Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

# SPA アルゴリズムの半教師あり学習への応用

松島 慎[†1] 清水 伸幸[†1] 佐藤 一誠[†1]
二宮 崇[†1] 中川 裕志[†1]

ここではオンライン学習における半教師あり学習の学習アルゴリズムを提案する．オンライン学習はデータが大量に手に入る状況で有効利用される学習法であり，半教師あり学習はラベルのないデータが多く手に入る場合にも有効な学習法である．この 2 つの学習は共に現実的な問題設定を指向して考案された学習方法であるが，オンラインであり，かつ半教師ありである場合の学習モデルは今まで提案されていなかった．代表的な教師ありの設定におけるオンライン学習は，正解ラベルのデータによる条件付き確率を推定していると考えることができる．そのため，これらの枠組みで正解ラベルのないデータを簡単に利用することはできなかった．本論文ではオンライン教師あり学習の有効な方法である SPA アルゴリズムを応用してオンライン半教師あり学習の方法を提案する．

# Online Semisupervised Learning on Multiclass Classificatio

SHIN MATSUSHIMA,[†1] NOBUYUKI SHIMIZU ,[†1] ISSEI SATO ,[†1]
TAKASHI NINOMIYA [†1] and HIROSHI NAKAGAWA[†1]

We analyze the combination of online learning and semi-supervised learning. Online learning is an approach to supervised machine learning, which is effective at processing large amount of training data. Semi-supervised learning is employed when large amount of unlabeled data. While both setting occurs frequently in real applications of machine learning, thus far, no one has analyzed the combination of online learning and semi-supervised learning. Our analysis provides valuable insights into the combination of two approaches

†1 東京大学
The University of Tokyo

## 1. はじめに

ある対象がどのカテゴリに属するかという識別を自動化することは，多くの場面で我々に利便性をもたらしてきた．そのための方法の多くは定められた規則に従って自動的に判断するアルゴリズムを用いて行われている．しかし，文字認識や情報検索などの場面での識別は，判断のための規則が非常に複雑であり，どのような判断に基づいて識別を行うべきかを正確に書き下し規則を定めることが困難であるため，簡単にはそのようなアルゴリズムを構成できない．このような場合に，データと各データに対して割り振られる正解クラスとの組で構成された教師データの集合を用意し，この集合から自動的に規則を生成することを考える．言い換えると，教師データを用いてデータの正解クラスを識別する規則（識別関数）をどのように学習するかという問題を考えることになる．これを識別問題といい，特に３つ以上のクラスから正解クラスを割り当てる問題を多クラス識別問題という．本論文では特に断らない限り，この多クラス識別問題を扱う．

多クラス識別問題において識別関数を学習するための方法は古くから研究されており，多くの方法が提案されているが[1)–3)]，それらの方法はオンライン学習とバッチ学習に大別することができる．オンライン学習とは逐次的に単一のデータを受け取る前提で，毎回次のデータのための識別関数を更新していく方法であり，バッチ学習は一度に全てのデータを受け取った前提で，一つの最適な識別関数を出力する方法である．多クラス識別問題におけるバッチ学習の代表的な方法としては，多クラス SVM や多クラスロジスティック回帰などがあげられ[3),4)] 一方オンライン学習の場合はパーセプトロン[5)] や PA アルゴリズム[6)] などの方法が提案されている．バッチ学習は同じデータを用いて学習を行った場合にオンライン学習に比べて精度のよい識別関数を導くことができるため，これまで重要視されてきた．しかし，昨今ではオンライン学習の重要性が再び認識されてきている．

現実的な問題として，バッチ学習は大量のデータを扱う場合や，追加のデータと既存のデータと合わせて識別関数を再び出力しなおす場合に多大な時間を要するということがある．昨今では大容量記憶媒体やインターネットの普及により，爆発的に大量のデータが得られるようになった．このようなデータを利用する際にも，バッチ学習では実用的な時間で学習ができないがオンライン学習においては計算可能である場合も多く，オンライン学習の有用性が期待されている．さらなるオンライン学習のメリットは，明確に訓練と予測のフェーズにわかれておらず識別を行いながら識別関数を学習することができる点である．ユビキタス環境においてセンサーを用いてデータを取りながらリアルタイムでそれらを識別しなけ

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

ればならない場合などにも，オンライン学習ならば識別関数の学習が可能である．さらに，バッチ学習では時系列で性質が変化する様なデータを扱うことができないのに対し，オンライン学習は時系列に沿った識別が可能であり，時間による状況の変化を考慮するのにも適した学習法といえる．

別の現実的な問題として，教師データは正解クラスを人手で判断して付与されており入手に多大なコストがかかる場合や，正解クラスが欠損しているデータが多く含まれる場合がある．そのため最近では教師データに加えてクラスの情報が与えられていないデータも利用しながら識別関数を学習する試みが注目されており，これらを教師データのみを用いる教師あり学習と区別して半教師あり学習と呼ぶ．実際に今まで様々なモデルの半教師あり学習の方法が提案されており[7)–9)]，実際に正解の付与されていないデータが学習の性能を向上させることが報告されている．しかしながら，これらの方法はバッチ学習の枠組みでの学習方法でありオンライン学習においてラベルなしデータを利用する枠組みは，未だ十分な議論がされていないのが現状である．

多クラス識別問題において，これらのオンライン学習と半教師あり学習はともに現実的な問題を指向して考案された学習法である．しかしながら，これらの問題が重畳した状況に対しては今まで十分に注目されて来なかった．画像認識における顔表情認識，人種・年代認識などのタスクや，ウェブブログの著者年代の識別問題など，これらの問題が同時に起こる場合も多く存在する．そこで我々は多クラス識別問題におけるオンライン半教師あり学習の手法として，多クラス識別問題の教師あり学習法のひとつである SPA アルゴリズムを半教師あり学習に拡張した方法を提案する．

バッチ学習において半教師あり学習を多クラス識別問題において有効に行う為の方法としていくつかの方法が提案されている．提案する手法では，同時に並列に存在する複数の識別関数が，正解のないラベルに対し同じ様な予測を行うように識別関数を逐次的に更新していく．PA アルゴリズムなどのオンライン学習が有用視されるもう一つの理由として，有限個の教師データ集合に対し，それが識別可能な集合である場合に関するいくつかの性能の理論的な保証がされていることが挙げられる．よってこれらの半教師あり学習に対しても同様に性能の理論的な保証がされることが望ましい．本提案手法では，半教師あり学習をオンラインで行うにあたって，教師データに対する誤識別回数が有限の値で上から抑えられることを証明した．この時の値は，6) や，10) の教師あり学習における上限の値と一致し，半教師あり学習がその限界を乱さないことがわかる．また，今まで性能が保証されているのは固定された教師データ集合に対する場合のみであったが，前述のように実際にオンライン学習を有効利用する場合は，次々とデータが生成されていくようなストリームデータに対する識別を行う場合である．本論文ではこのようなストリームデータに対する性能の保証に拡張できることを示し，また提案手法におけるストリームデータに対する識別の保証を行った．

具体的には，(1) 我々はオンライン学習の並列化を利用して，正解のないデータを用いても識別関数のパラメータが必ず収束するようにアルゴリズムを導き，(2) 上述の方法が，データを無限に受けとる極限でデータを分類する真の関数に近づくことを，教師あり・半教師あり学習の両方の枠組みにおいて証明した．さらに，(3) 実際のデータを用いた計算機実験により，本手法によるオンライン半教師あり学習法が有用であることを示した．

以下，2 章では SPA アルゴリズムと PA アルゴリズムの概要について説明する．3 章ではこれらの従来の収束の証明の内容と，それらを用いた半教師あり学習のアルゴリズムの導出を行う．4 章では，データを無限に受けとる極限での収束性を証明を述べる．5 章で実験結果を示し，6 章でまとめを行う．

## 2. PA アルゴリズムと SPA アルゴリズム

最初に，オンライン教師あり学習における予測と学習の設定を説明する．学習は次の工程を再帰的に繰り返すことによって行われる．

(1) $i$ 番目のデータ $\mathbf{x}_i \in X \subset \mathbb{R}^D$ を受け取る．

(2) $i$ 番目の識別関数 $h^{(i)}$ によって $\mathbf{x}_i$ に対する予測クラス $p_i \in Y = \{1,2,\cdots,K\}$ を計算する．

(3) $i$ 番目の正解クラス $y \in Y = \{1,2,\cdots,K\}$ を受け取り，次の識別関数 $h^{(i+1)}$ を求める．

ここで識別関数は線形関数のみを考える．すなわち，学習アルゴリズムが出力しうる識別関数 $h^{(i)} : X \to Y$ は $K$ 本の重みベクトル $\mathbf{w}_u^{(i)} \in \mathbb{R}^D (u = 1,2,\cdots K)$ を用いて，以下のように表されるとする．

$$p_i = \arg\max_{u \in Y} \left( \mathbf{w}_u^{(i)} \cdot \mathbf{x} \right) \tag{1}$$

ここで $\mathbf{w} = \{\mathbf{w}_u\}_{u \in Y}$ とした．以下でも誤解のない範囲でこの記法を用いる．式からわかるように，$\mathbf{x}$ を与えられたとき，識別関数 $h_{\mathbf{w}}$ は全てのクラスの重みベクトルと $\mathbf{x}$ との内積を求め，$\mathbf{w}_u \cdot \mathbf{x}$ が最大となる $u$ を選ぶ．この予測の際に比べる値 $\mathbf{w}_u \cdot \mathbf{x}$ を $u$ の（$\mathbf{x}$ における）スコアという．

PA アルゴリズムの枠組みにおける逐次学習の設定では，$\mathbf{x}_i$ に対応する正解ラベル $y_i$ を受け取り，このラベルと各クラスのスコアを用いて定義される損失を計算する．この損失を表

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

現する関数を損失関数と呼び，$\ell(\mathbf{w};\mathbf{x},y)$ と表現する．ここで損失は非負値とする．すなわち，$\ell(\mathbf{w};\mathbf{x},y)\geq 0$ である．PA アルゴリズムの枠組みでは，損失関数に従い，次の識別関数に関する重みベクトルを以下の最適化問題を用いて特徴づける．

$$\mathbf{w}^{(i+1)}=\min_{\mathbf{w}}\frac{1}{2}\sum_{u=1}^{K}\left\|\mathbf{w}_u-\mathbf{w}_u^{(i)}\right\|^2 \quad s.t.\ \ \ell(h_{\mathbf{w}}(\mathbf{x},y))=0 \tag{2}$$

この最適化問題問題の解は，損失が最小値 0 をとるような重みベクトルの中で最もデータを受け取った時点での重みベクトルからの変更が少ない重みベクトルとなる．一般的な PA アルゴリズムの枠組みにおいてはこの繰り返しで識別関数の学習を逐次的に行う．PA アルゴリズムの枠組みは損失関数を適切に定義することにより，2 値分類やランキングなどのタスクでも同様に識別関数を学習することが可能であるが，特に Crammer らは，$K$ クラス識別問題における PA アルゴリズム[6] で用いる損失関数を以下のように定義している．

$$\ell(\mathbf{w};(\mathbf{x},y))=\max_{u\neq y}\left[1-\left(\mathbf{w}_y\cdot\mathbf{x}-\mathbf{w}_u\cdot\mathbf{x}\right)\right]_+$$

$[\bullet]_+$ は $\max(\bullet,0)$ で定義される閾値関数である．この時，最適化問題 (2) は $K-1$ 本の線形制約式を持った凸最適化問題として以下のように定式化される．

$$\mathbf{w}^{(i+1)}=\arg\min_{\mathbf{w}}\frac{1}{2}\sum_{v=1}^{K}\left\|\mathbf{w}_v-\mathbf{w}_v^{(i)}\right\|^2 \quad \text{s.t.} \quad \left(\mathbf{w}_{y_i}\cdot\mathbf{x}_i-\mathbf{w}_u\cdot\mathbf{x}_i\right)\geq 1 \quad (\forall u\neq y_i). \tag{3}$$

これを，次のように近似し，以下の問題を解くのが PA アルゴリズムである．ここで $p_i$ は式 () で定義される予測クラスである．

$$\mathbf{w}^{(i+1)}=\arg\min_{\mathbf{w}}\frac{1}{2}\sum_{v=1}^{K}\left\|\mathbf{w}_v-\mathbf{w}_v^{(i)}\right\|^2 \quad \text{s.t.} \quad \left(\mathbf{w}_{y_i}\cdot\mathbf{x}_i-\mathbf{w}_{p_i}\cdot\mathbf{x}_i\right)\geq 1. \tag{4}$$

この解析解は以下のようになる．

$$\begin{aligned}
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)}+\tau\mathbf{x}_i \ \ (v=y_i)\\
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)}-\tau\mathbf{x}_i \ \ (v=p_i)\\
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)} \qquad (v\neq p_i,y_i)
\end{aligned}$$

ここで，

$$\tau=\left[\frac{1-\left(\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x}_i-\mathbf{w}_u^{(i)}\cdot\mathbf{x}_i\right)}{2\left\|\mathbf{x}_i\right\|^2}\right]_+.$$

である．

一方，我々は最適化問題 (3) を厳密に解くことができることを 10) で示した．この解に従い識別関数の重みベクトルを更新するのが SPA アルゴリズムという．この解析解は以下のようになる．

**表 1** PA アルゴリズムと SPA アルゴリズム
Table 1 Summary of PA and SPA algorithms

| | PA | SPA |
|---|---|---|
| Optimization | $\min_{\mathbf{w}}\frac{1}{2}\sum_{u=1}^{K}\left\|\mathbf{w}_u-\mathbf{w}_u^{(i)}\right\|^2$ s.t. $\mathbf{w}_{y_i}\cdot\mathbf{x}_i-\mathbf{w}_{p_i}\cdot\mathbf{x}_i\geq 1$, | $\min_{\mathbf{w}}\frac{1}{2}\sum_{u=1}^{K}\left\|\mathbf{w}_u-\mathbf{w}_u^{(i)}\right\|^2$ s.t. $\forall u\neq y_i, \mathbf{w}_{y_i}\cdot\mathbf{x}_i-\mathbf{w}_u\cdot\mathbf{x}_i\geq 1$ |
| Stepsize | $\tau=\frac{1-\left(\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x}_i-\mathbf{w}_u^{(i)}\cdot\mathbf{x}_i\right)}{2\left\|\mathbf{x}_i\right\|^2}$ | $\tau_v=\left\|\mathbf{x}_i\right\|^{-2}\left(\ell_v-\frac{1}{\lvert S\rvert+1}\sum_{u\in S}\ell_u\right)$ |
| Support class | $p_i$ | $\sigma(k):\ \sum_{j=1}^{k-1}\ell_{\sigma(j)}<k\ell_{\sigma(k)}$ |

ようになる．

$$\begin{aligned}
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)}+\sum_{u\in S}\tau_u\mathbf{x}_i \qquad (v=y_i)\\
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)}-\tau_v\mathbf{x}_i \qquad (v\in S)\\
&\mathbf{w}_v^{(i+1)}=\mathbf{w}_v^{(i)} \qquad (v\notin S).
\end{aligned}$$

ここで，

$$\begin{aligned}
&\tau_v=\frac{1}{\left\|\mathbf{x}_i\right\|^2}\left(\ell_v-\frac{1}{\lvert S\rvert+1}\sum_{u\in S}\ell_u\right)\\
&\ell_v=\left[1-\left(\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x}_i-\mathbf{w}_v^{(i)}\cdot\mathbf{x}_i\right)\right]_+\\
&S=\left\{\sigma(k)\in Y\,\middle|\,\sum_{j=1}^{k-1}\ell_{\sigma(j)}<k\ell_{\sigma(k)},\sigma:\{1\cdots K\}\to Y s.t.\left(k<k'\Rightarrow\ell_{\sigma(k)}>\ell_{\sigma(k')}\right)\right\}
\end{aligned}$$

である．この $S$ をサポートクラス集合と呼び，$S$ の要素クラスすべてに対し重みベクトルの更新を行う．$S$ の探索は簡単なアルゴリズムで達成できる．これを表 2 に示す．

## 3. 半教師あり学習への応用

本論文では，前述の教師あり学習の問題設定に加え，データ $\mathbf{x}$ のみが与えられ，それに対応する正解データ $y$ が与えられない場合にも何らかの更新を行い精度の向上を目指す半教師あり学習方法を提案する．具体的には，正解なしデータが与えられた時にそのデータに対するラベル付けが一貫するように同時に存在する識別関数がそれぞれの関数を調整する方法を提案する．これと同様の方法はバッチ学習における半教師あり学習でも考えられており，実際のタスクによる成果が 7)–9) などによって報告されている．

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

**表 2** SPA アルゴリズム
Table 2 SPA algorithm

**procedure** *SPA*
//Initialization
$\mathbf{w}_v^{(1)} = \mathbf{0}$ $(\forall v \in Y)$
**foreach** $i = 1, 2, \cdots$ **do**
//Receive an Instance $\mathbf{x}_i$
//Predict a Class $p_i$
$p_i = \arg\max_{u \in Y} (\mathbf{w}_u^{(i)} \cdot \mathbf{x}_i)$
//Receive a True Class $y_i$
//Compute Classwise Loss $\ell_u$
$\ell_v := \left[1 - \left(\mathbf{w}_v^{(i)} \cdot \mathbf{x}_i - \mathbf{w}_{y_i}^{(i)} \cdot \mathbf{x}_i\right)\right]_+ \quad (v \neq y_i);$
//Search support class $S$
Compute $j$-th class $\sigma(j) \in Y \backslash \{y_i\}$ in descending order of $\ell_v$ .
$S := \varnothing;$
**while** $\sum_{j=1}^{|S|} \frac{\ell_{\sigma(j)}}{|S|+1} < \ell_{\sigma(|S|)}$ **do**
$S := S \cup \sigma(|S|);$
**end while**
//Set Stepsize $\tau$
$\tau_v := \ell_v - \sum_{u \in S} \frac{\ell_u}{|S|+1} \qquad (v \in S)$
$\tau_v := 0 \qquad (v \notin S)$
//Update $\mathbf{w}$
$\mathbf{w}_v^{(i+1)} := \mathbf{w}_v^{(i)} + \sum_{u \neq y_i} \tau_u \mathbf{x}_i \qquad (v = y_i);$
$\mathbf{w}_v^{(i+1)} := \mathbf{w}_v^{(i)} - \tau_v \mathbf{x}_i \qquad (v \neq y_i);$
**end foreach**

**表 3** オンライン Bayes Point Machine(教師あり学習)
Table 3 Online Bayes Point Machines

**procedure** *OBPM*
$\mathbf{w}_{r,u}^{(1)} = \mathbf{0}, \ 0 < \pi < 1$
**foreach** $i = 1, 2, \cdots$ **do**
//Receive an Instance $\mathbf{x}_i$
//Predict a Class $p_i$
$\tilde{\mathbf{w}}_u^{(i)} = \sum_{r=1}^{M} \mathbf{w}_{r,u}^{(i)}$
$p_i = \arg\max_{u \in Y} (\tilde{\mathbf{w}}_u^{(i)} \cdot \mathbf{x}_i)$
//Receive a True Class $y_i$
//Update $\mathbf{w}$
**foreach** $r = 1, \cdots M$ **do**
Draw $\mathrm{Z} \sim$ **Bernoulli**$(\pi)$
**if** $\mathrm{Z} = 1$
$\{\mathbf{w}_{r,v}^{(i+1)}\}_{v \in Y} = \mathsf{Supervised\text{-}Learn}(\{\mathbf{w}_{r,v}^{(i+1)}\}_{v \in Y}, \mathbf{x}_i, y_i);$
**end foreach**
**end foreach**

### 3.1 並 列 化

オンライン学習を特定の教師データ集合について行い，一つの識別関数を導く方法にはいくつかの方法が知られている[11]．最も簡単にはデータ集合を繰り返し反復させて識別関数を更新させることにより識別関数が収束し，このときの識別関数の識別精度がよいことが経験的に知られている．しかし，現実的な問題として，オンライン学習法をストリームデータに適用する場合，反復をすることはできない．そのためオンライン線形分類アルゴリズムを用いて反復を用いずによい識別関数をを行うための方法がいくつか提案されており，Online Bayes Point Machine(OBPM) はその手法のひとつである[12]．この方法は基となる重みベクトルの学習手法を用いて，ベイズ的に最もよい識別関数を示す点である Bayes Point の近似値を求める方法である[13]．与えられた $\mathbf{x}$ の事前分布, 識別関数 $h$ 全体の集合 $\mathscr{H}$ とその事前分布, 及びロス関数 $l$ において，Bayes Point となる識別関数 $h_{bp}$ は次のように定義される．

$$h_{bp} = \arg\min_{h^* \in \mathscr{H}} \int\int l(h^*(\mathbf{x}), h(\mathbf{x}))\, p(\mathbf{x}) p(h) d\mathbf{x} dh$$

今多クラス線形分類器のみを考え，さらにデータの生成分布を一様であると仮定し，重みベクトル全体の空間 $\mathbb{R}^{KD}$ に対し $l(a,b) = [\![a = b]\!]$ を考える．また，与えられた教師データ集合の Version Space $V = \{\mathbf{w} \in \mathbb{R}^{KD} | \mathbf{w}_{y_i} \cdot \mathbf{x} = \max_{v \in Y} \mathbf{w}_v \cdot \mathbf{x}_i \ (1 \leq \forall i \leq N)\}$ が十分狭く，事後分布が鋭く尖っていると考えてよい場合，多クラス線形分類器における Bayes Point は，Version

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

**表 4** オンライン半教師あり学習
Table 4 Online-Semisupervised algorithm

**procedure** *OSS*
  **foreach** $i = 1,2,\cdots$ **do**
//Receive an Instance $\mathbf{x}_i$
//Predict a Class $p_i$
    $\tilde{\mathbf{w}}_u^{(i)} = \sum_{r=1}^{M} \mathbf{w}_{r,u}^{(i)}$
    $p_i = \arg\max_{u \in Y}(\tilde{\mathbf{w}}_u^{(i)} \cdot \mathbf{x}_i)$
//Update $\mathbf{w}$
  **If** Received a True Class $y_i$
    **foreach** $r = 1,\cdots M$ **do**
      Draw $\mathsf{Z} \sim$ **Bernoulli**$(\pi)$
      **if** $\mathsf{Z} = 1$
      $\{\mathbf{w}_{r,v}^{(i+1)}\}_{v \in Y} = \mathsf{Supervised\text{-}Learn}(\{\mathbf{w}_{r,v}^{(i)}\}_{v \in Y}, \mathbf{x}_i, y_i)$;
    **end foreach**
  **else**
    $\{\mathbf{w}_{r,v}^{(i+1)}\}_{v \in Y, 1 \le r \le M} := \mathsf{Semisupervised\text{-}Learn}(\{\mathbf{w}_{r,v}^{(i)}\}_{v \in Y, 1 \le r \le M}, \mathbf{x}_i)$
  **end foreach**

Space の重心でよく近似される．すなわち，

$$\mathbf{w}_{bp} \simeq \mathbf{w}_{cm} = \frac{\int_{\mathbf{w} \in V} \mathbf{w} p(\mathbf{w}) d\mathbf{w}}{\| \int_{\mathbf{w} \in V} \mathbf{w} p(\mathbf{w}) d\mathbf{w} \|}$$

となることを利用する．この $\mathbf{w}_{cm}$ の近似値を効率よく求める方法が提案されている．また，線形分類器のオンラインアルゴリズムを用いてこれをオンライン手法に拡張したのが OBPM である．OBPM は基となるオンライン線形分類器を自由に変えることで，別のアルゴリズムを簡単に導くことができる．この更新式の拡張は次で述べるように，元のアルゴリズムの決定的な逐次更新式を確率的な更新式に変形することで得られる．まず，基となる重みベクトルの更新式が次のように表されているとする．

$$\mathbf{w}_v^{(i+1)} = \mathbf{w}_v^{(i)} + \alpha_v^{(i)} \mathbf{x}_i \tag{5}$$

この時パラメータ $\pi$ をもったベルヌーイ分布に従う 0-1 値確率変数 $\mathsf{Z}$ を用い更新式を次のように確率化する．

$$\mathbf{w}_v^{(i+1)} = \mathbf{w}_v^{(i)} + \mathsf{Z} \alpha_v^{(i)} \mathbf{x}_i \tag{6}$$

このとき重みベクトルは確率変数となるが，実際には確率変数 $\mathsf{Z}$ は，毎回サンプリングすることによって得られた実現値を用いる．すなわち，初期値を $\mathbf{w}_v^{(1)} = \mathbf{0}$ $(v = 1 \cdots K)$ とする識別器を $M$ 個用意する．各識別関数の $i$ 番目の重みベクトルを $\mathbf{w}_{r,v}^{(i)}$ $(r = 1..M)$ と表し，これらを同時に式 (5) に従い更新する．このとき確率変数 $\mathsf{Z}$ の実現する値は多様性を持つことになるので，重みベクトル $\mathbf{w}_{r,v}^{(i)}$ も多様性を持つことになる．バッチ学習として識別関数を出力する方法はこれら重みベクトルの実現値の平均値を用いると，BayesPoint の近似値となり，理論上性能の向上が期待される．すなわち，

$$\tilde{\mathbf{w}}_v^{(i)} \mu \sum \mathbf{w}_{r,v}^{(i)} \tag{7}$$

とする．ここで，重みベクトル $\tilde{\mathbf{w}}_v^{(i)}$ は定数倍によって予測が変わらないことに注意されたい．この方法で学習する場合，実際の予測は逐次的に $\tilde{\mathbf{w}}_v^{(i)}$ を計算し予測はこの重みを使った識別関数によって行う．本手法も OBPM によって PA および SPA アルゴリズムの並列化を行う．このことによってストリームデータに対し効率的な識別関数を学習できるのと同時に，半教師あり学習においてラベル付けの一貫性を保つために利用する複数の並列の識別関数を得ることができる．このアルゴリズムをまとめると表 3 のようなアルゴリズムとなる．

### 3.2 提案手法

提案手法は OBPM による PA アルゴリズム及び SPA アルゴリズムの並列化に加え，正解なしデータを用いて，並列に存在するの重みベクトルが相互に影響を及ぼし，各識別関数の $\mathbf{x}$ に対するスコアが近づくように識別関数の更新を行う．12) では重みづけ平均の議論がされていたが，実際には単純平均を採用していた．本論文では，次の様な重みづけ平均をとる．

$$\tilde{\mathbf{w}}_v^{(i)} \mu \sum 2^{-\#miss(r)} \mathbf{w}_{r,v}^{(i)} \tag{8}$$

ここで，$\#miss(r)$ は $\mathbf{w}_r$ によって誤識別されたデータ数である．これらの並列化された重みベクトルを，正解データに関しては各識別関数が PA および SPA アルゴリズムの OBPM による更新を行う．そして，ラベルなしデータ $\mathbf{x}_i$ を受け取った場合，次のように更新を行う．

$$\mathbf{w}_{r,v}^{(i+1)} = \mathbf{w}_{r,v}^{(i)} + \alpha_{r,v}^{(i)} \mathbf{x}_i \tag{9}$$

ここで，

$$\alpha_{r,v}^{(i)} = C(\bar{m}_v^{(i)} - m_{r,v}^{(i)}) \tag{10}$$

である．また，

$$m_{r,v}^{(i)} = \frac{\mathbf{w}_{r,v}^{(i)} \cdot \mathbf{x}_i}{\|\mathbf{x}_i\|^2}, \bar{m}_v^{(i)} = \frac{\sum m_{r,v}^{(i)}}{M} \tag{11}$$

とした．$C$ は 0 から 2 までの値をとる超パラメータである．特に $C = 1$ の時各識別関数の $\mathbf{x}$ に関するスコアは一致する．本論文ではこの半教師ありオンラインアルゴリズムを，PA アルゴリズム SPA アルゴリズムを用いた場合それぞれに対し ss-PA，ss-SPA と呼び区別する．ss-PA,ss-SPA のアルゴリズムをまとめたものを表 4 に示す．

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

## 4. アルゴリズムの解析

### 4.1 有限データに関する誤識別回数の評価

本節では上で述べたアルゴリズムは半教師あり学習において，正解のないデータを用いても識別関数のパラメータが必ず収束することを証明することを目標とするが，それまでにいくつかの定理の紹介及び導出を行う．最初に，教師データ集合を固定した学習の場合，PA アルゴリズムには以下の定理が示されている．

**定理 1.** 有限個の要素からなる教師データ集合 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ に対し，正しい識別を行う重みベクトル $\{\mathbf{u}_v\}_{v\in Y}$ があるとする．すなわち，任意の $i$ で $\min_{v\neq y_i}(\mathbf{u}_{y_i}\cdot\mathbf{x}_i - \mathbf{u}_v\cdot\mathbf{x}_i) \geq 1$ とする．また，$\|\mathbf{x}_i\| \leq R$ であるとする．この教師データを用いて *PA* アルゴリズムを反復させて得た全ての段階での $\{\mathbf{u}_v^{(i)}\}_{v\in Y}$ に対し，総誤識別回数を $\#miss = \sum_{i=1}^N [\![\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x} \neq \max_{v\in Y}\mathbf{w}_v^{(i)}\cdot\mathbf{x}]\!]$ と定義する．すると *#miss* の上限が以下のように存在する．

$$\#miss \leq 2R^2 \sum_{v=1}^K \|\mathbf{u}_v\|^2 \tag{12}$$

ここで，[[P]] は命題 P が真の時に 1，偽の時に 0 をとる 0-1 値関数である．証明は文献 6) を参照されたい．一方，SPA アルゴリズムにおいても同様の証明が可能である．すなわち以下が成り立つ．

**定理 2.** 有限個の要素からなる教師データ集合 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ に対し，正しい識別を行う重みベクトル $\{\mathbf{u}_v\}_{v\in Y}$ があるとする．すなわち，任意の $i$ で $\min_{v\neq y_i}(\mathbf{u}_{y_i}\cdot\mathbf{x}_i - \mathbf{u}_v\cdot\mathbf{x}_i) \geq 1$ とする．また，$\|\mathbf{x}_i\| \leq R$ であるとする．この教師データを用いて *SPA* アルゴリズムを反復させて得た全ての段階での $\{\mathbf{w}_v^{(i)}\}_{v\in Y}$ に対し，総誤識別回数を $\#miss = \sum_{i=1}^N [\![\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x} \neq \max_{v\in Y}\mathbf{w}_v^{(i)}\cdot\mathbf{x}]\!]$ と定義する．すると *#miss* の上限が以下のように存在する．

$$\#miss < 2R^2 \sum_{v=1}^K \|\mathbf{u}_v\|^2 \tag{13}$$

となる．

証明の詳細は http://www.r.dl.itc.u-tokyo.ac.jp/~masin/paper/09mps.pdf に補遺があるのでこれを参照されたい．これらの定理は，PA 及び SPA アルゴリズムを用いて重みベクトルを学習したときに，すべての教師データ集合に対し正解クラスを割り当てられる重みベクトルにたどり着けることを意味する．したがって，アルゴリズムは必ず停止することを示している．また，定理の条件が，教師データ集合の順序や重複によらないことから，どのような順番で識別器がデータを受け取っても，また同じものを何回受け取っても定理の不等式を満たすことがわかる．したがって，OBPM による確率化を行った際に，各識別関数において $Z_{i,r} = 1$ となる系列に対しても誤識別の上限が存在する．ここで $Z_{i,r}$ は $i$ 番目のデータの $r$ 番目の重みベクトルに対するベルヌーイ確率変数の実現値である．すなわち，次の定理が成り立つ．

**定理 3.** 教師データ $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ と理想の識別を行う重みベクトル $\mathbf{u}$ があり，任意の $i$ で $\min_{v\neq y_i}(\mathbf{u}_{y_i}\cdot\mathbf{x}_i - \mathbf{u}_v\cdot\mathbf{x}_i) \geq 1$ かつ $\|\mathbf{x}_i\| \leq R$ であるとする．$Z_{i,r}$ は $i$ 番目のデータの $r$ 番目の重みベクトルに対するベルヌーイ確率変数の実現値とし，*PA* および *SPA* アルゴリズムに対し *OBPM* を用いたアルゴリズムを適用して並列化したアルゴリズムは，$\#miss = \sum_{i=1}^N \sum_{r=1}^M [\![\mathbf{w}_{r,y_i}\cdot\mathbf{x} \neq \max_{v\in Y}\mathbf{w}_{r,v}\cdot\mathbf{x}]\!]\,[\![Z_{i,r}=1]\!]$ に対し

$$\#miss \leq 2MR^2 \sum_{v=1}^K \|\mathbf{u}_v\|^2 \tag{14}$$

が成り立つ．

証明は http://www.r.dl.itc.u-tokyo.ac.jp/~masin/paper/09mps.pdf を参照されたい．これはベルヌーイ試行の事象が十分多くなればほとんどいたるところで各識別器は有限回で停止することを意味する．言い換えれば，Z のベルヌーイ分布の定義される確率空間において，確率 1 で各識別器が停止する．この定理と本手法による正解なしデータに対する更新式の性質により，次の定理が成り立つ．

**定理 4.** 教師データ $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ に対し，ある $\mathbf{u}$ があり，$\min_{v\neq y_i}(\mathbf{u}_{y_i}\cdot\mathbf{x}_i - \mathbf{u}_v\cdot\mathbf{x}_i) \geq 1$ を満たすとする．また $\|\mathbf{x}\| \leq R$ とする．この時正解なしデータを用いて *ss-PA* アルゴリズム及び *ss-SPA* アルゴリズムを反復させた場合，ラベルありデータを誤識別する回数 *#miss* の上限は次式で抑えられる．

$$\#miss < 2MR^2 \sum_{v=1}^K \|\mathbf{u}_v\|^2 \tag{15}$$

となる．

証明は http://www.r.dl.itc.u-tokyo.ac.jp/~masin/paper/09mps.pdf を参照されたい．これは半教師あり学習によって学習がうまくいかなくなることがないことを保証する定理となっている．

### 4.2 ストリームデータに関する誤識別率の漸近的評価

前節ではデータ集合が固定された有限の集合の場合において提案手法の性能を保証する定理であった．しかしながら実際にオンライン学習が有用であるのはストリームデータに対する予測を行う時であるので，反復のない状態での何らかの性能の保障が望ましい．この節では，そのようなオンライン学習の設定においても PA および SPA が有効な方法であることを証明した．また，提案手法である ss-PA，ss-SPA も同様に半教師オンライン学習の問題

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

設定で有効な学習が可能であることを証明した．まず，次の定理を証明した．

**定理 5.** $\mathbf{x}_i$ が閉集合上の分布 $P(\mathbf{x})$ に *i.i.d.* で従う確率変数とする．ある $\mathbf{u}$ があり，$y$ は $\min_{v\neq y}(\mathbf{u}_y-\mathbf{u}_v)\cdot\mathbf{x}>0$ となるように定められた $Y$ 値確率変数とする．ここで $\mathbf{x}$ の確率密度 $p$ は連続微分可能で $p(\{\min_{v\neq y}(\mathbf{u}_y-\mathbf{u}_v)\cdot\mathbf{x}\leq 0\})=0$ とする．このストリームデータを用いて *PA* アルゴリズムを逐次的に学習させた場合，経験誤識別率 $\frac{\#miss}{N}$ は 0 に確率収束する．すなわち，任意の $\varepsilon>0$ に対し，

$$\lim_{N\to\infty} P\left(\frac{\#miss}{N}\geq\varepsilon\right)=0 \tag{16}$$

となる．

これも証明については http://www.r.dl.itc.u-tokyo.ac.jp/~masin/paper/09mps.pdf に掲載したので参照されたい．仮定の $p(\{\min_{v\neq y_i}(\mathbf{u}_{y_i}\cdot\mathbf{x}_i-\mathbf{u}_v\cdot\mathbf{x}_i)\leq 0\})=0$ は境界線上でデータが生成される確率は 0 であることを意味しているので，ストリームデータにおける線形分離可能の仮定であるといえる．また，SPA アルゴリズムについても，定理 2．より $\#miss\leq\frac{2R^2\sum_{v=1}^{K}\|\mathbf{u}_v\|^2}{\min_{i=1\cdots n}|\min_{v\neq y_i}(\mathbf{u}_{y_i}-\mathbf{u}_v)\cdot\mathbf{x}_i|^2}$ が同じように示すことができているので，同様の定理が成り立つことがわかる．また，半教師あり学習に対しても同様に上限が定理 4 によって示されているので，次がわかる．

**定理 6.** $\mathbf{x}_i$ が閉集合上の分布 $P(\mathbf{x})$ に *i.i.d.* で従う確率変数とする．ある $\mathbf{u}$ があり，$y$ は $\min_{v\neq y}(\mathbf{u}_y-\mathbf{u}_v)\cdot\mathbf{x}>0$ となるように定められた $Y$ 値確率変数とする．ここで $\mathbf{x}$ の確率密度 $p$ は連続微分可能で $p(\{\min_{v\neq y}(\mathbf{u}_y-\mathbf{u}_v)\cdot\mathbf{x}\leq 0\})=0$ とする．このストリームデータを用いて *SPA* アルゴリズム，*ss-PA* アルゴリズム，及び *ss-SPA* アルゴリズムにより識別関数を逐次的に学習させた場合，経験誤識別率 $\frac{\#miss}{N}$ は 0 に確率収束する．すなわち，任意の $\varepsilon>0$ に対し，

$$\lim_{N\to\infty} P\left(\frac{\#miss}{N}\geq\varepsilon\right)=0 \tag{17}$$

となる．

## 5. 実　　験

これらのアルゴリズムの性能を Machine Learning Group Datasets[⋆1] の 20 Newsgroups corpus を用いて評価した．20 Newsgroups corpus は約 20,000 の文書からなり，20 の異なるグループに分けられている．この文書集合の中に用意されている四つのセクションを用いた．それぞれ”sb-8-1”,”sb-8-2”,”sb-7-1”,”sb-7-2”と呼ばれているもので，それぞれ 8 つと 7 つのジャンルを示すグループに分けられている．素性は BoW で与えられている．すなわち，全ての単語に対応する次元があり，文書においてはそれらの単語の出現回数をその次元の値とする．したがってこれらは非常に高い次元をもつ疎なベクトルとなる．今回の実験では $M=30,\pi=0.8$ とした．

⋆1 http://mlg.ucd.ie/datasets

**表 5** PA による教師あり学習と半教師あり学習の識別精度 (%)
Table 5 The Testset Accuracy of PA and ss-PA(%)

| PA | 教師あり | 半教師あり |
|---|---|---|
| sb-7-1 | 77.66 | 79.77 (+2.11) |
| sb-7-2 | 77.26 | 77.89 (+0.63) |
| sb-8-1 | 77.78 | 78.60 (+0.82) |
| sb-8-2 | 74.78 | 76.05 (+1.27) |

**表 6** SPA による教師あり学習と半教師あり学習の識別精度 (%)
Table 6 The Testset Accuracy of SPA and ss-SPA(%)

| SPA | 教師あり | 半教師あり |
|---|---|---|
| sb-7-1 | 84.83 | 85.23 (+0.40) |
| sb-7-2 | 87.00 | 87.66 (+0.66) |
| sb-8-1 | 83.78 | 84.33 (+0.55) |
| sb-8-2 | 82.38 | 83.33 (+0.95) |

今回解析した PA 及び SPA アルゴリズムの半教師あり学習を用いて，オンライン学習を想定した実験を行った．データセットのデータには全て正解クラスが割り振られているが，大部分が正解クラスが割り振られていないと仮定して，正解なしデータを受け取った場合に提案した半教師あり学習を行う方法と，正解データのみを用いて学習する方法の識別精度を比較した．この時ラベルなしデータがどれくらいあるかを変えて，データ量に対する提案手法の性質を検証した．評価は 10 等分における交差検証 (クロスバリデーション) を行い，そのうちの一つを無作為に選び超パラメータ（$C$）の調節を行った．表に示す値はそれらの識別精度の平均値である．ラベルなしデータはラベルありデータ比で 4 の場合を 10 回交差検証した．超パラメータは実験に用いないデータ集合を用いて調節した．

表 5,6 に，PA および SPA の識別精度 (%) を示す．識別精度は，高いほうが良い精度となる．PA,SPA ともに，半教師あり学習の性能は向上している．PA では，最大 2% 以上の増加が見られる．SPA は，教師データのみを使った場合ですでに，PA に比べ 10 %程度識別性能が高いが，教師なしデータを使うことで，さらに最大 1 %弱の性能が向上していることが分かる．

Vol.2009-MPS-76 No.21
Vol.2009-BIO-19 No.21
2009/12/17

## 6. ま と め

本論文ではオンライン学習における半教師あり学習の方法として，オンライン学習の分野において詳しく解析される誤識別回数の上限に着目した２種類のアルゴリズムを提案した．これらのアルゴリズムは依然として教師データに対する誤識別回数に対して上限を持つことを証明した．誤識別回数が上限をもつということは，識別関数が一定回数以上誤識別を行わないことを保証するので，非常に有用な概念である．提案手法がこの保証を保ったまま正解なしデータによる更新が可能であることは，非常に望ましい点であるといえる．また，有限の教師データ集合に対する誤識別回数の上限が示されることと，ストリームデータに対し識別関数がデータを分類する真の関数に近づくことが密接に関係していることが本論文で示され，データが際限なく与えられるような実用的な場面においても有限の教師データに関する解析が有用であることがわかった．さらに，実際のデータセットを用いた実験結果でも正解なしデータの利用による識別精度の向上がみられ，本手法がオンライン半教師あり学習の手法として有用であることがわかった．

一般に半教師あり学習を有効に行う為には，正解なしデータが正解付きの教師データに関する何らかの情報を共有している必要がある．したがってこれらのデータに何らかの仮定を置かなければならないと考えられるが，どのような仮定の下で学習ができるかは未だに議論の余地がある問題となっている．したがってより有効な半教師あり学習を行うためには，しかるべき仮定を置いたのちにそれを利用した半教師あり学習を構築することが今後の課題となるが，この時にも誤識別回数の上限が抑えられる方法が望ましい．

## 参 考 文 献

1) K.Crammer and Y.Singer. A new family of online algorithms for category ranking. *Jornal of Machine Learning Research*, 3:1025–1058, 2003.
2) K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Jornal of Machine Learning Research*, 3:951–991, 2003.
3) N.Cristianini and J.Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
4) K.Crammer and Y.Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Jornal of Machine Learning Research*, 2:265–292, 2002.
5) F.Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.
6) Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.
7) deSaV.R. Learning classificatio with unlabeled data. In *Proc. of ICML*, 2008.
8) U.Brefeld, T.Gaertner T.Scheffer, and S.Wrobel. Efficien co-regularized least squares regression. In *Proc. of ICML*, 2006.
9) V.Sindhwani, P.Niyogi, and M.Belkin. A co-regularized approach to semi-supervised learning with multiple views. In *Proc. of the 22th ICML workshop on Learning with Multiple views*, 2008.
10) Shin Matsushima, Nobuyuki Shimizu, Kazuhiro Yoshida, Takashi Ninomiya, and Hiroshi Nakagawa. Multi-class passive-aggressive algorithm with support classes. In *IPSJ SIGNL Technical Report,192-12*, 2009.
11) Offer Dekel and Yoram Singer. Data-driven online to batch conversions. In *Advances in Neural Information Processing Systems 18*, 2005.
12) Edward Harrington, Ralf Herbrich, John C.Platt JyrkiKivinen, and RobertC. Williamson. Online bayes point machines. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2005.
13) Ralf Herbrich, Thore Graepel, and Colin Campbell. Bayes point machines. *Jornal of Machine Learning Research*, 2001.