D-007

# 偽陽性率に着目したオンライン学習を用いたスパム判別
# Spam Detection Using Online Learning Focused on False Positive Rate

数原 良彦 † 鈴木 潤 ‡ 片岡 良治 †
Yoshihiko Suhara Jun Suzuki Ryoji Kataoka

## 1. はじめに

ウェブの普及に伴い，検索エンジンの検索結果に不適切なコンテンツを混在させようとするスパム業者や SEO が増加の一途を辿っている．スパム業者は，検索結果にユーザにとって意味のないコンテンツや，キーワードを散りばめたページを表示する工夫を行うため，検索エンジン提供側はこのようなウェブページを検索結果に表示しないよう，極力排除する必要がある．検索エンジンに限らず，ウェブ情報を情報源として利用する情報抽出やトレンド分析といった応用技術においてもスパムデータの影響により，解析精度が低下するおそれがある．

スパム判別のアプローチとしては，あらかじめ人手によってウェブページに対して付与されたスパム (spam) ラベル，非スパム (ham) ラベルを用いて教師あり機械学習の枠組みで判別器を生成し，未判別の文書に対して判別器を用いてスパム判別を行う方法が一般的に用いられている [1]．教師あり機械学習においては，一般的にはロジスティック回帰のような識別モデルや，SVM のような識別関数の学習手法が，ナイーブベイズなどの生成モデルに比べて高精度に学習が可能である．また，省メモリで高速に学習が可能なため，大規模なデータに対してはオンライン学習が用いられることが多い．識別関数のオンライン学習手法としては，パーセプトロンや Passive-Aggressive アルゴリズム (PA) が挙げられ，多くの派生アルゴリズムが提案されている．

実用面の観点から，(1) 大規模データにスケール可能であること，(2) 追加学習が可能であること，(3) 高速に分類が可能であることの 3 つの要件を満たすオンラインの線形識別モデルを用いることとし，スパム判別に合わせた改善を試みる．

実サービスにおいては，非スパムデータをスパムに誤分類する偽陽性 (False positive; FP) が問題となる．スパム判別においては偽陽性率を可能な限り低く抑えつつ，全体の判別正解率を向上することが望ましい．また，ラベル付きデータセット構築においては spam と ham を正確に等しくサンプリングすることは困難であるため，spam 事例と ham 事例の数に偏りが生じることがある．このような事例の偏りは，識別学習においてはしばしば問題となる．

本研究では，先述の 2 つの問題の解決を目指したオンライン学習手法を提案する．具体的にはマージン識別学習を行う Passive-Aggressive (PA) [2] においてクラス毎に異なるマージンサイズの設定を行い，False Positive (FP) を抑えた分類を目指す．また，spam クラスと ham クラスから最大損失を与える事例を交互に選択することにより，訓練データの偏りの影響を排除した学習を可能とする戦略を提案する．

†日本電信電話株式会社 NTT サービスエボリューション研究所
‡日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

**Algorithm 1** PA with Altenate Max-loss Update

**Input:** $D$, $T$, $C$, $U$, $E_{+1}$, $E_{-1}$
**Output:** $\mathbf{w}^*$
$\mathbf{w}_0 \leftarrow \mathbf{0}$
**for** $i = 1$ to $T$ **do**
  **for** $j = 1$ to $U$ **do**
    $D' \leftarrow D$
    **if** $U(i-1)+j \mod 2 = 0$ **then**
      $l \leftarrow +1$
    **else**
      $l \leftarrow -1$
    **end if**
    $(\mathbf{x}_t, y_t) = \underset{(\mathbf{x},y)\in D' \cap y=l}{\operatorname{argmax}} \ell_t(\mathbf{w}_t; \mathbf{x}, y)$
    $\tau_t = \min\left\{C, \frac{\ell_t}{\|\mathbf{x}_t\|^2}\right\}$
    $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t \mathbf{x}_t$
    $D' \leftarrow D' \setminus (\mathbf{x}_t, y_t)$
  **end for**
**end for**
$\mathbf{w}^* = \frac{1}{T \cdot U} \sum_{t=1}^{T \cdot U} \mathbf{w}_t$
**return** $\mathbf{w}^*$

## 2.Passive-Aggressive

既存の PA について説明を行う．PA では，重みベクトル $\mathbf{w}_t$ の更新を以下の最適化問題として定式化する:

$$\mathbf{w}_{t+1} = \underset{\mathbf{w}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 \quad (1)$$
$$\text{s.t.} \quad \ell_t(\mathbf{w}_t; \mathbf{x}_t, y_t) = 0.$$

ここで $\ell_t = 0$ ならば何もせず，$\ell_t > 0$ ならば，更新の大きさが最小になるように $\mathbf{w}$ を更新する．式 (1) の最適化問題は，ラグランジュの未定乗数法を用いて解くことにより，閉じた解で $\mathbf{w}_{t+1}$ を求めることができる．このように 1 回の更新を閉じた解で求めることができるのが PA の利点のひとつである．また，誤りを許容するためにスラック変数 $\xi$ を導入した場合も，同様に閉じた解で重みベクトルの更新が可能である．正則化項を $C\xi$ と設定した手法は PA-I と呼ばれ，PA と同様に閉じた解で更新が可能である [2].

$t$ 試行目の更新に用いる事例を $(y_t, \mathbf{x}_t)$ とすると損失 $\ell_t$ は，

$$\ell_t(\mathbf{w}_t; \mathbf{x}_t, y_t) = \begin{cases} 0 & \mathbf{w}_t \cdot \mathbf{x}_t \geq 1 \\ 1 - \mathbf{w}_t \cdot \mathbf{x}_t & \text{otherwise} \end{cases}$$

という hinge 損失によって計算される．

## 3. 提案手法: PA-AMU

通常の PA ではクラスによらず等しいマージンを設定し，同じ損失を利用するが，スパム分類においては FP を可能な限り小さくしたいという要求がある．そこ

表 1: 実験結果

| | | Spam | | | Ham | | |
|---|---|---|---|---|---|---|---|
| Method | Accuracy | Precision | Recall | F1 | Precision | Recall | F1 |
| LIBLINEAR | 0.686 | 0.429 | **0.930** | 0.583 | **0.968** | 0.614 | 0.746 |
| PA-I | 0.830 | 0.621 | 0.674 | 0.646 | 0.900 | 0.876 | 0.888 |
| PA-AMU | **0.857** | **0.718** | 0.624 | **0.667** | 0.892 | **0.927** | **0.909** |

で本研究では，ham クラス側のマージンサイズを spam 側に比べて大きく設定することで，ham を spam と誤分類する FP の減少を目指す．具体的には，損失 $\ell_t$ の計算を

$$\ell_t(\mathbf{w}_t; \mathbf{x}_t, y_t) = \begin{cases} 0 & \mathbf{w}_t \cdot \mathbf{x}_t \geq E(y_t) \\ E(y_t) - \mathbf{w}_t \cdot \mathbf{x}_t & \text{otherwise} \end{cases}$$

とし，クラス毎にマージンサイズを $E(y)$ で与える．

PA では最大損失を与える事例を選択的に用いて更新する戦略 (max-loss update) [2] を用いており，経験的に収束が速くなることが知られている．しかしながら，スパム分類の学習においては一般的に訓練データに含まれる各クラスの事例数が不均等であるため，最大損失を与えるクラスの事例がひとつのクラスに偏ってしまう．そのため，事例数が多いクラスの事例だけを用いてパラメータ更新を行い，分類精度が低下するおそれがある．Max-loss update の利点を活かしたまま，この問題を解決するため，本稿では spam クラスと ham クラスから最大損失を与える事例を交互に選択し，更新を行う戦略 (PA with Alternate Max-loss Update; PA-AMU) を提案する．PA-AMU では各イテレーションにおいて各クラスにおいて損失が大きい事例を $U/2$ 件ずつ選択し，それらに対して重みベクトルの更新を行う．ここで損失の計算と更新には先述の各クラスに対して設定されたマージンを利用する．

上記 2 つの改良を取り入れたアルゴリズムを Algorithm 1 に示す．PA-AMU は，訓練データ $D$，イテレーション回数 $T$，ソフトマージンパラメータ $C$，イテレーションあたりの事例選択数 $U$，マージンサイズ $E_{+1}$，$E_{-1}$ をパラメータとして与える．$T \times U$ 回だけ選択された事例について重みベクトルの更新を行った後，最後に重みベクトルの平均化を行う [3].

## 4. 評価

提案手法の有効性を検証するため評価実験を行った．スパム判別のデータセットには日本語ブログ記事に対して被験者によりスパム判定結果を付与したデータセットを利用した．各ブログ記事のタイトルと本文を単語で分割し，bag-of-words を特徴とした．その結果，本データセットは spam が 1,892 件, ham が 6,328 件で構成され，特徴ベクトルは 136,669 次元であった．

本実験では SVM をベースライン手法とした．SVM の実装は線形カーネルを高速に学習可能な LIBLINEAR[§] を利用し，訓練データにおけるサンプル数の偏りを考慮する一般的な方法であり，少ないクラスに事例数を合わせる undersampling を訓練データに適用した [4]. また，提案手法におけるマージンサイズ変更が有効に働いているか検証するため，マージンサイズ変更を行わない通常の PA-I をベースライン手法に選択した．

実験ではデータセットを 5 分割し，3 ブロックを訓練データ，1 ブロックを検証データ，残り 1 ブロックをテストデータとする 5-fold cross validation で評価を行った．LIBLINEAR，PA-I，PA-AMU における $C$ パラメータの選択 ($\{10.0, 1.0, 0.1, 0.01, 0.001\}$)，PA-I，PA-AMU のイテレーション回数 ($\{10000, 50000, 100000, 500000\}$) は検証セットにおける正解率が最大となる値を選択した．PA-AMU のパラメータは $U = 10$，$E_{+1} = 1.0$，$E_{-1} = 1.5$ とした．評価指標として正解率，各クラスに対する precision, recall と F1 値を用いた．なお，スパム判別における偽陽性率の低さは ham クラスの recall で評価することができる．

実験結果を表 1 に示す．表 1 より，PA-AMU がベースライン手法に比べて正解率，ham クラスの F1 値，spam クラスの F1 値において高い値を示した．また偽陽性の観点では，ベースライン手法に比べて PA-AMU において ham の recall が最も高く，偽陽性率が最も低い分類を実現していることを確認した．これより，LIBLINEAR や PA-I に比べて，PA-AMU によって偏りのあるデータセットについても偽陽性率を抑えた学習が可能であることが示された．

## 5. おわりに

本稿では，訓練データに偏りがある場合でも偽陽性を抑えた学習が可能なオンライン学習手法である PA-AMU を提案した．スパムブログデータに対する評価実験を通じて，LIBLINEAR や通常の PA-I に比べて両クラスの F1 値，正解率が高い結果を示した．PA-AMU は調整可能なパラメータが多いため，パラメータの変化に対して正解率と偽陽性率がどのように変化するかという分析を行い，安定した性能を持ち，調整が容易な学習アルゴリズムの開発を今後の課題とする．

## 参考文献

[1] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. WWW '06*, pp. 83–92, 2006.

[2] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithm. *Mach. Learn.*, Vol. 7, pp. 551–585, 2006.

[3] M. Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proc. EMNLP '02*, pp. 1–8, 2002.

[4] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, Vol. 21, No. 9, pp. 1263–1284, 2009.

[§] http://www.csie.ntu.edu.tw/ cjlin/liblinear/