Semi-Supervised Spam Filtering using Aggressive Consistency Learning

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

ABSTRACT

A graph based semi-supervised method for email spam filtering, based on the local and global consistency method, yields low error rates with very few labeled examples. The motivating application of this method is spam filters with access to very few labeled message. For example, during the initial deployment of a spam filter, only a handful of labeled examples are available but unlabeled examples are plentiful. We demonstrate the performance of our approach on TREC 2007 and CEAS 2008 email corpora. Our results compare favorably with the best-known methods, using as few as just two labeled examples: one spam and one non-spam.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Experimentation, Measurement

Keywords

Spam, Email, Filtering, Classification

1. INTRODUCTION

Semi-supervised methods are of special interest when there are very few training samples available. In many machine learning applications, there is always great human effort involved in labeling samples, while obtaining unlabeled data is fairly simple. This is the case for spam filters. During the initial deployment of spam filters, a normal user may be willing to provide only a few labeled examples for training but will still expect correct classification of a large number of emails. Another application is personalized spam filtering with low label cost, using per-user semi-supervised filters with few labeled examples to augment a global filter.

In this paper we address the problem of email spam filtering with very few correct training samples using graph based semi-supervised learning methods. Previous semi-supervised methods such as Transductive SVM and Logistic Regression and Dynamic Markov Compression with self training for spam filtering have yielded mixed results [4]. In this paper we are focused on the special situation in which the first handful of messages are labeled and used to filter the rest.

We present an aggressive graph-based iterative solution modeled after the local and global consistency learning method of Zhou *et al.* [5]. The same method is applied for detecting web spam in [3]. Local consistency guarantees that the nearby points are likely to have the same label; while the global consistency guarantees that the points on the same structure are likely to have the same label. We have also applied Single Value Decomposition to find the most informative terms. Our experiments show a comparatively high performance of our method in the presence of very few training samples.

2. AGGRESSIVE CONSISTENCY LEARNING METHOD

Given a sequence of n email messages and labels denoting the true class – *spam* or *nonspam* – of each of the first $n_{labeled} \ll n$, we consider the problem of finding the class of the remaining $n - n_{labeled}$ messages. Algorithm 1 demonstrates the details of our method. The input matrix X_{n*m} represents the feature vector of the messages; n is number of messages and m is the number of terms, and Y_{n*1} is the labels of messages; $\{y_i \in \{-1, 1\}$ for $i \leq n_{labeled}$ and $y_i = 0$ for $i > n_{labeled}\}$. The output of the algorithm is $\{y_i \in \{-1, 1\}$ for $i > n_{labeled}\}$.

The $n \times n$ symmetric Gaussian affinity matrix A captures the similarity between each pair of messages $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the Euclidian distance between messages $\mathbf{x}_i$ and $\mathbf{x}_j$. A is then normalized by constructing $L = D^{-1/2}AD^{-1/2}$ [5]. The $\alpha \in (0, 1)$ parameter in line 4 of the algorithm, determines the relative amount of information that each node in the graph receives from its neighbors. It is worth mentioning that self-reinforcement is avoided since the diagonal elements of the affinity matrix are set to zero in the first step.

The main contribution of this algorithm is the aggressive approach in updating the affinity matrix. A large number of elements in the affinity matrix are approximately zero due to the large Euclidean distances between messages meaning that the messages do not share many terms. In our aggressive definition of affinity matrix, for all zero rows or columns in equation (1), a "1" (equivalently, a link in the graph) is inserted where the distance between the two corresponding messages is minimum in that column or row. Although adding a link in this case may seem too "aggressive", the simulation results show the improved performance.

Moreover, in order to better handle the sparsity of the

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Algorithm 1 Aggressive Consistency Learning Method (ACLM)

```
Input: $X$, $Y$, $\alpha$, $\sigma$
1: Compute Affinity matrix
   $$A_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}} & \text{for } i \neq j \\ 0 & \text{for } i = j, \end{cases} \quad (1)$$
2: For all $j$ such that $\sum_i A_{i,j} \approx 0$ : $A_{rj} = 1$ where $r = \arg\min_j \|\mathbf{x}_r - \mathbf{x}_j\|$
3: Compute $L = D^{-1/2} A D^{-1/2}$ where
   $$D_{ii} = \sum_{j=1}^{n} A_{ij} \,. \quad (2)$$
4: $\mathbf{Y} = (\mathbf{1} - \alpha)(\mathbf{I} - \alpha \mathbf{L})^{-\mathbf{1}} \mathbf{L}$
```

affinity matrix A, we also propose to reduce the dimensionality of matrix X. By applying Singular Value Decomposition (SVD)[2] on matrix X; we find the most informative terms in X and replace X with its approximate. In other words, $X' = U \Lambda V^{-1}$, we only keep the *rank* highest singular values of X; so $\{\Lambda_{i,i} = 0 \, \forall \, i > rank\}$.

3. EXPERIMENTS AND RESULTS

We compare the effectiveness of ACLM with the supervised and transductive modes of SVM^{light} [1] (denoted SVM and TSVM). We have compared these methods on two email corpora, TREC 2007 Public Corpus [1] and CEAS 2008 Public Corpus [2]. From each corpus we have selected the first $10,000$ from which the first 1000 were used for tuning purposes to figure out the three main parameters σ, α, and *rank*.

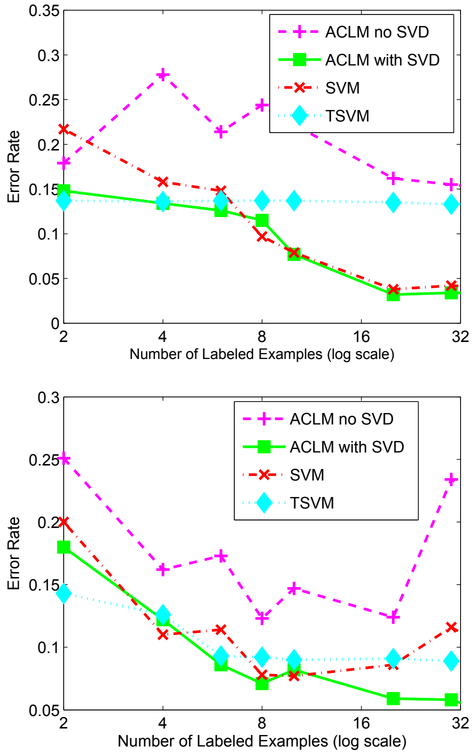
For the actual experiment, we divided the remaining 9000 messages into batches of 1000, getting 9 batches. For each batch we used the first 100 messages to select a balanced training set (same number of spam and non-spam) and the remaining 900 messages as the test set. We report mean error rate, as average over all batches.

Each message was abstracted as a binary feature vector representing word occurrences within the whole email, including headers. We removed terms with document frequency of less than 5 in the training and test sets combined. Binary term frequency was then used for the terms. Raw term frequency was also investigated, but did not provide better results than binary weights.

For parameters of SVM and TSVM, several values were adjusted but no improvement over their default values was observed. The p parameter in TSVM, representing the proportion of spam messages to be expected, was tuned using our tuning set of emails.

Fig. 1 shows the results of the methods on CEAS08 and TREC07 corpora. ACLM with SVD gives best performance of all methods between 4 and 32 labeled examples, mostly having less than 0.01 error rate. TSVM only performs best with fewer than 4 examples. We have previously seen similar results in [4] where TSVM was performing better than SVM *only* when the train and test sets were from two completely different sources. SVM does not give best performance on CEAS08 even with 30 labeled examples.

Figure 1: Error rate for ACLM (with SVD and without), SVM, TSVM on CEAS08 (up) and TREC07 (bottom) corpora



4. REFERENCES

[1] *SVM Light*. http://svmlight.joachims.org/.

[2] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. In *Proc Natl Acad Sci*, USA, 2000.

[3] C. Castillo, D. Donato, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *30st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Netherlands, 2007.

[4] M. Mojdeh and G. Cormack. Semi supervised spam filtering: Does it work? In *31st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, 2008.

[5] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pages 321–328. MIT Press.

[1] trec.nist.gov/data/spam.html

[2] www.ceas.cc/challenge