F-023

Passive-Aggressive アルゴリズムにおける更新パラメータの改良
Improvement of Update Parameter in Passive-Aggressive Algorithm

吉田 幸平 † 大村 英史 ‡ 太原 育夫 ‡
Kohei Yoshida Hidefumi Ohmura Ikuo Tahara

1. はじめに

機械学習を利用した自然言語処理の一分野に文書分類があるが，自然言語処理の分野では言葉の曖昧性や表現の多様性によって処理するデータ量は膨大になってしまう欠点がある．ここではブログ記事の分類における次元の削減と分類精度の維持を目的としてオンライン学習アルゴリズムの一つである Passive-Aggressive アルゴリズム [1] の更新パラメータの改良を検討する．

2. ブログ記事の分類

2 つのブログ記事から特徴語を抽出し，それらを用いて分類し，精度を検証する．手順は以下のとおりである．

1. ブログ記事を形態素解析し，品詞ごとに分解する．
2. 分解した単語群を tf-idf に基いて出現頻度からフィルタリングして特徴語辞書を作成する．
3. 文書をベクトルとして表すモデルである Bag of Words により，特徴語辞書を用いて特徴語ベクトルを作成する．
4. Passive-Aggressive によりクラスを分類する．
5. 分類したものから精度を測定する．

3. Passive-Aggressive アルゴリズム

あるラウンド t において，受け取る事例ベクトルを $\mathbf{x}_t$，正解ラベルを y_t，事例を分類するための重みベクトルを n 次元の実数ベクトル $\mathbf{w}$ と定める．また分類が正しかった際の確信度を $y_t(\mathbf{w}_t \cdot \mathbf{x}_t)$ で表し，確信度がある程度大きい場合 (この場合は 1 以上) は更新を行わず，確信度が 1 より小さい場合に，以下の損失関数により損失を受けることとする．

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

これを簡略化し，$l_t(= l(\mathbf{w}_t;(\mathbf{x}_t,y_t)))$ と表記する．

重みベクトルの更新を以下の式の最適化問題の解として定義する．

$$\begin{aligned} \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in R^n} \frac{1}{2} \parallel \mathbf{w} - \mathbf{w}_t \parallel^2 \\ \text{s.t.} \quad l(\mathbf{w};(\mathbf{x}_t,y_t)) = 0 \end{aligned} \tag{1}$$

この式では $l_t = 0$ ならば更新は行わず $\mathbf{w}_{t+1} = \mathbf{w}_t$ となり (passive)，そうでない場合は条件の範囲内で更新を行い $\mathbf{w}_{t+1}$ となる (aggressive)．

この最適化問題は以下のように，ラグランジュの未定乗数法を用いて解くことが出来る．この問題におけるラグランジュ関数は，ラグランジュ乗数を $\tau \geq 0$ とすると，

$$L(\mathbf{w},\tau) = \frac{1}{2} \parallel \mathbf{w} - \mathbf{w}_t \parallel^2 + \tau(1 - y_t(\mathbf{w}\cdot\mathbf{x}_t))$$

である．この関数の偏導関数が 0 になる点を求めればよいので，

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t \quad \text{where} \quad \tau_t = \frac{l_t}{\parallel \mathbf{x}_t \parallel^2} \tag{PA}$$

とすることでパラメータの更新を行うことができる．

ところで，ノイズの影響を避け，パラメータの更新を穏やかにするために式 (1) にスラック変数 ξ を導入すると，

$$\begin{aligned} \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in R^n} \frac{1}{2} \parallel \mathbf{w} - \mathbf{w}_t \parallel^2 + C\xi \\ \text{s.t.} \quad l(\mathbf{w};(\mathbf{x}_t,y_t)) \leq \xi \ \text{ and } \ \xi \geq 0 \end{aligned} \tag{2}$$

となる．ここで，C はスラック変数がパラメータ更新に与える影響の大きさを調整するための正のパラメータである．このとき，τ_t として以下の値が得られる．

$$\tau_t = \min\left\{C, \frac{l_t}{\parallel \mathbf{x}_t \parallel^2}\right\} \tag{PA − I}$$

この式より，PA-I では，損失によるパラメータの更新の値の上限を C により抑えていることになる．

また，ξ^2 を考えて，

$$\begin{aligned} \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in R^n} \frac{1}{2} \parallel \mathbf{w} - \mathbf{w}_t \parallel^2 + C\xi^2 \\ \text{s.t.} \quad l(\mathbf{w};(\mathbf{x}_t,y_t)) \leq \xi \end{aligned} \tag{3}$$

とすれば，

$$\tau_t = \frac{l_t}{\parallel \mathbf{x}_t \parallel^2 + \frac{1}{2C}} \tag{PA − II}$$

が得られる．

PA-I，PA-II は更新を抑えるのみで，ソフトマージンという考え方は変わらない．また精度を上げるには C の値を細かく設定する必要がある．

データを学習する際はクラス毎に連続してデータが与えられるので，前回と似た値を更新することで数値の均衡化を行うことができ，バランス良く更新できると考えられる．そこで，新たな τ_t として以下の式を提案する．

$$\tau_t = \frac{n \cdot l_t + l_{t-1}}{n \cdot \parallel \mathbf{x}_t \parallel^2 + \parallel \mathbf{x}_{t-1} \parallel^2}$$
$$(t = 0 \text{ のとき } l_t = 0, \mathbf{x}_t = 0;\ n \text{ は実数})$$

この式は，前回 ($t-1$ 回目) の損失である l_{t-1} と前回の事例の二乗ノルム $\mathbf{x}_{t-1}$ を継承している．そして，変数 n の値に応じて t と $t-1$ のバランスを調整している．

†東京理科大学大学院理工学研究科情報科学専攻
‡東京理科大学理工学部情報科学科

4. 分類実験

Passive-Aggressive によるクラスの分類は PA,PA-I, 提案手法で行い，特徴語辞書作成に用いる記事数を各 100 記事から各 25 記事に減らすことで次元を削減し，精度を維持することが出来るかを検証する．またこのとき，学習させるデータは 50 記事，精度の検証に用いるテスト用データは 400 記事とする．

辞書の作成やテストに行ったブログ記事はクリエイティブ・コモンズライセンスのものを livedoor ニュースコーパスから引用した [2]．以下では 2 つのブログを，ブログ A，ブログ B とする．

まず，各 100 記事の場合である．この時の特徴語ベクトルの次元数は 671 次元となった．PA と PA-I の場合について検証を行った結果を表 1 に示す．精度に大きな違いは見られなかった．

表 1: 各 100 記事による実験 - 正解率 (PA-I)

変数 C	正解率 (ブログ A)	正解率 (ブログ B)
0.020	98.5 %	85.5 %
0.030	97.5 %	90.5 %
0.040	95.5 %	93.5 %
PA	94.5 %	93.5 %

次に，各 50 記事の場合について，PA,PA-I, 提案手法での結果を表 2 に示す．このときの特徴語ベクトルは 502 次元となった．また，このときの PA-I について，更新の上限を抑える変数 C の値は 0.021 なのだが，τ の値が 0.0204 であったため，これ以上の精度の調整のためには C を小数点第 5 位以下でチューニングする必要がある．

提案手法は，変数 n の小数点第 1 位でのチューニングでバランスの良い正解率を保っていることが確認できる．

表 2: 各 50 記事による実験 - 正解率 (提案手法)

変数 n	正解率 (ブログ A)	正解率 (ブログ B)
1.5	89.5 %	89.0 %
1.7	85.5 %	90.0 %
1.9	86.0 %	86.5 %
2.1	87.0 %	85.5 %
2.3	87.5 %	84.0 %
PA	100.0 %	47.5 %
PA-I (C=0.021)	100.0%	47.5%

最後に，各 25 記事の場合について結果を表 3 に示す．このときの特徴語ベクトルは 275 次元となり，初めの各 100 記事の約 40% である．PA や PA-I では精度が大きく下がっているが，提案手法では各記事について精度 80%以上を保っていることが確認できる．

以上より，PA-I を用いてチューニング中に小数第 3 位や 4 位で頭打ちになった場合，それ以上小さな桁数まで調整する必要があるが，提案手法ではその必要が無く，精度を向上させることができた．

表 3: 各 25 記事による実験 - 正解率 (提案手法)

変数 n	正解率 (ブログ A)	正解率 (ブログ B)
1.1	83.5 %	87.5 %
1.3	71.5 %	92.0 %
1.5	81.0 %	88.0 %
1.7	79.5 %	87.5 %
PA	83.0 %	73.5 %
PA-I (C=0.0355)	83.5%	73.5%

また，PA-I はソフトマージンであり更新を抑えるのみであるが，場合によっては値を大きく更新する必要もあるので，提案手法により精度が向上したものと考えられる．

さらに，PA-I の min 関数によるソフトマージンが初回更新時や前半のみにしか適用されていない場合が多かったことから，提案手法の，$t = 0$ のとき $l_t = 0, \mathbf{x}_t = 0$ の条件により，初回の更新を和らげていることが PA-I と同様の効果を出していると考えられる．

5. おわりに

Passive-Aggressive アルゴリズムにおいて，チューニング可能なパラメータ更新式を提案し，検討を行った．実験の結果，次元を削減した場合での精度の向上とチューニングの簡略化を示すことができた．

課題としては，PA-I の変数 C の値を細かくチューニングした場合と精度を比較し検証を行いたい．

参考文献

[1] K.Crammer，O.Dekel，J.Keshet，S.Shwartz，and Y.Singer，“Online Passive-Aggressive Algorithms，”Journal of Machine Learning Research，vol.7，pp.551-585，2006.

[2] livedoor ニュースコーパス `http://www.rondhuit.com/download.html#ldcc` 2015 年 2 月 4 日に閲覧.