Semi-supervised Spam Filtering: Does it Work?

Mona Mojdeh and Gordon V. Cormack
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
{mmojdeh,gvcormac}@uwaterloo.ca

ABSTRACT

The results of the 2006 ECML/PKDD Discovery Challenge suggest that semi-supervised learning methods work well for spam filtering when the source of available labeled examples differs from those to be classified. We have attempted to reproduce these results using data from the 2005 and 2007 TREC Spam Track, and have found the opposite effect: methods like self-training and transductive support vector machines yield inferior classifiers to those constructed using supervised learning on the labeled data alone. We investigate differences between the ECML/PKDD and TREC data sets and methodologies that may account for the opposite results.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Information filtering

General Terms: Experimentation, measurement.

Keywords: Spam, semi-supervised learning, transductive learning, self-feedback.

1. INTRODUCTION

A spam filter is necessarily trained on messages from a different sample space than those it is asked to classify. Obviously, training examples are from the past, while the filter is called upon to classify future messages. If the training examples are recent, their characteristics may differ little from those to be classified. However, acquiring labels is costly and time consuming, with the net effect that the training data may not well represent the messages to be classified. Spammers exploit this fact by launching spam campaigns in which millions of messages in a never-before-used format are sent in an effort to circumvent filtering in the interval before the new format is learned.

A second source of disparity between training and target data occurs because every user's email has somewhat different characteristics. A user may be unable or unwilling to collect and label a sample of his or her own messages, leaving no alternative but to train the filter on something else. Furthermore, a brand-new user will have received no messages with which to train the filter; ideally, the filter should perform well "out of the box" in spite of the lack of user-specific training data. There is furthermore the promise that collaborative filtering – the use of labeled data delivered to other users – might provide superior performance to personal filtering; for example, in the early detection of spam campaigns.

The TREC Spam Track results [6, 5] show that personal spam filters – those trained only on past email received by the recipient – achieve very strong results, with typical AUC scores of 0.9999 or better [i.e. (1-AUC) = 0.01%]. It may be argued that these results are unrealistically good because no user would provide perfect feedback; TREC experiments to measure the effect of delayed and missing feedback show some degradation but still very strong results. Further degradation – but still strong performance with 1-AUC scores of 0.1% – was observed when training was effected using messages from users separate from those for whom filter performance was tested. None of the well-performing filters at TREC harnessed unlabeled data, although it was available in the form of previously classified messages.

ECML/PKDD Discovery Challenge [1] provided training and test data sets from diverse sources, thus explicitly modeling the situation in which user-specific training data are not available. Each message was abstracted as a feature vector representing word frequencies within the message body. Filters had no knowledge of the message header, the textual representation of the message, the order of the words in the message, or the words themselves. All of the well-performing entries used some form of semi-supervised learning, with the best achieving 1-AUC scores of about 5% – orders of magnitude worse than the TREC results, but substantially better than the baseline supervised learning methods, which achieved 1-AUC scores of about 15%.

Our hypothesis in undertaking the experiments reported here was that semi-supervised learning methods might similarly improve the performance of TREC filters, so as to overcome or mitigate the performance degradation due to reduced feedback or cross-user training. What we found was the opposite. All the semi-supervised methods we investigated – including two that were successful for the Discovery Challenge – showed inferior performance to the baseline supervised methods. Only using TREC 2005 training data and TREC 2007 evaluation data were we able to reproduce the ECML/PKDD results.

2. EXPERIMENTS

We evaluated supervised and semi-supervised variants of three spam filter methods known to perform well at TREC and ECML/PKDD: SVM and transductive SVM with default parameters [8], DMC [2] and logistic regression [7] with self-training [4]. The TREC 2007 methods and data sets were adapted for batch as opposed to on-line filtering (cf.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

[3]) so as to maximize the opportunity for semi-supervised learning. The *delayed feedback* task from TREC was modeled by using the first 10,000 messages for training and the remaining messages for evaluation, divided into six batches of 10,000. The remaining 5000 messages were not used. The *partial feedback (cross-user)* task was modeled by using the 30338 messages for which feedback was given (the messages addressed to a particular subset of users) as a training set, and the remaining 45081 messages (the messages addressed to other users) for evaluation.

Method	On-line with feedback
DMC	0.007
LR	0.040

Table 1: Baseline results 1-AUC (%)

Table 1 shows the baseline performance of the DMC and logistic regression filters with full feedback; that is, the filters are operated on-line and the correct label for each message is communicated to the filter immediately after classification. All results are presented as the average of 1-AUC over all evaluation sets (smaller 1-AUC indicates better performance). Table 2 shows the results of batch evaluation in which no feedback is given. The DMC and LR results in the first column show that both filters suffer from lack of feedback; DMC moreso. Our SVM filter – which was not amenable to on-line operation – showed intermediate performance. None of the filters was improved by semi-supervised methods that worked well for the ECML/PKDD challenge – self-training and transductive learning. The parenthesized result (0.053) indicates the result that SVM would have achieved with an optimal parameter setting indicating the prevalence of spam. Since this parameter was computed with knowledge of the data, it indicates an upper bound rather than an achieved result.

Method	Supervised	semi-super.
DMC	0.016	0.090
LR	0.049	0.046
SVM	0.030	0.230 (0.053)

Table 2: Delayed feedback results 1-AUC (%)

Table 3 shows the results of cross-user training on the same messages. The performance of the supervised filters is compromised far more substantially by cross-user training than by training delay. The semi-supervised methods make matters worse.

Method	Supervised	semi-super.
DMC	2.17	9.97
LR	1.00	10.72
SVM	1.06	24.3 (1.89)

Table 3: Cross-user results 1-AUC (%)

The results of these experiments led us to conduct a cross-corpus experiment in an effort to reproduce the results of the ECML/PKDD Discovery Challenge. The training set consisted of the first 10000 messages of the TREC 2005 corpus; the evaluation sets consisted of the TREC 2007 corpus, split into 10000 message segments. The results of this cross-corpus evaluation are shown in table 4. All the filters perform poorly – at least compared to the other tasks – but still better than chance. Self-training is again harmful; however, SVM shows considerable improvement, yielding the best performance on this dataset – even with the default prevalence parameter.

Method	Supervised	semi-super.
DMC	14.9	49.1
LR	17.8	43.0
SVM	23.5	13.1 (10.3)

Table 4: Cross-corpus results 1-AUC (%)

3. CONCLUSIONS

The performance of state-of-the-art on-line spam filters is compromised by delayed or cross-user training. For training sets delayed by a matter of weeks or months, or derived from different users during a similar time frame, self-training and transductive SVM appear to exacerbate rather than mitigate the problem. Only when cross-training between data sets differing chronologically by years, and containing messages from entirely unrelated systems and users, does transductive SVM appear to yield an improvement.

The datasets we have used are easily derived from the TREC public corpora. We advance them as standards for evaluating new semi-supervised methods, and suggest that the first two – delayed and cross-user training – model the most realistic scenarios.

4. REFERENCES

[1] Bickel, S. ECML-PKDD Discovery Challenge 2006 Overview. In *Proc. ECML/PKDD Discovery Challenge Workshop* (2006).

[2] Bratko, A., Cormack, G. V., Filipic, B., Lynam, T. R., and Zupan, B. Spam filtering using statistical data compression models. *Journal of Machine Learning Research 6* (2006), 2673–2698.

[3] Cormack, G., and Bratko, A. Batch and on-line spam filter evaluation. In *CEAS 2006: The Third Conference on Email and Anti-Spam* (Mountain View, CA, 2006).

[4] Cormack, G. V. Harnessing unlabeled examples through iterative application of Dynamic Markov Modeling. In *Proc. ECML/PKDD Discovery Challenge Workshop* (2006).

[5] Cormack, G. V. TREC 2007 Spam Track Overview. In *Sixteenth Text REtrieval Conference (TREC-2007)* (Gaithersburg, MD, 2007), NIST.

[6] Cormack, G. V., and Lynam, T. R. TREC 2005 Spam Track overview. In *Fourteenth Text REtrieval Conference (TREC-2005)* (Gaithersburg, MD, 2005), NIST.

[7] Goodman, J., and tau Yih, W. Online discriminative spam filter training. In *The Third Conference on Email and Anti-Spam* (Mountain View, CA, 2006).

[8] Joachims, T. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.