大規模文書分類のための SGD SVM の高速化

Acceleration of SGD SVMs for Large-Scale Text Classification

小野 正貴[*1]　David McAllester[*2]　佐々木裕[*1]

[*1]豊田工業大学　[*2]Toyota Technological Institute at Chicago

1. はじめに

近年，インターネットの普及や医療カルテの電子化などによって，多くのテキストがコンピュータ上で利用可能となっている．それらを利用したタスクとして，文書の自動分類があり，機械学習を用いてその実現を行う．SVM(Support Vector Machine)は，テキスト分類に対し有効な機械学習手法であり，精度・学習速度共に良好である[1]．しかしながら，臨床試験計画書の MeSH カテゴリ分類[2]や，PASCAL LSHTC3 CHALLENGE [3]などの階層的多クラス分類の場合には，膨大なカテゴリ数に対応するモデルを学習する必要があり，学習には多くの時間を要する．

本研究では，SVM の分類モデル学習法のうち，高速に学習ができる確率的勾配降下法（SGD: Stochastic Gradient Decent）について分析し，改良手法についての評価を行った．

2. 先行研究

2.1. SVM

SVM は機械学習手法の一種であり，線形 2 値分類器を学習する．学習データとして以下のものを考える．

$$x_i \in \Re^D, y_i \in \{\pm 1\} (i = 1 \ldots n)$$

このデータから以下の目的関数を最小化し，分類器 $w\left(\in \Re^D\right)$を学習する．

$$E(w) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \max\left(0, 1 - y_i w^T x_i\right) + \frac{\lambda}{2} \|w\|^2$$

$$\lambda : \mathrm{trade-off\ parameter}$$

2.2. 確率的勾配降下法[4]

確率的勾配降下法とは，期待値の形で表される目的関数の最適化問題に適応できる解法アルゴリズムである．直接目的関数の勾配を計算・更新するのではなく，データ集合からランダムに 1 サンプルに対する勾配を計算・更新を行う．本解法を SVM 学習に適用したものを以下に示す．

$$\Delta E = y_i x_i \left[1 > y_i w^T x_i\right] + \lambda w$$

$$\begin{aligned} w_{t+1} &= w_t - \eta \Delta E \\ &= w_t - \eta \left(y_i x_i \left[1 > y_i w^T x_i\right] + \lambda w \right) \end{aligned}$$

$$w : \mathrm{classifier\ vector}$$

$$t : \mathrm{iteration\ number}$$

$$\eta = \frac{1}{\lambda (t + t_0)} : \mathrm{learning\ rate}$$

t は試行が進むにつれ増加するため，η は減衰していく．

3. 高速化のための提案

3.1. 既存手法の問題点

学習データがテキストデータ等から生成されるとすると，ベクトルの各要素の大半がゼロ，すなわちスパースである場合が多い．また，学習データ内における，各要素の出現頻度もまた大きく異なる場合がある．しかしここで更新式に着目すると，η は全ての要素に対して一様に作用する．この更新量の不均衡性は最適値を得る妨げになる可能性があると考えられる．

3.2. 提案手法

そこで我々は，要素毎に学習率を設定し，個別に減衰させることを提案する．

$$w_{t+1} = w_t - \begin{pmatrix} \eta_{t1} \\ \vdots \\ \eta_{tD} \end{pmatrix}^T \left(y_i x_i \left[1 > y_i w_t^T x_i \right] - \lambda w_t \right)$$

$$\eta_j = \frac{\left[x_{ij} != 0 \right]}{\lambda \left(t_j + t_0 \right)}$$

4. 評価実験

4.1. 実験方法

従来手法及び提案手法の性能を，学習時間とテストデータに対する精度について比較する．利用したデータセットはそれぞれ，①臨床試験計画書の MeSH カテゴリ分類，②LSHTC3 を元にしたものである．どちらも階層的多クラス分類であるが，今回はある 1 クラスにサンプルが属すがどうかの 2 値分類問題としてデータセットを生成した．利用したデータの素性を表 1 に示した．また，精度の評価の指標としては F 値を用いる．

表 1：データセット概要

	データ①	データ②
次元数	1,363,944	2,085,164
非ゼロ要素の平均数	261	47
正例の占める割合	54%	0.25%
学習データサンプル数	68,684	411,197
テストデータ サンプル数	7,632	45,689

4.2. 結果及び考察

実験結果を図 1 に示した．データ①に対しては，従来手法のほうが早い時間でよりよい分類モデルを得られる事がわかる．また，F 値の最大値も向上している．一方データ②に対しては，二つの手法に優位な差は見られなかった．

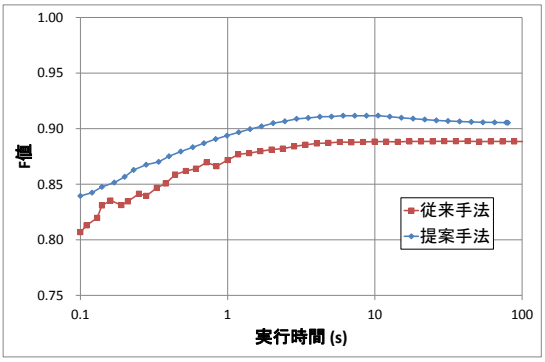
これらのことから，一部のデータに対しては提案手法が従来手法より良好な結果が得られることが分かったものの，データによっては提案手法の効果はなく，提案手法を適用した分の計算量増大によって学習時間が伸びる場合がある．

5. 終わりに

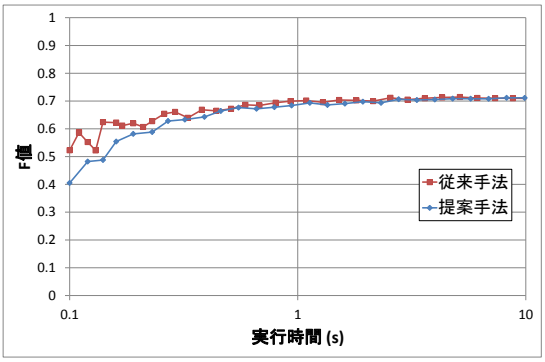
本論文では，確率的勾配降下法を用いた SVM 学習法の高速化に関する提案とその結果について述べた．学習率を要素毎に減少させることで，あるデータセットに対して良好な結果が得られたものの，もう一つのデータセットに対しては提案手法の有用性は確認できなかった．

今後の方針として，さらに多くの特性の異なるデータセットを用いて性能の評価を行い，提案手法の優位性と，その優位性に影響を与えるデータセットの指標の評価があげられる．また，異なる手法での学習と

なるため，収束性に関する解析的な議論も必要であると考えられる．



(a) データセット①



(b) データセット②

図１：各データに対する学習時間とテストの F 値の比較

これらのことから，一部のデータに対しては提案手法が従来手法より良好な結果が得られることが分かったものの，データによっては提案手法の効果はなく，提案手法を適用した分の計算量増大によって学習時間が伸びる場合がある．

6. 終わりに

本論文では，確率的勾配降下法を用いた SVM 学習法の高速化に関する提案とその結果について述べた．学習率を要素毎に減少させることで，あるデータセットに対して良好な結果が得られたものの，もう一つのデータセットに対しては提案手法の有用性は確認できなかった．

今後の方針として，さらに多くの特性の異なるデータセットを用いて性能の評価を行い，提案手法の優位性と，その優位性に影響を与えるデータセットの指標の評価があげられる．また，異なる手法での学習となるため，収束性に関する解析的な議論も必要であると考えられる．

参考文献

[1] Y. Yang, X. Liu: A re-examination of text categorization methods, SIGIR, 1999.

[2] 佐々木裕: 臨床試験計画書の MeSH カテゴリーへの自動分類，言語処理学会年次大会発表論文集, 2010.

[3] LSHTC PASCAL CHALLENGE (http://lshtc.iit.demokritos.gr/forum/17)

[4] Léon Bottou: “Large-Scale Machine Learning with Stochastic Gradient Descent”, Proceedings of the 19th International Conference on Computational Statistics, pp. 177–187, 2010.