情報処理学会論文誌　データベース　Vol.6　No.2　51–60 (Mar. 2013)

不均衡データにおける偽陽性率を考慮したスパム判別器のオンライン学習

数原 良彦[1,a)]　鈴木 潤[2,b)]　鷲崎 誠司[1,c)]

受付日 2012年9月20日, 採録日 2012年11月8日

概要：ウェブスパム判別においては，あらかじめラベル付けされた訓練データを用いて機械学習の枠組みでスパム判別器を生成する方法が広く用いられている．本稿では，ウェブスパム判別において特に課題となる偽陽性率に着目し，偏りのある訓練データを用いた場合においても偽陽性率を抑えつつ，高精度な判別が可能となるマージン識別器のオンライン学習手法を提案する．提案手法では学習時にスパムと非スパム側に異なるマージンサイズを設定することで偽陽性率を抑え，クラスを確率的に選択したうえで当該クラスにおいて最大損失を与える事例を更新に用いることで，訓練データの偏りの影響を排除しつつ高精度な学習を可能とする．本稿ではスパムブログデータセットを用いて訓練データの事例数に偏りがある場合においても提案手法によって偽陽性率を抑えた高精度なスパム判別が可能であることを示す．

キーワード：スパム分類，オンライン学習，不均衡データ

Spam Detection Using Online Learning From Imbalanced Data with the Focus on False Positive Rate

Yoshihiko Suhara[1,a)]　Jun Suzuki[2,b)]　Seiji Susaki[1,c)]

Received: September 20, 2012, Accepted: November 8, 2012

Abstract: Web spam detection systems often use supervised learning algorithms; the classifier is created from a labeled training dataset. This paper focuses on minimizing the false positive rate, a key goal in web spam detection. We propose an online learning algorithm for margin classifiers that can suppress false positives. Our method prepares different margin sizes for spam and non-spam instances to suppress the false positive rate; it stochastically selects the class to choose the instance that has maximum-loss in the selected class to eliminate the effect of imbalance in the training data and achieve high classification accuracy. We use real splog datasets to verify that our method can achieve high accuracy and low-false-positive-rate spam classification even when the training data is imbalanced.

Keywords: spam detection, online learning, imbalanced data

1. はじめに

ウェブの普及にともない，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が増加の一途をたどっている．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

[1] 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
NTT Service Evolution Laboratories, Yokosuka, Kanagawa 239–0847, Japan
[2] 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
NTT Communication Science Laboratories, Soraku-gun, Kyoto 619–0237, Japan
a) suhara.yoshihiko@lab.ntt.co.jp
b) suzuki.jun@lab.ntt.co.jp
c) suzaki.seiji@lab.ntt.co.jp

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 [13]．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 [7] などがあげられ，多くの派生アルゴリズムが提案されている．

検索エンジンにおいては常時ユーザからの膨大な数のスパムページ報告が行われていており，それらを即時反映したスパム判別器を生成したいという実用上の要件がある．また現在は，クローラにより常時大量にウェブページが取得されている状況であり，判別速度はスパム判別器の重要な性能要件といえる．このような実用面の観点から本稿では，(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，実用的なスパム判別を目指した改善を試みる．

検索エンジンなどのサービスにおいては，非スパムページをスパムと判別し，排除してしまうと，本来ユーザにとって有益な検索結果を提示できなくなる．よって，非スパムデータをスパムに誤分類することはユーザ満足度の低下につながる大きな要因である．このように非スパムデータをスパムに誤分類する偽陽性（false positive; FP）を可能な限り避ける必要がある．そのため，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させることが望ましい．たとえば spam を正例，ham を負例とすると，分類器の分類結果は正しく spam と判別した真陽性（true positive; TP），正しく ham と判別した真陰性（true negative; TN），誤って spam と判別した偽陽性（false positive; FP），誤って ham と判別した偽陰性（false negative; FN）の 4 種類に分けることができる．スパム判別の分類精度は一般的に正解率（accuracy），各クラスの適合率（precision），再現率（recall），F1 値といった指標で評価が行われる．分類の正解率は，すべての判別結果の中で正しく判別された事例の比率 $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$，spam クラスの適合率は，spam と判別された事例のうち spam である比率 $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$，spam クラスの再現率は，spam の全事例のうち，正しく spam と判別された比率 $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ で求めることができる．各クラスの F1 値は適合率と再現率の調和平均により求める．偽陽性率は ham クラスの再現率 $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ で評価することができる．本稿ではこのうち特に正解率と偽陽性に着目し，分類器の精度向上を目指す．

また，もう 1 つの課題として，ラベル付きデータセット構築を行うために spam と ham を等しくサンプリングすることは困難であるため，構築されたラベル付きデータセットにおいて spam 事例と ham 事例の数に少なからず偏りが生じることがあげられる．このような訓練事例の偏りは，識別学習においてはしばしば問題となる [9]．

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 アルゴリズムにおいてクラスごとに異なるマージンサイズの設定を行い，また，spam クラスと ham クラスから最大損失を与える事例を確率的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つつ偽陽性を抑えた高精度な学習を可能とする手法を提案する．本稿では実データから作成されたスパムブログデータセットを用いて提案手法の有効性を検証する．

本研究の貢献は以下のとおりである：

- 正例と負例から最大損失の事例を確率的にサンプリングすることにより，サンプル偏りの影響を排除しつつ高精度な学習を可能とする戦略を提案する．
- 正例と負例に対するマージンサイズを変更することにより，偽陽性率を抑えた分類を可能とし，上記戦略とあわせたオンライン学習手法を開発する．
- 実データから作成されたスパムブログデータセットにおける評価実験を通じて提案手法の組合せを網羅的に検証し，有効性の検証と傾向の分析を行う．

本稿の構成は以下のとおりである．2 章で機械学習手法を用いたスパム判別の従来手法について述べ，3 章で提案手法の詳細とアルゴリズムを述べる．4 章で評価実験について述べたのちに 5 章で考察を行い，6 章で関連研究を述べ，7 章でまとめる．

2. 機械学習を用いたスパム判別

機械学習を用いたスパム判別手法では，スパム（+1），非スパム（−1）の 2 値ラベルが付与された訓練データセットを用いて二値分類モデルの生成を行う．訓練データセット D は，N 個の事例 $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ から構成される．ここで $y \in \{-1, +1\}$ は事例に対するラベル．$\mathbf{x} \in \mathbb{R}^m$ は m 次元の特徴ベクトルを表す．この際，特徴としてはウェブページの単語の出現情報である bag-of-words 表現などを用いる．そして訓練データセット D を用いて，教師あり機械学習の枠組みで入力事例に対して予測を行うスパム分類器を学習する．

本稿で対象とする線形識別モデルは，特徴ベクトル次元数と同じ次元数[*1]を持つ重みベクトル $\mathbf{w}$ を学習によって獲得する．入力事例に対する予測は当該事例の特徴ベク

[*1] バイアス項を学習する場合には特徴ベクトル次元数より 1 つ多い次元数

トルと重みベクトルの内積によって計算し，内積の符号 $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})$ を予測ラベルとして用いる．

教師あり機械学習は訓練データの利用方法の観点で大きく 2 つに分けられ，訓練データを同時にすべて用いてパラメータ学習を行うバッチ学習と，訓練データの中から選択された事例を用いてパラメータ学習を行うオンライン学習に分けられる．一般的にはオンライン学習の方が大規模データに対して適用可能であり，かつ高速な学習が可能である．

2.1 Passive-Aggressive アルゴリズム

本節では提案手法のベースとなるオンライン学習手法である Passive-Aggressive（PA）アルゴリズムについて述べる．PA では，t 回目の更新における重みベクトル $\mathbf{w}_t$ の更新を以下の最適化問題として定式化する：

$$\mathbf{w}_{t+1} = \underset{\mathbf{w}\in\mathbb{R}^n}{\mathrm{argmin}} \frac{1}{2}||\mathbf{w}-\mathbf{w}_t||_2^2 \quad \text{s.t.} \quad \ell_t(\mathbf{w}_t;\mathbf{x}_t,y_t)=0. \tag{1}$$

ここで ℓ_t は現在の重みベクトルにおける事例の損失を表し，$\ell_t=0$ 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 1 つである．また，誤りを許容するためにスラック変数 ξ 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 と呼び [7]，よりロバストな学習が可能であることが知られている．本稿ではこれより PA-I を用いる．

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 ℓ_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．

更新式では事例の選択方法については規定されておらず，たとえば通常のオンライン学習と同様に各試行において事例をランダムに選択する方法が用いられる．また，PA では各試行において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）[7] も提案されており，収束が速くなり，予測精度が向上する場合もあることが経験的に知られている．

3. 提案手法

本稿では，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させることと，識別学習における訓練データの偏りへの対処という 2 つの課題を考慮したオンライン学習を実現する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）[7] においてクラスごとにマージンサイズの設定と，学習を行う事例の選択戦略を工夫することにより，スパム分類に適した学習アルゴリズムの開発を目指す．

3.1 マージンサイズの変更

通常の PA では正例に対しても負例に対しても同じ損失を利用するが，スパム分類においては偽陽性を可能な限り小さくしたいという要求がある．そこで提案手法では，ham 側のマージンサイズを spam 側に比べて大きくすることで ham 事例をより正しく判別するような識別超平面の学習を行い，偽陽性の減少を目指す．具体的には，損失 ℓ_t の計算を

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

のように，クラスごとにマージンサイズ $E(y)$ を設定し，これを非対称にすることによって，偽陽性を減少させるモデルの学習を目指す．マージンサイズ $E(y)$ はアルゴリズムに与えられるものとし，たとえば検証データセットを用いてパラメータ選択を行う．Li ら [12] は，パーセプトロンにおいてマージンサイズを変更する方法を提案しており，本稿ではこれを PA に取り入れ，スパム判別における有効性を検証する．

3.2 クラス選択を用いた最大損失事例の選択

2.1 節で述べたとおり，PA では最大損失を与えるペアを選択的に用いて更新する Max-loss update を用いることがある．しかしながら，スパム分類器の学習においては一般的に訓練データに含まれる事例数がしばしば不均等であるため，最大損失を与える事例のクラスが事例数が多いクラスに偏ってしまい，そのような事例を用いてパラメータ更新を行うと事例が少ないクラスの再現率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，クラスを確率的に選択し，選択されたクラスの中から最大損失を与える事例を選択して更新を行う戦略を提案し，マージンサイズの変更とともにこの戦略を PA に取り入れたアルゴリズム（Flip-Flop Max-loss update PA; FFM-PA）を開発する．

提案手法のイメージを**図 1** に示す．FFM-PA では各イテレーションにおいて，ランダムにクラスを選択し，選択されたクラスにおける最大損失を与える事例を選択し，選択された事例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各クラスに対して設定されたマージンが利用されることに注意されたい．

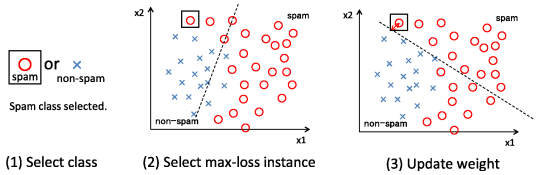

図 1　Flip-Flop Max-loss update の概要
Fig. 1　Overview of Flip-Flop Max-loss update.

Algorithm `Flip-Flop Max-loss update PA (FFM-PA)`

Input: $(\mathbf{x}_n, y_n) \in D$, T, C, $E(+1)$, $E(-1)$
Output: $\mathbf{w}^*$

1: **Initialize**: $\mathbf{w}_0 = \mathbf{0}$
2: **FOR** t in 1 to T
3: Obtain random value p from $[0, 1]$
4: **IF** $p > 0.5$ **THEN**
5: $(\mathbf{x}_t, y_t) = \underset{(\mathbf{x},y)\in D \cap y=+1}{\operatorname{argmax}} \ell_t^{uneven}(\mathbf{w}_t; \mathbf{x}, y)$
6: **ELSE**
7: $(\mathbf{x}_t, y_t) = \underset{(\mathbf{x},y)\in D \cap y=-1}{\operatorname{argmax}} \ell_t^{uneven}(\mathbf{w}_t; \mathbf{x}, y)$
8: **ENDIF**
9: Calculate $\tau_t = \min\left\{C, \frac{\ell_t^{uneven}}{||\mathbf{x}_t||_2^2}\right\}$
10: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$
11: **ENDFOR**
12: $\mathbf{w}^* = \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i$
13: **RETURN** $\mathbf{w}^*$

図 2　Flip-Flop Max-loss update PA アルゴリズム
Fig. 2　Flip-Flop Max-loss update PA algorithm.

3.3　アルゴリズム

以上の戦略を用いた PA である提案手法のアルゴリズムを**図 2** に示す．提案手法では入力として，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，正例に対するマージンサイズ $E(+1)$，負例に対するマージンサイズ $E(-1)$ を受け取り，学習後の重みベクトル $\mathbf{w}^*$ を返す．各イテレーションにおいて，クラスをランダムに選択し，選択されたクラスにおいて現在の重みベクトル $\mathbf{w}_t$ によって最大損失を与える事例を選択する（ステップ 5 またはステップ 7）．そして，選択された事例を用いて重みベクトルの更新を行う．すべての試行を終えたのちに，学習の後半で選択された事例の影響を緩和し，性能を安定させるために重みベクトルの平均化を行う [4]．

4.　実験

提案手法の有効性を検証するため評価実験を行った．

4.1　データセット

スパム判別のデータセットには日本語ブログ記事 50,000 件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．スパム判定のために，ユーザが自分自身で閲覧者にとって有用なコンテンツを記述しているかという判定基準を設け，判定基準に従って被験者が判定を行った．本判定基準におけるスパム記事の例としてコピペ，アフィリエイト，アダルト，ギャンブルに該当するブログ記事があげられる．本データセットは spam が 28,675 件，ham が 21,325 件からなる（ALL）．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．単語分割には MeCab [*2] を利用し，MeCab の辞書には IPA 辞書を用いた．また，bag-of-words とは別に当該ページのリンク出次数を特徴として用いた．ham の訓練データが spam のそれよりも多い場合の検証を行うため，上記データセットから spam ラベルが付与されたブログ記事をランダムに 20,000 件除去したデータセットを用意した（REDUCED）．2 つのデータセットの情報を**表 1** に示す．ここで #spam は spam ラベルが付与された事例数，#ham は ham ラベルが付与された事例数，#feature は特徴次元数を表す．

表 1　データセットの概要
Table 1　Dataset summary.

Dataset	#spam	#ham	#feature
ALL	28,675	21,325	199,367
REDUCED	8,675	21,325	155,344

4.2　実験条件

本実験では提案手法におけるマージンサイズ変更とクラス選択を用いた最大損失選択が有効に働いているか検証するため，以下の 2 × 3 の組合せ 6 通りについて検証を行った．

- マージンサイズ：
  - (1) 変更なし（Even margin）
  - (2) マージンサイズ変更（Uneven margin）
- 事例選択：
  - (a) ランダムサンプリング（PA）
  - (b) 最大損失選択（PA + Max-loss）
  - (c) クラス選択を用いた最大損失選択（PA + FF Max-loss）

ここで PA はすべて PA-I を利用し，ベースラインも提案手法と同様に重みベクトルの平均化（図 2 のステップ 12）を用いた．通常の PA は，(1) と (a) の組合せであり，従来の Max-loss update を用いる PA は (1) と (b) の組合せに該当する．提案手法は，(2) と (c) の組合せである．

また本実験ではバッチ学習手法である SVM [15] もベースライン手法とした．SVM の実装には線形カーネルを高速に学習可能な LIBLINEAR [*3] を利用した．LIBLINEAR のソルバには L2-regularized L2-loss SVM（dual）を用い

[*2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[*3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

た．LIBLINEAR についてはコスト考慮型学習が可能であるため，ベースラインとして 2 つのクラスを等しく扱う手法（Even cost）と 2 つのクラスのコストを変更したコスト考慮型学習（Cost-sensitive）の 2 通りを用いた．

上記すべての手法について，データセットをそのまま利用する方法（NO SAMPLING），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 under-sampling（UNDER）[9]，多いクラスに事例数を合わせる over-sampling（OVER）[9] をデータセットに適用したものを訓練データとした場合の実験も行った．すなわち，ALL，REDUCED についてそれぞれ 3 通りの訓練データ利用方法について 6 手法の比較を行った．なお，提案手法は訓練データにおいて事例の偏りがある場合の利用を想定しているため，UNDER と OVER における提案手法の評価値はあくまで参考として算出し，手法の評価には利用しない．

評価指標として分類結果の正解率，各クラスに対する適合率，再現率とそれらの調和平均である F1 値，そして ROC 曲線の下側面積である Area Under Curve（AUC）値を利用した [2]．たとえば spam クラスの事例が非常に多い際に正解率に基づいて評価すると，事例のほとんどを spam クラスと判別し，偽陽性を多く発生させるモデルが良い分類器であると判定されてしまう．また，分類器の出力に対して判別閾値を設定することにより，正解率の値は変化する．AUC は，分類器の出力値の降順に並べ替えた結果を用いて偽陽性に対する真陽性の増加具合に基づいて計算される．このため，事例の偏りや判別閾値の影響を受けずに分類器の評価を行うことが可能であり，スパム判別の評価指標でも用いられる [1]．

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，PA + Max-loss，PA + FF Max-loss における C の選択 ($\{10.0, 1.0, 0.1, 0.01, 0.001\}$)，PA のイテレーション回数 ($\{10000, 50000, 100000, 500000\}$)，PA + Max-loss，PA + FF Max-loss のイテレーション回数 ($\{1000, 5000, 10000\}$) は検証セットにおける AUC が最大となる値を選択した．また Uneven margin には $E(+1) = 1.0$ とし，$E(-1) \in \{1.5, 2.0, 3.0\}$ から同様に選択し，LIBLINEAR の Cost-sensitive では同様に spam クラスのコストを 1.0 とし，ham クラスのコストを $\{1.5, 2.0, 3.0\}$ から検証データセットを用いて選択した．

テストデータに対する各評価結果に対して，平均の差に有意差があるか Wilcoxon の順位和検定を用いて有意水準 0.05，0.01，0.001 の検定を行った．

4.3　実験 1：ALL データセットを用いた実験

実験結果を**表 2**，**表 3**，**表 4** に示す．簡単のため，各手法の組合せに記号を付与した．表 2，3，4 において，太字はその他すべての手法の結果に対して有意に高い数字であること，下線は Uneven Margin PA + FF Max-Loss を除くすべての手法の結果に対して有意に高い数字であることを表している．これらの表より，手法 (A)–(X) について以下の結果を確認した．ここで，「すべての提案手法」と言及した場合には FF Max loss update を行う (C)，(F)，(K)，(N)，(S)，(V) を指し，「すべてのベースライン手法」と言及した場合には上記提案手法以外の手法すべてを指すものとする．

- 対称マージンを用いた提案手法である (C) において，正解率，AUC，spam クラスの F1 値，ham クラスの F1 値について，提案手法を除くすべてのベースライン手法に対して高い値を示した．(J) に対しては上述の評価値すべてについて有意水準 0.05，(B) の正解率，spam クラスの F1 値においては有意水準 0.05．それ以外のベースライン手法の上記評価指標すべてについては有意水準 0.01 であった．
- 提案手法 (F) において，正解率，AUC，spam クラスの F1 値，ham クラスの F1 値について，提案手法を除くすべてのベースライン手法に対して高い値を示した（有意水準 0.01）．
- マージンサイズ変更の有無を検証するために提案手法 (F) を (C) と比較すると，正解率，AUC，spam クラスの適合率，F1 値，ham クラスの再現率，F1 値において高い値を示しているが，有意差は確認できなかった．

4.4　実験 2：REDUCED データセットを用いた実験

実験 1 では訓練データにおいて spam 事例が ham 事例よりも多いデータセットにおける実験を行った．実験 2 では反対に ham 事例が spam 事例よりも多いデータセットにおいて，同様に偽陽性率を抑えた学習が可能であるか検証するため，REDUCED データセットを用いて実験を行った．利用したデータセット以外，すべて実験 1 と同じ条件で実験を行った．

実験結果を**表 5**，**表 6**，**表 7** に示す．簡単のため，各手法の組合せに記号を付与した．各手法の記号付与規則と呼び方は実験 1 と同様である．また，表 5，6，7 における太字と下線の意味も実験 1 の結果と同様である．これらの表より，手法 (A)–(X) について以下の結果を確認した．実験 1 と同様に，ここで「すべての提案手法」と言及した場合には FF Max loss update を行う (C)，(F)，(K)，(N)，(S)，(V) を指し，「すべてのベースライン手法」と言及した場合には上記提案手法以外の手法すべてを指すものとする．

- 提案手法 (C) において，正解率についてすべてのベースライン手法に対して高い値を示した．ただし (B) に対しては有意差が見られず，それ以外のすべてについて有意水準 0.01 であった．AUC については (C) がす

表 2　ALL（NO SAMPLING）データセットにおける実験結果
Table 2　Experiment: results for ALL (NO SAMPLING) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(A)	Even margin	PA	0.8814	0.9698	0.8505	0.9629	0.9031	0.9398	0.7718	0.8471
(B)		PA + Max-loss	0.8813	0.9550	0.8791	0.9203	0.8991	0.8848	0.8288	0.8557
(C)		PA + FF Max-loss	0.9327	0.9827	0.9309	0.9534	0.9420	0.9355	0.9047	0.9198
(D)	Uneven margin	PA	0.8821	0.9697	0.8524	0.9612	0.9034	0.9376	0.7755	0.8484
(E)		PA + Max-loss	0.4265	0.5508	0.0667	0.0000	0.0001	0.4265	0.9999	0.5979
(F)		PA + FF Max-loss	**0.9379**	**0.9845**	0.9398	0.9527	**0.9462**	0.9353	0.9179	**0.9265**
(G)	LIBLINEAR	Even cost	0.8599	0.9634	0.8183	0.9715	0.8883	0.9487	0.7098	0.8120
(H)		Cost-sensitive	0.8961	0.9480	0.9395	0.8752	0.9062	0.8463	0.9242	0.8835

表 3　ALL（UNDER）データセットにおける実験結果
Table 3　Experiment: results for ALL (UNDER) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(I)	Even margin	PA	0.9021	0.9552	0.9332	0.8932	0.9128	0.8643	0.9141	0.8884
(J)		PA + Max-loss	0.8962	0.9654	0.8971	0.9259	0.9111	0.8954	0.8563	0.8752
(K)		PA + FF Max-loss	0.9367	0.9832	0.9400	0.9502	0.9450	0.9322	0.9185	0.9253
(L)	Uneven margin	PA	0.9018	0.9543	0.9352	0.8904	0.9122	0.8616	0.9170	0.8884
(M)		PA + Max-loss	0.4264	0.5516	0.0500	0.0000	0.0001	0.4265	0.9998	0.5979
(N)		PA + FF Max-loss	0.9369	0.9614	0.9388	0.9519	0.9453	0.9344	0.9165	0.9253
(O)	LIBLINEAR	Even cost	0.8956	0.9634	0.9150	0.9017	0.9083	0.8704	0.8873	0.8788
(P)		Cost-sensitive	0.8738	0.9480	0.9728	0.8023	0.8794	0.7849	0.9698	0.8676

表 4　ALL（OVER）データセットにおける実験結果
Table 4　Experiment: results for ALL (OVER) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(Q)	Even margin	PA	0.9033	0.9560	0.9336	0.8950	0.9139	0.8662	0.9144	0.8897
(R)		PA + Max-loss	0.8743	0.9525	0.8676	0.9243	0.8946	0.8864	0.8073	0.8439
(S)		PA + FF Max-loss	0.9327	0.9839	0.9309	0.9534	0.9420	0.9355	0.9047	0.9198
(T)	Uneven margin	PA	0.9031	0.9552	0.9356	0.8924	0.9135	0.8639	0.9174	0.8898
(U)		PA + Max-loss	0.4264	0.5511	0.0500	0.0000	0.0001	0.4265	0.9998	0.5979
(V)		PA + FF Max-loss	0.9373	0.9841	0.9378	0.9539	0.9457	0.9366	0.9150	0.9256
(W)	LIBLINEAR	Even cost	0.8975	0.9634	0.9182	0.9016	0.9098	0.8708	0.8919	0.8812
(X)		Cost-sensitive	0.8755	0.9480	0.9728	0.8053	0.8812	0.7875	0.9698	0.8692

べてのベースライン手法に対して高い値を示した（有意水準 0.01）．また，ham の F1 値については Max-loss update を用いたベースライン手法である (B) を除くすべてのベースライン手法に対して有意に高い値を示した（有意水準 0.01）．(B) に比べて値は高いものの，ham の F1 値については有意差が見られなかった．spam の F1 値については (B) を除くすべてのベースライン手法に対して高い値を示した（有意水準 0.01）．

- 提案手法 (F) において，正解率についてすべてのベースライン手法に対して高い値を示した．ただし (B) については有意差が見られず，それ以外は有意水準 0.01 であった．AUC について，(F) がすべてのベースライン手法に対して高い値を示した．ただし，(B) と (J) に対しては有意水準 0.05，それ以外は有意水準 0.01 であった．ham の F1 値については Max-loss update を用いたベースライン手法である (B) を除くすべてのベースライン手法に対して (F) が有意に高い値を示した（有意水準 0.01）．(B) に比べて値は高いものの，有意差が見られなかった．spam の F1 値については，すべてのベースライン手法に対して (F) が高い値を示し

表 5　REDUCED（NO SAMPLING）における実験結果
Table 5　Experiment: results for REDUCED (NO SAMPLING) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(A)	Even margin	PA	0.8746	0.8900	0.9707	0.5844	0.7291	0.8545	0.9928	0.9184
(B)		PA + Max-loss	0.8987	0.9618	0.7875	0.8991	0.8386	0.9560	0.8985	0.9261
(C)		PA + FF Max-Loss	0.9450	**0.9863**	0.8683	0.9548	0.9095	0.9808	0.9410	0.9605
(D)	Uneven margin	PA	0.8697	0.8843	0.9734	0.5648	0.7147	0.8488	0.9937	0.9155
(E)		PA + Max-loss	0.7107	0.5591	0.0000	0.0000	0.0000	0.7108	0.9998	0.8309
(F)		PA + FF Max-Loss	**0.9475**	0.9859	0.8755	0.9544	**0.9131**	0.9808	0.9446	**0.9624**
(G)	LIBLINEAR	Even cost	0.8700	0.8810	0.9753	0.5650	0.7153	0.8489	0.9942	0.9158
(H)		Cost-sensitive	0.8396	0.8441	0.9854	0.4523	0.6199	0.8173	0.9973	0.8984

表 6　REDUCED（UNDER）における実験結果
Table 6　Experiment: results for REDUCED (UNDER) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(I)	Even margin	PA	0.9100	0.9558	0.8148	0.8916	0.8514	0.9541	0.9175	0.9355
(J)		PA + Max-loss	0.8771	0.9673	0.7279	0.9272	0.8147	0.9664	0.8569	0.9080
(K)		PA + FF Max-Loss	0.9357	0.9876	0.8429	0.9557	0.8957	0.9811	0.9274	0.9535
(L)	Uneven margin	PA	0.9110	0.9549	0.8192	0.8884	0.8524	0.9530	0.9202	0.9363
(M)		PA + Max-loss	0.7107	0.5572	0.0000	0.0000	0.0000	0.7108	0.9999	0.8309
(N)		PA + FF Max-Loss	0.9331	0.9877	0.8379	0.9532	0.8917	0.9800	0.9249	0.9516
(O)	LIBLINEAR	Even cost	0.8934	0.8810	0.7699	0.9009	0.8301	0.9567	0.8904	0.9223
(P)		Cost-sensitive	0.9215	0.8441	0.9178	0.8001	0.8549	0.9227	0.9709	0.9462

表 7　REDUCED（OVER）における実験結果
Table 7　Experiment: results for REDUCED (OVER) dataset.

					Spam			Ham		
			acc.	AUC	prec.	rec.	F1	prec.	rec.	F1
(Q)	Even margin	PA	0.9111	0.9561	0.8159	0.8941	0.8532	0.9552	0.9180	0.9362
(R)		PA + Max-loss	0.8467	0.9564	0.6791	0.9271	0.7813	0.9645	0.8145	0.8816
(S)		PA + FF Max-Loss	0.9319	0.9880	0.8335	0.9554	0.8902	0.9808	0.9224	0.9507
(T)	Uneven margin	PA	0.9120	0.9554	0.8201	0.8912	0.8542	0.9541	0.9205	0.9370
(U)		PA + Max-loss	0.7107	0.5569	0.0000	0.0000	0.0000	0.7108	0.9999	0.8309
(V)		PA + FF Max-Loss	0.9355	0.9890	0.8410	0.9580	0.8956	0.9819	0.9263	0.9533
(W)	LIBLINEAR	Even cost	0.8966	0.8810	0.7771	0.9009	0.8344	0.9569	0.8949	0.9248
(X)		Cost-sensitive	0.9226	0.8441	0.9186	0.8034	0.8572	0.9239	0.9711	0.9469

た．ただし，(B) については有意差見られず，それ以外のすべてについて有意水準 0.01 であった．

- マージンサイズ変更の有無を検証するために提案手法 (F) を (C) と比較すると，正解率，spam クラスの適合率，F1 値，ham クラスの再現率，F1 値において高い値を示しているが，有意差は見られなかった．

5. 考察

本章では実験結果から得られた結果をもとに提案手法についてクラス選択を用いた最大損失選択の効果とマージンサイズ変更の効果に関する考察を行い，最後に提案手法の高速化のための並列化について行った実験と検討について述べる．

5.1 クラス選択の効果

提案手法は事前にクラスを確率的に選択し，当該クラスにおける最大損失の事例を選択する戦略を用いている．これは一見，事例数が多いクラスに少ないクラスの事例数を合わせる over-sampling や，少ないクラスに合わせる under-sampling を適用した訓練データに対して単純な

Max-loss update 戦略を用いた PA と等しいアルゴリズムであるように思える．しかしながら，評価実験より under-sampling，over-sampling を適用した訓練データにおける Max-loss update に比べて，提案手法が有意に高い精度を示した．そこで over-sampling や under-sampling における Max-loss update との違いについて考察を行い，FFM-PA が事例数の偏りを排除した訓練データに対する Max-loss update と性質が異なるアルゴリズムであることを示す．

under-sampling を訓練データに適用する場合，事例数が多いクラスの事例を除去するため，情報量が失われてしまう．そのため，サンプリングを行わない訓練データに対して FFM-PA を適用することにより，情報量を失うことなく，正例負例に対してバランス良く事例選択が可能となる．

また over-sampling を適用した訓練データに対する Max-loss update 戦略では，一見サンプリングなしの訓練データに対する FFM-PA と同じアルゴリズムに見えるものの，over-sampling ではすでに存在する事例のコピーを増やしているため，サンプリングなしの訓練データに対する Max-loss update と実質的に同じアルゴリズムとなる．すなわち，更新に利用される事例がランダムに選択される場合には，over-sampling によってクラスごとの事例数が等しくなるため，選択される事例のクラスの偏りが解消されるものの，Max-loss update を用いて最大損失を与える事例を選択する場合，サンプリング前の事例のコピーが増えたところで最大損失を与える事例には変化がない．そのため，over-sampling を適用した訓練データにおける Max-loss update では，偏りがある訓練データに対して偏りを排除するような効果が期待できず，over-sampling の恩恵が得られないと考えられる．

また，クラスを確率的に選択した後に Max-loss update を利用しない方法は UNDER や OVER に対して通常の PA を適用したものと等価であるため，クラスを確率的に選択した後に Max-loss update を行う提案手法により，ベースライン手法に比べて高精度に学習が可能であることを示すことができた．

5.2　マージンサイズ変更の効果

マージンサイズ変更の効果について考察を行う．PA，PA + Max-loss，PA + FF Max-loss について以下の傾向が得られた．

- PA：実験 1，2 における (A) と (D)，(I) と (L)，(Q) と (T) を比較すると，対称マージン/非対称マージンでほぼ同じ結果となり，すべての評価指標において有意な差を確認することができなかった．
- PA + Max-loss：非対称マージンにおいて ham に対する損失が大きく設定されたため，全体における最大損失を与える事例がほとんど ham となり，選択された ham に対して正しく分類するような超平面が学習されたため，適切に学習することができなかった．
- PA + FF Max-loss：実験 1，2 において非対称マージンの提案手法 (F) が正解率，AUC（実験 1 のみ），spam の F1 値，ham の F1 値において最高精度を達成した．また，有意差は見られなかったものの，対称マージンの提案手法 (C) に比べても上述の評価指標において高い値を示した．

これより，提案手法においてマージンサイズを変更することにより，偽陽性率を抑え，高精度な学習が可能であることを部分的に示すことができた．また，通常の PA に非対称マージンを適用することで Max-loss update 戦略の利点を得られなくなるが，提案手法ではこの問題が発生せず，安定した性能を得ることが可能であることを示した．

5.3　FFM-PA の並列化

FFM-PA において最大損失事例を選択する場合には，各イテレーションごとに選択されたクラスの全事例に対して損失を計算する必要があり，訓練データ量の増加に応じて計算コストの増加が課題となる．各事例に対する損失の計算は事例ごとに独立しているため，簡単に並列化することが可能である．具体的にはアルゴリズム（図 2）におけるステップ 5 とステップ 7 における argmax 計算の並列化を行う．実験では `OpenMP` [*4] を用いて最大損失計算のための各事例に対する損失計算の並列化を行った．実験に用いた計算機は Intel Xeon 5570 2.93 GHz CPU x 2，48 GB メモリで OS は Scientific Linux（Linux Kernel 2.6.32），gcc 4.4.6 と OpenMP 200805 を利用し，コンパイルには O3 オプションを用いた．Xeon5570 は 4 コア HT であるため，最大 16 スレッドまで同時に利用可能である．

ALL データセットすべてを訓練データとして利用し，提案手法のイテレーション回数は 1,000 とし，スレッド数は 1 から 16 までの 16 通りの計算を行った．重みベクトル更新にかかる時間を `gettimeofday(2)` を用いて計測を 10 回ずつ行った．平均時間と標準偏差を**図 3** に示す．

図 3 より，11 スレッドまでは訓練時間が減少し，12 スレッド以上の場合においてはかえって訓練時間が増加するという結果が得られた．これはスレッド生成のオーバヘッドなどのコストがスレッド分割による効果を上回ったためだと考えられる．また，スレッド数 11 の際に平均 11.97 秒という結果が得られ，並列化なしの平均 30.06 秒の約 2.5 倍の速度で計算することができた．ただし，ランダムサンプリングの PA ではイテレーション回数が 500,000 において平均 0.47 秒という速度であるため，速度面では改善の余地がある．

[*4] http://openmp.org/wp/

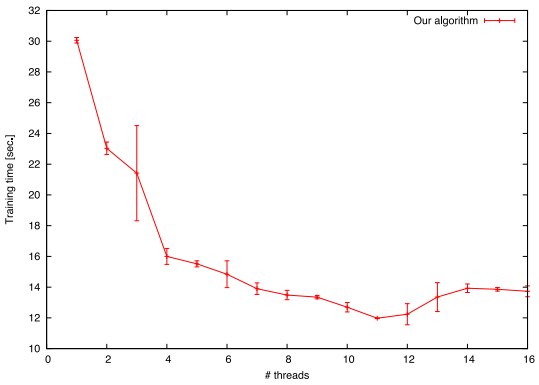

図 3　FFM-PA のスレッド数ごとの訓練時間

Fig. 3　Training time of FFM-PA with different numbers of threads.

6. 関連研究

ウェブスパムの自動判別手法は内容解析，リンク解析などさまざまなアプローチで行われてきた [14]．内容解析のアプローチでは，ウェブページのコンテンツから抽出した特徴を用いて分類器で自動的に判別を行う [5], [13]．またスパムブログを対象にした研究もある [10]．リンク情報を用いたアプローチでは，リンクファームの情報を用いるもの [17] や，非スパムページはスパムページにリンクしにくいという性質を利用した TrustRank などの研究があげられる [3], [8]．その他の方法として，スパムページを取得する際の HTTP サーバの応答ヘッダを利用することで内容を解析せずにスパムページを判別する手法 [6], [16] もある．しかしながら，HTTP ヘッダを利用する手法では，たとえばブログサービスのように非スパムブログもスパムブログも同じサーバが応答するため，必ずしも汎用的ではないため，本研究の対象外とする．

不均衡データに対する学習手法の一般的な方法として少ないクラスの事例を復元抽出して多いクラスに合わせる方法（over sampling），多いクラスの事例をサンプリングして少ないクラスに合わせる方法（under sampling）の 2 つがあげられる [9]．オンライン学習における不均衡データの既存研究として Li ら [11] は情報抽出タスクにおいて偏りのある訓練データに対して非対称なマージンを用いた SVM とパーセプトロンを適用している．本研究では，非対称マージンを PA に取り入れた点，そしてスパム判別タスクにおける非対称マージンの有効性を検証した点において異なる．

7. 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線形識別モデルのオンライン学習手法である FFM-PA を提案した．

FFM-PA では，正例と負例に対するマージンサイズを変更し，かつイテレーションごとにランダムに選択されたクラスから最大損失を与える事例を重みベクトルの更新に用いて PA に基づくモデル学習を行うことにより，偏りのある訓練データにおいても高精度な学習を可能とする．

スパムブログデータに対する評価実験を通じて，under-sampling や over-sampling を行った訓練データを用いた通常の PA，Max-loss update を用いた PA，コスト考慮型学習を用いた LIBLINEAR に比べて正解率，AUC，両クラスの F1 値において高い値を示した．考察を通じて，FFM-PA が under-sampling，over-sampling を行った訓練データに対する Max-loss update と異なることを確認し，また並列化による高速化について検証を行った．今後の課題としては，全訓練データに対して最大損失を与える事例を選択するのではなく，近似的に最大損失を選択することでモデルの性能を保ちつつ，高速に学習な方法を検討する予定である．

参考文献

[1] Abernethy, J., Chapelle, O. and Castillo, C.: Web spam identification through content and hyperlinks, *Proc. 4th international workshop on Adversarial information retrieval on the web, AIRWeb '08*, pp.41–44 (2008).
[2] Büttcher, S., Clarke, C.L.A. and Cormack, G.V.: *Information Retrieval: Implementing and Evaluating Search Engines*, MIT Press (2009).
[3] Castillo, C., Donato, D., Gionis, A., Murdock, V. and Silvestri, F.: Know your neighbors: Web spam detection using the web topology, *SIGIR '07: Proc. 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.423–430 (2007).
[4] Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms, *Proc. EMNLP '02*, pp.1–8 (2002).
[5] Cormack, G.V.: Content-based web spam detection, *AIRWeb '07: Proc. 3rd International Workshop on Adversarial Information Retrieval on the Web* (2007).
[6] Cormack, G.V., Smucker, M.D. and Clarke, C.L.A.: Efficient and effective spam filtering and re-ranking for large web datasets, *Inf. Retr.*, Vol.14, No.5, pp.441–465 (2011).
[7] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Singer, Y.: Online Passive-Aggressive Algorithm, *Mach. Learn.*, Vol.7, pp.551–585 (2006).
[8] Gyongyi, Z., Berkhin, P., Garcia-Molina, H. and Pedersen, J.: Link spam detection based on mass estimation, *VLDB '06: Proc. 32nd international conference on Very large data bases, VLDB Endowment*, pp.439–450 (2006).
[9] He, H. and Garcia, E.A.: Learning from Imbalanced Data, *IEEE Trans. Knowl. Data Eng.*, Vol.21, No.9, pp.1263–1284 (2009).
[10] Kolari, P., Java, A., Finin, T., Oates, T. and Joshi, A.: Detecting spam blogs: A machine learning approach, *Proc. 21st national conference on Artificial intelligence AAAI'06*, Vol.2, pp.1351–1356, AAAI Press (2006).
[11] Li, Y., Bontcheva, K. and Cunningham, H.: Using uneven margins SVM and perceptron for information ex-

traction, *Proc. 9th Conference on Computational Natural Language Learning, CONLL '05*, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.72–79 (2005).
[12] Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J. and Kandola, J.S.: The Perceptron Algorithm with Uneven Margins, *Proc. 19th International Conference on Machine Learning, ICML '02*, San Francisco, CA, USA, pp.379–386, Morgan Kaufmann Publishers Inc. (2002).
[13] Ntoulas, A., Najork, M., Manasse, M. and Fetterly, D.: Detecting spam web pages through content analysis, *Proc. WWW '06*, pp.83–92 (2006).
[14] Spirin, N. and Han, J.: Survey on web spam detection: Principles and algorithms, *SIGKDD Explor. Newsl.*, Vol.13, No.2, pp.50–64 (2012).
[15] Vapnik, V.: *Statistical learning theory*, Wiley (1998).
[16] Webb, S., Caverlee, J. and Pu, C.: Predicting web spam with HTTP session information, *CIKM '08: Proc. 17th ACM conference on Information and knowledge mining*, pp.339–348 (2008).
[17] Wu, B. and Davison, B.D.: Identifying link farm spam pages, *WWW '05: Proc. World Wide Web conference*, pp.820–829 (2005).

数原 良彦　（正会員）

2006 年慶應義塾大学理工学部管理工学科卒業．2008 年同大学大学院理工学研究科開放環境科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスエボリューション研究所に所属．主として情報検索，機械学習に関する研究開発に従事．人工知能学会，言語処理学会各会員．

鈴木 潤　（正会員）

1999 年慶應義塾大学理工学部数理科学科卒業．2001 年同大学大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話株式会社入社．2005 年奈良先端大学院大学博士後期課程修了．2008～2009 年 MIT CSAIL 客員研究員．現在，NTT コミュニケーション科学基礎研究所に所属．博士（工学）．主として自然言語処理，機械学習に関する研究に従事．ACL，言語処理学会各会員．

鷲崎 誠司　（正会員）

1988 年名古屋大学理学部数学科卒業．同年日本電信電話株式会社入社．自然言語処理，情報検索に関する研究開発に従事．現在，NTT サービスエボリューション研究所主幹研究員．

（担当編集委員　比戸 将平）