject:cialis" and "x-mailer:outlook". In order to reduce noise, only tokens that occur in at least 1% of the messages in the training pool and at most 99% of the messages in the training pool were retained.

3.2 Clustering

Clustering has been combined with active learning in other application domains [7]. In this work, clustering is used to select an initial set of email messages to be labeled as training examples. The Partitioning Around Medoids (PAM) algorithm [5] was used to cluster a uniform random sample of 25% of the messages in the training pool. PAM is an implementation of the k-medoids algorithm. PAM selects the "k" most centrally located messages for its initial model, then iteratively assigns other messages to the nearest medoid and updates the medoid for each cluster. It is similar to the k-means clustering algorithm, but is less sensitive to the presence of outliers (unusual messages). The parameters for the model include the choice of distance measure and the number of clusters, "k". For these experiments, Euclidean distance was used to measure the difference between email messages, and "k" was chosen to be the number of messages to be labeled; e.g. if we wanted to label only 10 messages for our initial spam detection model, "k" was chosen to be 10.

3.3 Random Forest

After the cluster prototype messages were selected for training, feature selection and model selection were performed using leave-one-out cross validation. In repeated experiments, the model with the best performance was a Random Forest [1]. Table 1 shows a comparison of cross validation performance using 10 cluster prototypes for training. The performance measure is Area Under the receiver operating characteristic Curve (AUC) [4].

Method	AUC
Random Forest	95.2%
Naive Bayes	66.7%
SVM	66.7%
kNN	66.7%

Table 1. Comparison of Cross Validation Performance

A Random Forest is an ensemble of decision trees that vote on the spam/ham label for new messages. For each tree, a bootstrap sample is drawn from the labeled data and a decision tree is constructed by considering a random subset of features for each decision node in the tree. An example of a decision would be: "normalized TF/IDF value for token" > threshold. The strengths of the Random Forest method include feature selection and consideration of many feature subsets (instead of focusing on just a few features that best separate the training data). The key parameters for the Random Forest model include the number of trees to build and the number of features to consider for each decision node: 1,000 trees and 13 features were used as parameter values for these experiments.

3.4 Active Learning

Once the initial Random Forest model is constructed, additional messages are selected for labeling by choosing examples from the training pool where the probability of spam assigned by the Random Forest model is closest to 0.5. The probability of spam is computed as the proportion of decision trees assigning the spam label. Once labeled, the selected messages are then added to the cluster prototypes and the Random Forest is retrained.

4. EXPERIMENTS

The 2007 TREC Spam Email Corpus contains 75,419 messages received by an email server at the University of Waterloo between April 8 and July 6, 2007. For the experimental results reported here, the first week of data (9,535 messages) was used as a training pool and the next 12 weeks of messages were used for testing. Of the 573,670 distinct tokens found in the training pool, 4,515 were retained after removing tokens that appear in too few or too many messages; i.e. the retained tokens had to appear in at least 1% of the messages (to avoid over fitting) and at most 99% of the messages (to avoid "stop" words).

The Receiver Operating Characteristic curves in Figure 3 compare the performance of the following message selection algorithms on the test set:

- Cluster10 (dotted): unlabeled messages in the training pool were clustered into 10 groups and the medoids were selected [Area Under the Curve (AUC): 91.3%]
- Random10 (solid): 10 messages from the training pool were randomly selected [AUC: 68.2%]
- Cluster10+Uncertainty10 (dotted): Cluster10 was used to construct an initial model, then 10 more messages from the training pool were selected based on the assigned probability of spam [AUC: 92.9%]
- Random10+Uncertain10 (solid): Random10 was used to construct an initial model, then 10 more messages from the training pool were selected based on the assigned probability of spam [AUC: 65.7%]
- Cluster100 (dotted): unlabeled messages in the training pool were clustered into 100 groups and the medoids were selected [AUC: 99.7%]
- Random100 (solid): 100 messages from the training pool were randomly selected [AUC: 98.6%]

5. CONCLUSIONS AND NEXT STEPS

The most novel observation from this work in progress is the ability of cluster prototypes to efficiently represent compact, well separated clusters. For example, a burst of 2,532 messages from April 13th (over 25% of the training pool) can be effectively represented by a single prototype (file=inmail.5413). This cluster was a phishing attempt disguised as an important notice. Another example of a compact, yet well separated, cluster was an advertisement for a pharmaceutical product (represented by file=inmail.3258).

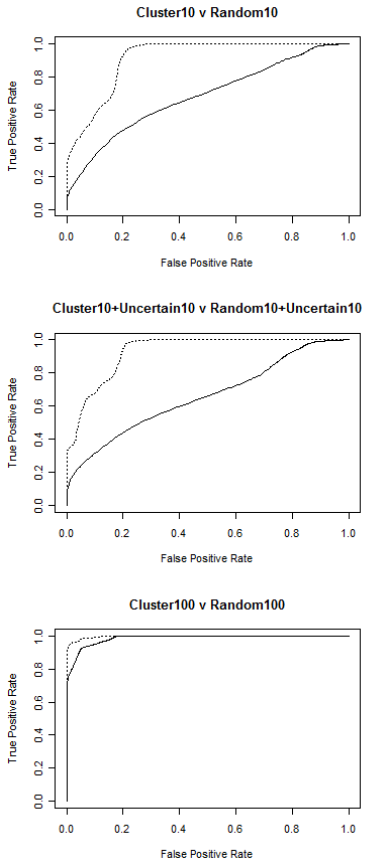

Figure 3. Receiver Operating Characteristic Curves

Cluster prototypes provide a useful representation of the training pool for the Random Forest algorithm. Choosing an initial set of messages for labeling based on cluster prototypes, then choosing an additional set of messages for model refinement based on uncertainty offers improved performance. With as few as 100 labeled messages from one week of data, performance is competitive with other reported results [2].

Our next step will focus on selecting messages for labeling based on the emergence of new clusters in the data stream for a deployed model. Our representation will also be enhanced to include information about links and images contained in the messages.

6. REFERENCES

[1] Breiman, L., "Random Forests", Machine Learning, Vol 45, Iss 1, Oct 2001, pp. 5-32, http://www.springerlink.com/content/u0p06167n6173512/fulltext.pdf.

[2] Cormack, G.V., "TREC 2007 Spam Track Overview", NIST Special Publication 500-274, The 16th Text REtrieval Conference (TREC) Proceedings, 2007, http://trec.nist.gov/pubs/trec16/papers/SPAM.OVERVIEW16.pdf.

[3] Crocker, D.H., "Standard for the Format of ARPA Internet Text Messages", ARPANET Request For Comments (RFC) No. 822, Aug 1982, http://www.ietf.org/rfc/rfc0822.txt.

[4] Fawcett, T. "An Introduction to ROC Analysis", Pattern Recognition Letters, Vol 27, Iss 8, Jun 2006, pp. 861-874.

[5] Kaufman, L. and Rousseeuw, P.J., "Partitioning Around Medoids", Finding Groups in Data, Wiley-Interscience, 2005, pp. 68-125.

[6] Manning, C.D., Raghavan, P., and Schutze, H. "Scoring, Term Weighting, and the Vector Space Model", Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2008, pp. 109-133, http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

[7] Nguyen, H.T. and Smeulders, A. "Active Learning Using Pre-clustering", Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 623-630, http://www.aicml.cs.ualberta.ca/_banff04/icml/pages/papers/94.pdf.

[8] Settles, B. "Active Learning Literature Survey", Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Jan 2009, http://pages.cs.wisc.edu/~bsettles/pub/settles.activelearning.pdf.

[9] TREC 2007 Public Spam Corpus, http://plg.uwaterloo.ca/~gvcormac/treccorpus07/.