3O1-5in

ロジスティック回帰におけるパラメータ平均化

Averaging Parameters in Logistic Regression

秋山健人 *1　　二宮崇 *2
Kento AKIYAMA　　Takashi NINOMIYA

*1愛媛大学工学部情報工学科
Department of Computer Science, Ehime University

*2愛媛大学大学院理工学研究科
Graduate School of Science and Engineering, Ehime University

This research proposes a method of averaging parameters in logistic regression (LR), which is a typical discriminative classifier often used in the area of machine learning. Learning of LR starts by initializing parameters to zero in many cases because the loss function is a convex and global minimum can be obtained. This research indicates that the performance of the LR can change by varying the initial values of the parameters, and also show that it can be improved by averaging a bundle of parameters that are obtained by learning with various initial values of the parameters.

1. はじめに

機械学習は，音声や文字などのパターン認識，メールクライアントのスパムフィルタ，医療診断など実社会において様々な場面に使われている．機械学習で用いられるものの一つであるロジスティック回帰は，“はい”，“いいえ”など出力が 2 値で表わされるデータを用い，SVM と並んでよく使われる代表的な識別モデルによる分類手法である．

ロジスティック回帰の学習は，素性の重みベクトル $\mathbf{w}$ と訓練データ D に対する目的関数を最小化するように $\mathbf{w}$ を更新することで行われる．ロジスティック回帰の目的関数は $\mathbf{w}$ に対し凸関数であるため，反復学習により大域的最適解が得られる．そのため，多くの実装では，$\mathbf{w}$ は 0 で初期化されており、学習を行うことで大域的最適解に収束して最適な $\mathbf{w}$ を得ている．しかし，最適解がただ一つではなく複数あること，つまり，目的関数の値が同じ最小値となる複数の解が存在することが考えられ，その中の一つだけしか得られていないことが考えられる．

本研究では，ロジスティック回帰において $\mathbf{w}$ の初期値を乱数で設定し，複数回学習したものを平均化することで，より精度が高くなることを目指す．

本論文では，2 節でロジスティック回帰についての説明をする．3 節で提案手法であるロジスティック回帰の高精度化を実現するための手法として，$\mathbf{w}$ の初期値の変更と，学習後の $\mathbf{w}$ の平均化について説明する．4 節で実際に計算機で行った実験結果を示す．

2. ロジスティック回帰

ロジスティック回帰は 2 クラス分類問題における一般化線形モデルの一種である．入力 $\mathbf{x} = (x_1, ..., x_n)$ と出力 $y \in \{-1, 1\}$ が与えられたとき，ロジスティック回帰の確率モデルは次式で与えられる．

$$p(y|\mathbf{x}) = \frac{1}{1 + e^{-y\mathbf{w}^T \cdot \mathbf{x}}} \tag{1}$$

ただし，$\mathbf{w}$ はロジスティック回帰のパラメータであり，重みベクトルと呼ばれる．重みベクトルは学習により得られる．ある未知の入力 $\mathbf{x}$ に対する出力 y の推定は次式により行われる．

$$\hat{y} = \arg\max_y p(y|\mathbf{x}) \tag{2}$$

訓練データ $D = (\mathbf{x}_i, y_i)_{i=1}^l$ が与えられたとき，重みベクトル $\mathbf{w}$ は次式により推定される．

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} g(\mathbf{w}, D) \tag{3}$$

ここで g は目的関数と呼ばれ，次式で与えられる [1]．

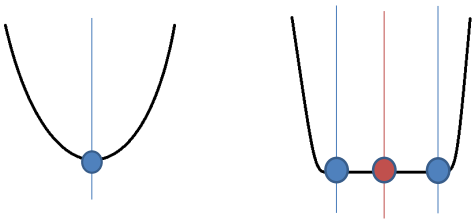
$$g(\mathbf{w}, D) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \log(1 + e^{-y_i \mathbf{w}^T x_i}) \tag{4}$$

ここで，x_i は訓練データの i 番目の素性ベクトル，y_i は x_i の正解ラベル，C はハイパーパラメータ，l はデータセットの大きさ，$\mathbf{w}$ は重みベクトルである．

式 4 の第 1 項は正則化項と呼ばれ，重みの平滑化を行い，過学習を防ぐための関数である．この正則化項は L2 正則化項と呼ばれ，重みの二乗和により与えられる．式 4 の第 2 項は損失関数と呼ばれ，これは確率モデルの負対数に対応する．損失関数は訓練データに対するパラメータ推定の悪さを定義した関数である．

式 4 に示すロジスティック回帰の損失関数および目的関数は凸関数である．凸関数に対する最適化は大域的最適解に収束するので，図 1 の (1) に示す 2 次元での例のように収束し一意に定まる．しかし，実際には図 1 の (1) のような形だけではなく，図 1 の (2) のような形の場合もあり，適した解がただ一つではなくその中の一つしか得られていないことがある．理想的には図 1 の (2) の赤丸のように中心に近い重みベクトルが得られることが望ましい．

連絡先: 秋山健人，愛媛大学工学部情報工学科
akiyama@ai.cs.ehime-u.ac.jp



(1) 理想　　(2) 予想

図 1: 2 次元での $\mathbf{w}$ 最適化イメージ

3. ロジスティック回帰における重みベクトルの平均化

本手法では，まず，実数の乱数で初期化された重みベクトルを生成する．次に，用意したデータセットを用いて学習し，複数の識別器を得る．最後に，識別器が学習した重みベクトルを平均化して新たな重みベクトルを得る．

3.1 重みベクトルの初期化設定

重みベクトル $\mathbf{w}$ の初期値を乱数で設定する．乱数は $(i)[-1,1]$ と $(ii)[-10,10]$ の 2 通りの実数範囲を用いる．

3.2 重みベクトルの平均化

パーセプトロンのパラメータの初期値を乱数で与えることにより、複数のパーセプトロンを学習し、これらの平均化を行うと性能が良くなることが知られている [2]．本研究ではこの考え方を元にし，重みベクトルを平均化することにより，重みベクトルの平滑化を行う．まず，初期値をランダムに変更しながら 100 回繰り返し学習を行う．学習が終わった後に，精度の高い順に k 個の重みベクトルを足しあわせて平均化する．l 回目の学習で生成された重みベクトルを $\mathbf{w}^{(l)}$ とすると，最終的に求める重みベクトル $\mathbf{w}$ は以下のようになる．

$$\mathbf{w} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{w}^{(sort(i))} \tag{5}$$

ただし，$sort(i)$ は i を引数として，i 番目に精度が良い重みベクトルのラベル (l) を返す関数である．

4. 実験

計算機を用いて，提案手法により重みベクトルを生成し，精度評価をした実験結果を以下に示す．提案手法の実装はオープンソースである $LIBLINEAR$[1] を改変して行った．また，従来法の評価には改変していない $LIBLINEAR$ を用い，L2 正則化項付きロジスティック回帰の性能を評価した．

4.1 データセット

$LIBSVMData$[*1] の $(1)a8a$ と $(2)news20binary$ を実験データとして用いる．(1) は 32,561 個のデータポイントから成り，訓練データとして 22,696 個，パラメータ調整用データとして 4,932 個，テストデータとして 4,933 個に分割して使用する．(2) は 19,996 個のデータポイントから成り，訓練データとして 10,000 個，パラメータ調整用データとして 4,998 個，テストデータとして 4,998 個に分割して使用する．訓練データは学習する際に使用するデータ，パラメータ調整用データはハイパーパラメータの推定と生成した重みベクトルの精度評価に使用するデータ，テストデータは未知データであり最終的な精度評価に使用するデータとして使用する．また，(1) の素性ベクトルは 123 次元から成り，(2) の素性ベクトルは 1,355,191 次元から成る．

*1 http://www.csie.ntu.edu.tw/ cjlin/libsvm/

4.2 ハイパーパラメータの推定

4.2.1 C パラメータの推定

式 4 における C の値は 2^z $(z \in \{-14,-13,...,14\})$ の中から探索し，パラメータを変更しながら 29 通りの重みベクトルを生成し，パラメータ調整用データでの精度を測定する．ここでは重みベクトル $\mathbf{w}$ の初期値は 0 とする．

$a8a$ データセットでの測定結果を図 2 に示す．図 2 からわかるように，$C = 0.5$ の時に 84.8135% となり，最も精度が高いので $a8a$ のデータセットでは $C = 0.5$ を用いることとする．また，この時のテストデータでの精度は 85.9518% である．このテストデータでの精度を $a8a$ の従来法の精度とし，提案手法との比較を行う．

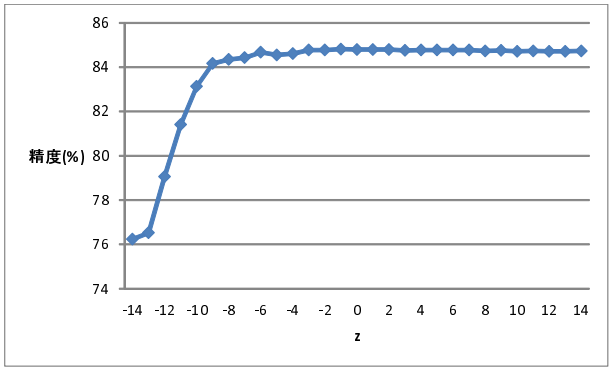


図 2: $a8a$ の C パラメータ推定

$news20binary$ データセットでの測定結果を図 3 に示す．図 3 からわかるように，$C = 16384$ の時に 96.3185% となり，最も精度が高いので $news20binary$ のデータセットでは $C = 16384$ を用いることとする．また，この時のテストデータでの精度は 95.7183% である．このテストデータでの精度を $news20binary$ の従来法の精度とし，提案手法との比較を行う．

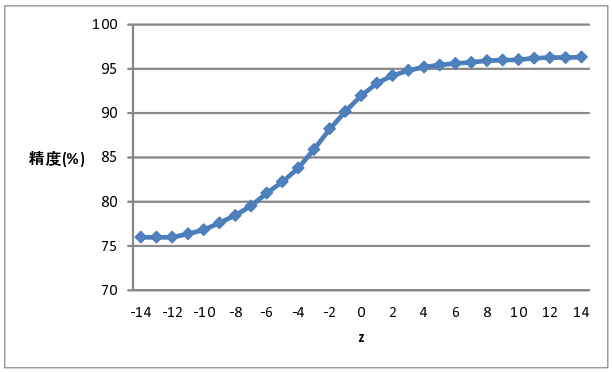


図 3: $news20binary$ の C パラメータ推定

4.2.2 k パラメータの推定

平均数は $10,20,30,40,50,60,70,80,90,100$ とする．3.2 節で示した式を用いてそれぞれの平均数毎の重みベクトルを生成

し，それらの重みベクトルを用いてパラメータ調整用データでの精度を測定したときに，最も精度が高い平均数を k とする．

$a8a$ データセットを用い，$[-1,1]$ と $[-10,10]$ の実数範囲で初期化し平均化した精度を図 4 に示す．図 4 より，$[-1,1]$ のときは 30，$[-10,10]$ のときは 40 を k とし，このときの精度を従来法と比較する．

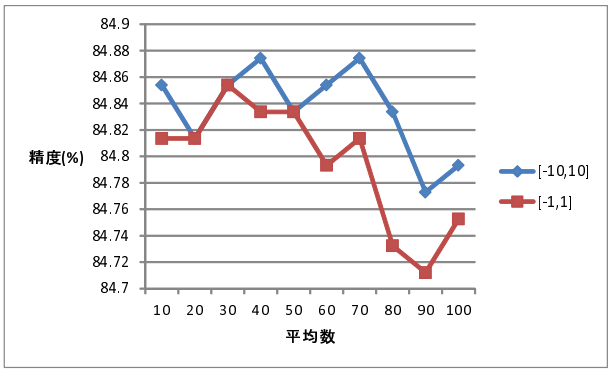


図 4: $a8a$ の k パラメータ推定

次に，$news20binary$ データセットを用い，[-1,1] と [-10,10] の実数範囲で初期化し平均化した精度を図 5 に示す．図 5 より，$[-1,1]$ のときは 30，$[-10,10]$ のときは 60 を k とし，このときの精度を従来法と比較する．

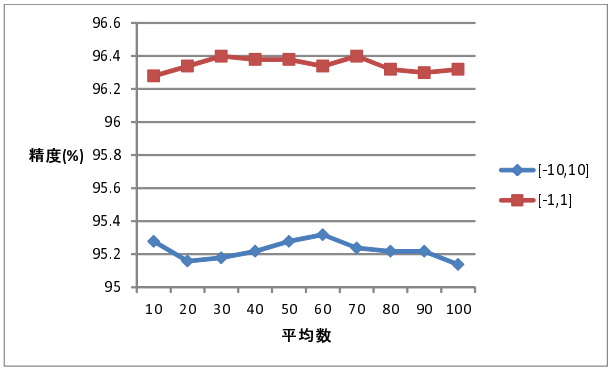


図 5: $news20binary$ の k パラメータ推定

4.3 実験結果

二つのデータセット $a8a$ と $news20binary$ に対し，重みベクトルを $(i)[-1,1]$ と $(ii)[-10,10]$ の二通りの実数範囲で初期化して実験を行った．従来法（重みベクトルの初期値が 0）と提案手法の比較を表 1 に示す．表に示す精度は，訓練データを用いて生成された重みベクトルのうち，パラメータ調整用データを分類した時の最大精度のものでテストデータを分類した結果である．従来法と比較すると，(i) の時，$a8a$ は平均化前は同じであるが平均化後では 0.081% 向上している．$news20binary$ では平均化前は 0.08% 下がっているが，平均化後は 0.1% 向上している．(ii) の時，$a8a$ は平均化前は 0.0608%，平均化後は 0.0405% 向上している．$news20binary$ では平均化前は 1.7407%，平均化後では 1.0804% 下がっている．

$a8a$ データセットでは $(i),(ii)$ 共に精度の向上が見られるが，$news20binary$ データセットでは (i) の時は精度が向上しているが，(ii) の時に精度が下がっている．

4.4 実験結果の解析

(i) 重みベクトルの初期値が $[-1,1]$ の実数の場合

重みベクトルの初期値を $[-1,1]$ の間の実数に設定し，平均化した重みベクトルを生成する．平均する数は $1,10,20,30,40,50,60,70,80,90,100$ とする．平均数 1 は，平均前の重みベクトルの内で最大精度であるものを示している．

$a8a$ データセットを用い，平均化した重みベクトルをテストデータにより精度測定した結果を図 6 に示す．図 6 より，平均数 10 と 30 の時に 86.0328% で最大精度である．平均数 1 の最大精度が 85.9518% であるので，平均化することで 0.081% 上がっている．また，平均数が $1,80$ 以外の場合で従来法よりも良い精度となっている．

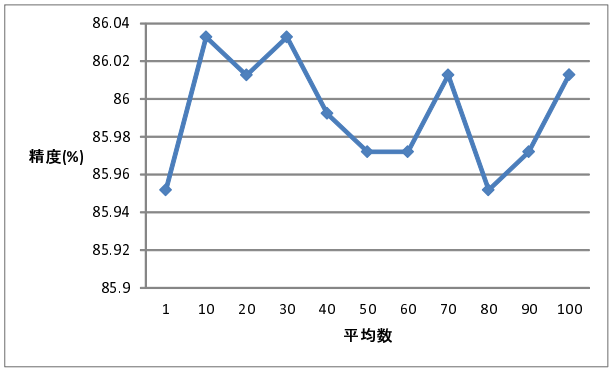


図 6: (i) の時の $a8a$ のテストデータでの精度

次に，$news20binary$ データセットを用い，平均化した重みベクトルをテストデータにより精度測定した結果を図 7 に示す．図 7 より，平均数 30 の時に 95.8183% で最大精度である．平均数 1 の最大精度が 95.6383% であるので，平均化することで 0.18% 上がっている．また，平均数が $1,90$ 以外のすべての平均数の場合で従来法よりも良い精度となっている．

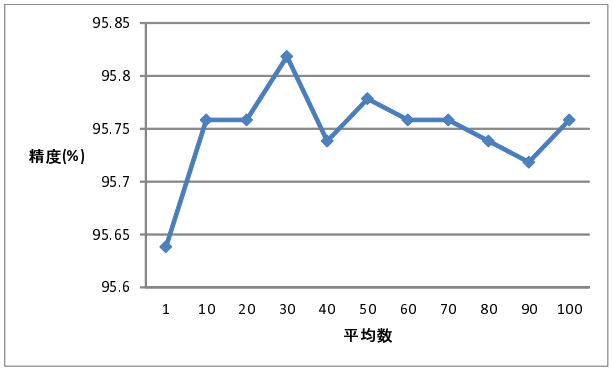


図 7: (i) の時の $news20binary$ のテストデータでの精度

(ii) 重みベクトルの初期値が $[-10,10]$ の範囲の実数の場合

重みベクトルの初期値を $[-10,10]$ の間の実数に設定し，平均化した重みベクトルを生成する．4.4 節と同様に，平均する数は $1,10,20,30,40,\ 50,60,70,80,90,100$ とする．

$a8a$ データセットを用い，平均化した重みベクトルをテストデータにより精度測定した結果を図 8 に示す．図 8 より，平均数 30 の時に 86.0531% で最大精度である．平均数 1 の最大

表 1: テストデータでの実験結果

	従来法	$(i)[-1,1]$		$(ii)[-10,10]$	
		平均化前	平均化後	平均化前	平均化後
$a8a$	85.9518%	85.9518%	86.0328%	86.0126%	85.9923%
$news20binary$	95.7183%	95.6383%	95.8183%	93.9776%	94.6379%

精度が 86.0126% であるので，平均化することで 0.0405% 上がっている．また，平均数が $20, 70, 90$ 以外の場合で従来法よりも良い精度となっている．

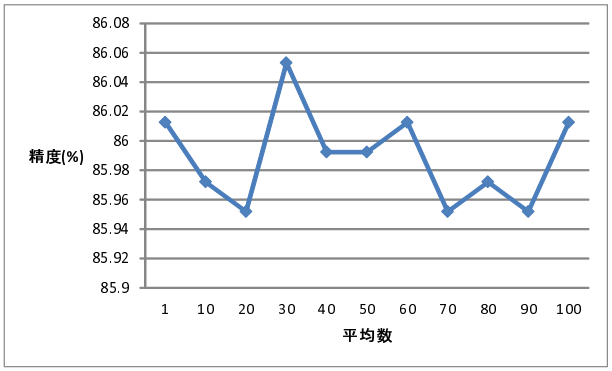


図 8: (ii) の時の $a8a$ のテストデータでの精度

次に，$news20binary$ データセットを用い，平均化した重みベクトルをテストデータにより精度測定した結果を図 9 に示す．図 9 より，平均数 10 の時に 94.7779% で最大精度である．平均数 1 の最大精度が 93.9776% であるので，平均化することで 0.8003% 上がっている．しかし，すべての平均数の場合で従来法の精度を下回っている．

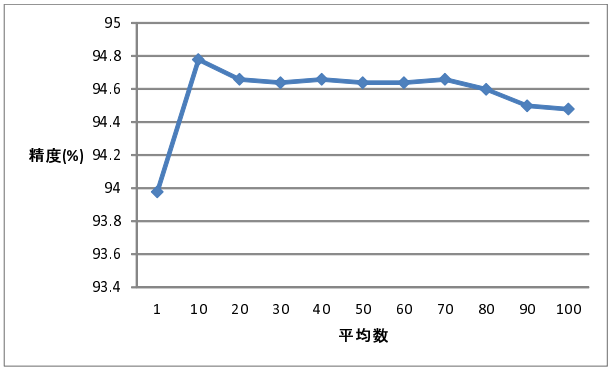


図 9: (ii) の時の $news20binary$ のテストデータでの精度

5. まとめ

本論文では，ロジスティック回帰において，初期値が乱数で設定された重みベクトルを用いて学習し生成された重みベクトルを平均化することで精度の向上を目指した．$a8a$ データセットでは重みベクトルの初期値が $[-1,1]$ と $[-10,10]$ の両方で，$news20binary$ データセットでは重みベクトルの初期値が $[-1,1]$ の場合で精度が向上した．

目的関数が最適解として取るパラメータがただ一つではないということが予想されるが，計算機実験から，複数回学習して重みベクトルを生成し，それぞれ精度に違いが生じたことから，今回の実験で用いたデータではその予想が正しいことを確認できた．また，それらの重みベクトルを平均化することで精度が向上することも確認できた．重みベクトルの初期値が $[-1,1]$ の場合，平均数が 30 の時に最大精度となった．重みベクトルの初期値が $[-10,10]$ の場合，平均数に共通点は確認できなかった．$news20binary$ データセットは，重みベクトルの初期値が $[-1,1]$ の時に精度が向上し，$[-10,10]$ の時に精度が下がった．これは，最適解を得るための重みベクトルの初期値が $[-1,1]$ の範囲に集中しているからと考えられる．また、これらの実験結果から，重みベクトルの初期値は $[-1,1]$ の範囲，平均数は 30 付近が最適ではないかと考えられる．

今後の課題として，重みベクトルの初期値や平均数を細かく設定し，さらなる精度の改善を行いたい．

参考文献

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. The Journal of Machine Learning Research. vol.9, pp.1871-1874 2008.

[2] Ralf Herbrich,Thore Graepel and Colon Campbell. Bayes Point Machines. The Journal of Machine Learning Research. vol.9, pp.245-279 2001.