多クラス識別問題におけるオンライン学習のための厳密なPAアルゴリズム

松島慎† 清水伸幸† 吉田和弘 二宮崇† 中川裕志†
†東京大学

1 概要

多くの機械学習ではデータ全体に対して1つの最適な関数を求めるため，大量のデータに対して学習を行う際に計算コストの面から困難が生じる場合があるが，オンライン学習法は，データを1つずつ受け取りながら逐次的に識別関数を学習していく手法であり，低い計算コストで学習可能である．Crammerらによって提案されたPassive-Aggressiveアルゴリズム（PAアルゴリズム）は代表的なオンライン学習アルゴリズムであるが[1]，多クラス識別問題においては，厳密なPAアルゴリズムの枠組みから外れた近似解法が用いられてきた．本研究は本来のPAアルゴリズムの厳密解をサポートクラスという概念を用いて導出し，これらに基づく識別関数の更新を効率的に行う為のアルゴリズムを提案する．この手法に対する文書分類や手書き文字認識の実験を行い，既存のPAアルゴリズムより精度が高く，SVM(Support Vector Machines)に比べ同程度の精度をより高速に実現することを確認した．

2 問題設定

多クラス線形識別関数 $h_{\mathbf{w}^{(i)}}$ とは，$\mathbf{x}_i \in \mathbb{R}^D$ に対し $p_i \in Y=\{1,2,\cdots,K\}$ を返す関数であり，K 本の重みベクトル $\{\mathbf{w}_u^{(i)}\}_{u\in Y}$ $(\mathbf{w}_u^{(i)} \in \mathbb{R}^D)$ を用いて，以下のように表されるとする．

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

ここで $\mathbf{w}^{(i)}=\{\mathbf{w}_u^{(i)}\}_{u\in Y}$ とした．一般的なオンライン学習の枠組みにおいてはこの関数の学習を次の工程を繰り返すことによって行う．

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．

Exact PA Algorithms for online learning on multi-class classification problem

Shin Matsushima†, Nobuyuki Shimizu †,Kazuhiro Yoshida, Takashi Ninomiya†and Hiroshi Nakagawa†
†The University of Tokyo
{masin, shimizu, ninomi}@r.dl.itc.u-tokyo.ac.jp,
kyoshiddha@gmail.com, n3@dl.itc.u-tokyo.ac.jp

3 提案手法

PAアルゴリズムの枠組みにおけるオンライン学習の設定では，$\mathbf{x}_i$ に対応する正解ラベル y_i を受け取り，損失 $\ell(\mathbf{w};\mathbf{x}_i,y_i)\geq 0$ を計算する．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．

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

表1: 導出されたSPAアルゴリズム

$\mathbf{w}_v^{(1)}=\mathbf{0}\ \ (\forall v\in Y)$
foreach $i=1,2,\cdots N$ **do**
　$\ell_v := \left[1-\left(\mathbf{w}_{y_i}^{(i)}\cdot\mathbf{x}_i - \mathbf{w}_v^{(i)}\cdot\mathbf{x}_i\right)\right]_+ \quad (v\neq y_i)$;
　Compute j-th class $\sigma(j)\in Y\backslash\{y_i\}$ in descending order of ℓ_v for $^\forall j\leq K-1$.
　$S:=\varnothing$
　while $\sum_{j=1}^{|S|}\frac{\ell_{\sigma(j)}}{|S|+1} < \ell_{\sigma(|S|)}$ **do**
　　$S:=S\cup\sigma(|S|)$
　end while
　$\tau_v := \ell_v - \sum_{u\in S}\frac{\ell_u}{|S|+1} \quad (v\in S)$
　$\tau_v := 0 \quad (v\notin S)$
　$\mathbf{w}_v^{(i+1)} := \mathbf{w}_v^{(i)} + \sum_{u\neq y_i}\tau_u\mathbf{x}_i \quad (v=y_i)$
　$\mathbf{w}_v^{(i+1)} := \mathbf{w}_v^{(i)} - \tau_v\mathbf{x}_i \quad (v\neq y_i)$
end foreach

この最適化問題の解は，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．Crammerらは，K クラス識別問題におけるPAアルゴリズム[1]で用いる損失関数を以下のように定義している．

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

表 3: 各データセットの最低誤識別率 (%, 括弧内は反復回数)

	PA	PA-I	PA-II	SPA	SPA-I	SPA-II	Perc	LR	M-SVM
USPS	6.82(34)	6.83(34)	6.82(40)	6.02(34)	5.73(39)	**4.95**(35)	6.45(30)	4.98	**4.83**
Reuters	3.80(14)	3.80(14)	3.81(15)	3.01(14)	3.03(39)	**2.96**(32)	3.99(25)	3.50	**3.18**
News20	22.98(4)	21.49(7)	21.02(8)	20.01(1)	16.84(5)	**15.83**(11)	22.63(26)	17.79	**15.94**

表 2: PA アルゴリズムと SPA アルゴリズム

	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	$\sigma(k):\ \sum_{j=1}^{k-1}\ell_{\sigma(j)}<k\ell_{\sigma(k)}$

表 4: 各学習器の反復なしの学習における誤識別率 (%)

	Perc	PA	SPA	PA-I	SPA-I	PA-II	SPA-II
USPS	11.74	14.02	9.15	11.05	8.68	11.74	**6.89**
reut20	7.88	7.69	5.32	7.44	5.46	7.95	**5.29**
News20	31.96	27.03	20.50	24.62	18.85	25.82	**17.23**

表 5: 各データセットにおける所要計算時間 (sec)

	PA	PA-I	PA-II	SPA	SPA-I	SPA-II	Perc	LR	M-SVM
USPS	5.5	5.5	5.5	5.9	5.9	5.9	1.5	53.6	68.5
Reuter	6.2	6.2	6.2	6.4	6.4	6.4	1.0	92.5	34.3
News20	15.7	15.9	15.8	16.8	17.1	17.2	7.9	405.6	75.7

$[\bullet]_+$ 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なる．Crammer らは計算の複雑性などを考慮してこれらの制約条件を緩和した問題の最適解を用いて識別関数を更新していた．本研究においては近似解を用いず厳密な最適解を解析的に導出し、その最適解を用いて識別関数を更新する。導出された解析解は以下のようになる。

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

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 の要素及び τ_v の値は簡単なアルゴリズムで計算できる．これを本稿では SPA アルゴリズムと呼び概要を表 1 に示す．また，Crammer らによる手法との違いを表 2 に示す．PA-I,PA-II アルゴリズム [1] に関しても同様に厳密解による更新アルゴリズムを導出した．これを本稿では SPA-I,SPA-II アルゴリズムと呼ぶ．

4 実験

20Newsgroups[1], Reuters[2], USPS[3]のデータセットを用いて実験を行った．これらのデータセットはそれぞれ脚注のサイトで入手可能なものである．20Newsgroups,Reuters は文書分類，USPS は文字認識のタスクになっている．PA,PA-I,PA-II,SPA,SPA-I,SPA-II 及びパーセプトロンに対しこれらを 40 回まで反復し，各時点でのテストセットに対する識別精度の比較実験を行った．そしてこれらのオンライン学習法のアルゴリズムに加えて，2 つのバッチ学習法，多クラスロジスティック回帰（表中"LR"）及び多クラス SVM[2, 3, 4]（表中"SVM"）[4]における精度も検証し比較した．結果を表 3 に示す．実験では 10 等分における交差検証を行い，超パラメータの調節はそのうちの一つを無作為に選んで行った．またこの時の所要時間を表 5 に示した．またオンライン学習器について反復なし，パラメータ調節なしの場合の精度も比較した（表 4）．

提案手法はどのデータセットにおいてもそれぞれの既存の手法に比べ良い性能を示している．USPS ではバッチ学習の方が良い精度を得ているが，20Newsgroups および Reuters のデータセットでは SPA-I,SPA-II アルゴリズムがバッチ学習器よりよい精度を達成した．また，SPA アルゴリズムは各バッチ手法に比べ計算時間は格段に短く，PA アルゴリズムにおける計算時間とはほとんど差がないことがわかった．

5 結論

本稿では，多クラス PA アルゴリズムの効率的な厳密解法を提案した．結果として得られた更新を用いたアルゴリズムは優良であり，今後計算コスト的にも精度面においても優良な学習のためにオンライン学習が用いられうることが示唆できたと考えられる．

参考文献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *JMLR*, 7:551–585, 2006.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[3] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.

[4] Crammer Koby and Singer Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.

[1] http://mlg.ucd.ie/datasets

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[4] 実装は http://svmlight.joachims.org/svm_multiclass.html によった．