G-08

大幾何マージン最小分類誤り学習法を用いた音声認識に関する実験的評価

Experimental Evaluation of Speech Recognition using Large Geometric Margin Minimum Classification Error Training

松廣 達也 †　北岡 見生代 †　ア デイビッド †　渡辺 秀行 ‡　片桐 滋 †　大崎 美穂 †
Tatsuya Matsuhiro　Mikiyo Kitaoka　David Ha　Hideyuki Watanabe　Shigeru Katagiri　Miho Osaki

1　はじめに

パターン認識における分類器設計法のひとつに最小分類誤り (MCE: Minimum Classification Error) 学習法がある [1,2]．MCE 学習法は，学習標本集合に対する分類誤り数を平滑な 0-1 損失関数を用いて近似的に表現し，その損失関数を最小化することによって，分類器設計の究極の目標である最小分類誤り確率状態に対応する分類器（クラスモデル）パラメータの状態の達成を直接的に目指す学習法である．また，この MCE 学習法に幾何マージン増大の機構を取り入れることによって，最小分類誤り確率状態を推定する能力をさらに高めた，大幾何マージン最小分類誤り (LGM-MCE: Large Geometric Margin-MCE) 学習法も提案されている [3]．以下本稿では，この 2 つの学習法を区別するために，初めの MCE 学習法を関数マージン最小分類誤り (FM-MCE: Functional Margin-MCE) 学習法と呼ぶことにする．

FM-MCE 学習法は，固定次元ベクトルパターンだけでなく，音声データのような可変長（時系列）パターンにおける分類問題にも広く適用され [4]，その有用性が確認されてきた．一方で，LGM-MCE 学習法に対する有用性の評価は，固定次元パターンの分類を対象としたものが多く，可変長パターン分類に向けた LGM-MCE 学習法が定式化 [5,6] され始まってはいるものの，その有用性は未だ十分に確認されているとは言い難い．

こうした状況を受け，本稿では，先行研究で行われた評価実験 [5,6] の規模（データ数やクラス数）を拡大させた，より大きなタスクにおいて，FM-MCE 学習法と LGM-MCE 学習法の比較を行い，音声パターン分類器の学習法としての LGM-MCE 学習法の有用性をより詳細に検証する．

また MCE 学習法では従来，損失関数の最小化法として勾配法の一種である確率的降下（PD: Probabilistic Descent）法が広く用いられてきた．しかし，PD 法には，学習において分類器パラメータの更新程度を制御するハイパーパラメータ，学習係数が存在する．この係数の値は，実験による試行錯誤等を通して経験的に設定せざるを得ず，その設定には多くの時間を要する．そこで本稿では，この時間浪費的なハイパーパラメータ設定の回避を目指し，分類器パラメータの更新程度を自動的に調整する機構を備えた最適化手法，RPROP 法を，LGM-MCE 法における損失最小化手続きに利用する効果についても調査する ．

2　状態遷移モデルのための最小分類誤り学習法

この節ではまず，パターン分類器の構造の基本を成す，時系列パターン分類に便利な状態遷移型のクラスモデルについて説明する．続いて，評価実験に用いる，状態遷移型モデルのための FM-MCE 学習法と LGM-MCE 学習法について説明する．

2.1　状態遷移構造を持つクラスモデル

状態遷移モデルからなる分類器を用いて，可変長パターン標本 $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ を J 個のクラス（本稿の実験では単語に相当する）$\{C_j \ (j = 1, \cdots, J)\}$ のうちの 1 つに分類する問題を考える．ここで，このベクトル系列の要素である $\boldsymbol{x}_t$ は，系列中のフレーム時刻指標 t における d 次元の音響特徴ベクトルであり，T は $\boldsymbol{X}$ における音響特徴ベクトルの数（フレーム数）である．

単語に相当するクラスのモデル，クラスモデルは，状態遷移構造をもつ音素モデルが連結されて構成される．ここで，音素モデルは，S 個の状態を持ち，また各状態において I 個のプロトタイプを持つ．例えば，h 番目の音素の s 番目の状態の i 番目のプロトタイプは，$\boldsymbol{r}_i^{h,s}$ で表わされる．

FM-MCE 法および LGM-MCE 法における学習対象は，単語クラスモデルであり，延いては音素モデルである．便利のため，その学習対象パラメータを $\boldsymbol{\Lambda} = \{\boldsymbol{r}_i^{h,s}\}_{h=1,s=1,i=1}^{H,\ S,\ I_{h,s}}$ として略記する．なおここで，H と S，$I_{h,s}$ は，それぞれ，音素クラスの数と，音素クラスに用いられる状態遷移モデルの状態数，その第 h 音素クラスモデルの第 s 状態に配置されるプロトタイプ数を示している．

図 1 は，2 つの音素，/a/と/o/に対応する音素モデルを連結して構成した単語 “あお” に対するクラスモデルを図解している．

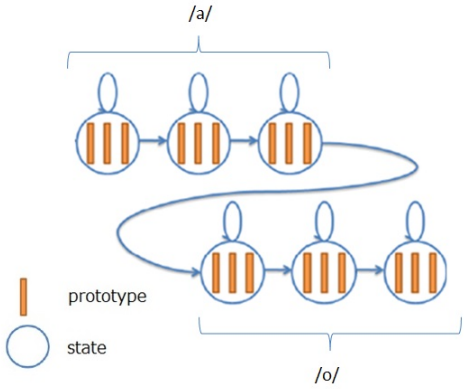

図 1　単語"あお"の状態遷移型クラスモデルの概念図.

固定次元ベクトルパターンの次元数と異なり，音声のような可変長パターンにおけるフレーム数はパターン毎に異なる．従って，そのようなパターンと（状態遷移型）モデルとの間の距離を測るさいには工夫が必要である．音声認識分野で広く行われてきたように，本稿で扱う分類器も，動的時間伸縮（DTW: Dynamic Time Warping）による距離計算を採用する．このとき，各単語クラスに対する識別関数を，DTW による最小累積距離として次のように定義する．

$$g_j(\boldsymbol{X}; \boldsymbol{\Lambda}) = \frac{1}{T} \sum_{t=1}^{T} \| \boldsymbol{x}_t - \boldsymbol{r}_{i(\varphi_{j,t}, \theta_{j,t}, t)}^{\varphi_{j,t}\ \theta_{j,t}} \|^2. \tag{1}$$

ここで，最適な経路 $\{(\varphi_{j,1}\ \theta_{j,1}); (\varphi_{j,2}\ \theta_{j,2}); \ldots; (\varphi_{j,T}\ \theta_{j,T})\}$ は，DTW によって自動的に見出される．また，$\varphi_{j,t}$ は j 番目のクラスにおける時刻 t での音素であり，$\theta_{j,t}$ は j 番目のクラスにおける時刻 t における状態の指標（index）である．さらに，$i(\varphi_{j,t}, \theta_{j,t}, t) = \arg\min_{i=1}^{I} \| \boldsymbol{x}_t - \boldsymbol{r}_i^{\varphi_{j,t}\ \theta_{j,t}} \|^2$ である．図 2 は，この DTW に基づく識別関数の計算過程を図解している．図は，横軸に入力パターンの音声特徴ベクトルを置き，縦軸に音声区間の始めと終わりに無音区間（"sil"で示す）を持つ，単語クラス"あお" の状態遷移モデルを置き，両者の間で行われる DTW の手続きを示している．

2.2　状態遷移モデルのための関数マージン最小分類誤り学習法

上述の DTW に基づく識別関数 (1) を得られるものとして，可変長音響特徴ベクトル列である音声入力パターン $\boldsymbol{X}$ を分類する課題は以下のように定形化することができる．

† 同志社大学，Doshisha University.
‡ 株式会社 国際電気通信技術研究所，ATR.

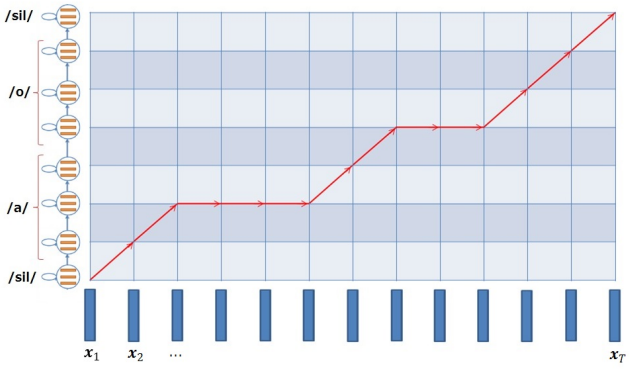

図 2　DTW による識別関数値の計算の概念．

$$C(\boldsymbol{X}) = C_i \quad i = \arg\min_j \; g_j(\boldsymbol{X};\boldsymbol{\Lambda}). \tag{2}$$

FM-MCE 学習は，まず，学習標本 $\boldsymbol{X}$（$\in C_y$）に対する，分類判断の正誤およびその程度を表す誤分類尺度を式 (1) の識別関数を用いて以下のように定義する．

$$d_y(\boldsymbol{X};\boldsymbol{\Lambda}) = g_y(\boldsymbol{X};\boldsymbol{\Lambda}) - \left[\frac{1}{J-1}\sum_{j,j\neq y} g_j(\boldsymbol{X};\boldsymbol{\Lambda})^{-\psi}\right]^{\frac{-1}{\psi}}, \tag{3}$$

ここで ψ は正の定数である．また，式 (3) を $\psi \to \infty$ としたとき，誤分類尺度は次のように簡単化される．

$$\begin{aligned} d_y(\boldsymbol{X};\boldsymbol{\Lambda}) &= g_y(\boldsymbol{X};\boldsymbol{\Lambda}) - g_{y^*}(\boldsymbol{X};\boldsymbol{\Lambda}), \\ y^* &= \arg\min_{j,j\neq y} g_j(\boldsymbol{X};\boldsymbol{\Lambda}). \end{aligned} \tag{4}$$

この誤分類尺度は，正値によって誤分類を，負値によって正分類を示す．従って，誤分類尺度の関数として以下の 0-1 損失関数を導入することによって，学習標本に対する分類誤りの個数を求めることが可能となる．

$$\ell_y(\boldsymbol{X};\boldsymbol{\Lambda}) = \begin{cases} 1 & \text{if} \quad d_y(\boldsymbol{X};\boldsymbol{\Lambda}) > 0 \\ 0 & \text{if} \quad d_y(\boldsymbol{X};\boldsymbol{\Lambda}) < 0 \end{cases}. \tag{5}$$

しかし，式 (5) は学習対象である分類器パラメータ $\boldsymbol{\Lambda}$ に関して微分不可能であり，さらにほとんど常にその勾配値はゼロであり，PD 法などの勾配法を用いて分類器パラメータの最小化をすることは難しい．そこで，式 (5) を，平滑な 0-1 損失関数であるシグモイド関数で近似する．近似された損失関数は，以下のように定義できる．

$$\ell_y(d_y(\boldsymbol{X};\boldsymbol{\Lambda})) = \frac{1}{1+\exp(-\alpha d_y(\boldsymbol{X};\boldsymbol{\Lambda}))}, \tag{6}$$

ここで α は正の定数であり，損失関数の傾きを表す．

個々の学習標本に対して定義された (6) の損失は，多数の学習標本群に対して最小化されなければならない．この最小化のために，個々の損失を学習標本群上で統合して，以下の経験的平均損失を定義する．

$$L(\boldsymbol{\Lambda}) = \frac{1}{N}\sum_{n=1}^{N} \ell_{y_n}(d_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})), \tag{7}$$

ここで $\boldsymbol{X}_n$ は n 番目の学習標本を表し，y_n はその学習標本の所属するクラスを表している．なお，この学習標本を表す指標に合わせて，$\boldsymbol{X}_n$ の要素に関しては $\boldsymbol{X}_n = [\boldsymbol{x}_1^n, \cdots, \boldsymbol{x}_t^n, \cdots, \boldsymbol{x}_{T_n}^n]$ と記述することとする．

FM-MCE 学習法は，こうして定義された $L(\boldsymbol{\Lambda})$ の，例えば PD 法などを用いた最小化を通して，その少なくとも極小点に対応する分類器パラメータ $\boldsymbol{\Lambda}$ の状態を求める．

分類器パラメータ $\boldsymbol{\Lambda}$ の最適化（損失の最小化）の手続きについては，LGM-MCE 学習法と合わせて第 3 節で説明する．

2.3　状態遷移モデルのための大幾何マージン最小分類誤り学習法

LGM-MCE 学習法は，FM-MCE 学習法が持つ，分類誤り数損失を最小化する学習に加えて，新しく幾何マージンを最大化する学習を行う．ここで幾何マージンとは，標本空間におけるクラス境界とそれに最も近い正分類される学習標本との距離である．幾何マージン増大化の概念が加えられることで，LGM-MCE 学習は，最小分類誤り確率状態に対応する分類器パラメータ状態の推定精度を高めることが期待される．なぜなら，LGM-MCE 学習法における幾何マージンの増加は，与えられた有限個の学習標本の周囲に仮想標本を生成し，その仮想標本の重なりによって，無限個の標本上で定義されるベイズ誤り，すなわち最小分類誤り確率状態を，自然な形で近似できるためである．

時系列パターンを扱う場合，固定次元ベクトルパターンの場合とは異なり，標本毎にフレーム（次元）数に違いが生じる．そのため，(音響特徴ベクトル列に対応する）パターン空間における幾何マージンの計算は，固定次元パターンの場合におけるパターン間のユークリッド距離の計算のようなわけにはいかない．本稿では，先の導出結果 [6] に従い，入力される音響特徴ベクトル系列（パターン）と，DTW の意味でそれに最も近い系列として状態遷移モデルから選ばれる音響特徴ベクトル系列（パターン）間の距離を用いて，可変長パターンに対する幾何マージン r を以下のように定義した．

$$\begin{aligned} D_y(\boldsymbol{X};\boldsymbol{\Lambda}) &= -r \\ &= \frac{\sqrt{T}\, d_y(\boldsymbol{X};\boldsymbol{\Lambda})}{2\sqrt{\sum_{t=1}^{T} \left\| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \right\|^2}}. \end{aligned} \tag{8}$$

なお，この (8) における $D_y(\boldsymbol{X};\boldsymbol{\Lambda})$ は，FM-MCE 学習法における式 (4) の誤分類尺度に代わる，LGM-MCE 学習法用の誤分類尺度である．

FM-MCE 学習法と LGM-MCE 学習法の主たる違いは，この誤分類尺度の定義の違いにある．LGM-MCE 学習における，誤分類尺度を変数とする（個別の）損失や経験的平均損失の定義は，基本的に FM-MCE 学習のそれと同じである．FM-MCE 学習法における式 (6) と式 (7) における $d_y(\boldsymbol{X};\boldsymbol{\Lambda})$ を $D_y(\boldsymbol{X};\boldsymbol{\Lambda})$ によって入れ替え，LGM-MCE 学習のための損失 (9) と経験的平均損失 (10) を得る．学習は，学習標本群上でこの経験的平均損失の最小化を行い，分類誤り数が小さく，かつ幾何マージンが大きな分類器パラメータ状態の発見を目指す．

$$\ell_y(D_y(\boldsymbol{X};\boldsymbol{\Lambda})) = \frac{1}{1+\exp(-\alpha D_y(\boldsymbol{X};\boldsymbol{\Lambda}))}. \tag{9}$$

$$L(\boldsymbol{\Lambda}) = \frac{1}{N}\sum_{n=1}^{N} \ell_{y_n}(D_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})). \tag{10}$$

3　確率的降下法と RPROP 法

この節では，本研究の評価実験で使用した 2 つの損失最小化法について説明し，さらにその中で FM-MCE 学習法と LGM-MCE 学習法の分類器パラメータの更新式についても詳述する．

3.1　確率的降下法

3.1.1　概要

PD 法は，最近では確率的勾配降下（SGD: Stochastic Gradient Desent）法と呼ばれることも多いが，60 年代のパターン分類器学習の研究において提案された，適応型あるいは継時型の勾配法の一種ある．勾配法とは，最小化対象である損失の勾配情報に基づいて，損失の超曲面の表面を下る方向にパラメータを更新することで損失の最小状態を探索する方法である．勾配法の概念を図 3 に図解する．図中，横軸はモデルパラメータであり，縦軸が損失の大きさである．全学習標本に対して定義される式 (7) や式 (10) で表す経験的平均損失の勾配情報に基づいてパラメータの更新を行う手法は最急降下法と呼ばれ，個々の学習標本に対して定義される式 (6) や式 (9) で表す（個別の）損失の勾配情報に基づいて逐次的にパラメータの更新を行う手法は確率

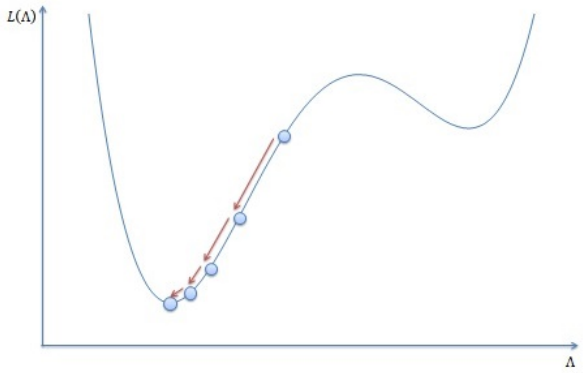

図 3　勾配法の概念図.

的降下（PD）法と呼ばれる．状態遷移型クラスモデルを持つ分類器のための FM-MCE 学習法では PD 法が用いられてきた．

式 (6) や式 (9) における一般的な（状態遷移モデルであることやプロトタイプ，あるいは重み係数などのようにパラメータの定義を特定しない）分類器パラメータ $\boldsymbol{\Lambda}$ に対して，PD 法はパラメータの更新式を以下のように与える．

$$\begin{aligned}
\boldsymbol{\Lambda}^{(m+1)} &= \boldsymbol{\Lambda}^{(m)} - \epsilon_m \nabla_{\boldsymbol{\Lambda}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda}^{(m)}). \qquad (11)\\
m &: \text{繰り返し計算のステップ番号 } (m = 1, 2, 3, \cdots),\\
\boldsymbol{\Lambda}^{(m)} &: \text{m ステップ目における更新前のパラメータ値,}\\
\boldsymbol{X} &: \text{m ステップ目に入力される学習標本,}\\
y &: \text{m ステップ目に入力される学習標本の正解クラスの指標,}\\
\epsilon_m &: \text{m ステップ目における学習係数，} \epsilon_m > 0,
\end{aligned}$$

$$\begin{aligned}
\nabla_{\boldsymbol{\Lambda}} \ell_y &= \frac{\partial \ell_y}{\partial \boldsymbol{\Lambda}} = \left[\frac{\partial \ell_y}{\partial \lambda_1} \frac{\partial \ell_y}{\partial \lambda_2} \cdots \frac{\partial \ell_y}{\partial \lambda_L}\right]^T\\
&\quad \text{(全パラメータによる微分),}\\
\boldsymbol{\Lambda} &= [\lambda_1 \lambda_2 \cdots \lambda_L]^T \quad L : \text{学習パラメータ数.}
\end{aligned}$$

3.1.2　状態遷移型分類器パラメータのための更新式

式 (11) を，本稿で扱う状態遷移型分類器パラメータ用に具体的に展開することで，状態遷移モデルの各状態に配置するプロトタイプ $\boldsymbol{r}_i^{h,s}$ 毎の PD 法による基本的な更新式を次のように得る．

$$\boldsymbol{r}_i^{h,s\,(m+1)} = \boldsymbol{r}_i^{h,s\,(m)} - \epsilon_m \nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda}^{(m)}). \qquad (12)$$

しかし，式 (12) におけるパラメータ更新量 $\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda}^{(m)})$ は，FM-MCE 学習法と LGM-MCE 学習法によって異なる．両学習法のそれぞれに対し，このパラメータ更新量の具体的な定義は以下のように与えられる．

・FM-MCE 学習法におけるパラメータ更新量

$\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ は鎖則より以下のように表すことができる．

$$\begin{aligned}
&\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})\\
&= \frac{\partial \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \left(\frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_y(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_y(\boldsymbol{X}; \boldsymbol{\Lambda}) \right.\\
&\qquad \left. + \frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda}) \right). \qquad (13)
\end{aligned}$$

また，式 (13) 中の損失関数の微分と誤分類尺度の微分，識別関数の微分は，それぞれ以下のようにして表すことができる．

$$\frac{\partial \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})} = \alpha \ell_y(d_y(\boldsymbol{X}; \boldsymbol{\Lambda}))\{1 - \ell_y(d_y(\boldsymbol{X}; \boldsymbol{\Lambda}))\}, \qquad (14)$$

$$\frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_y(\boldsymbol{X}; \boldsymbol{\Lambda})} = 1, \quad \frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda})} = -1, \qquad (15)$$

$$\begin{aligned}
\nabla_{\boldsymbol{r}_i^{h,s}} g_y(\boldsymbol{X}; \boldsymbol{\Lambda}) = -\frac{2}{T} \sum_{t=1}^{T} &\delta(\varphi_{y,t} - h)\delta(\theta_{y,t} - s)\\
&\cdot \delta(i(\varphi_{y,t}, \theta_{y,t}, t) - i) \left(\boldsymbol{x}_t - \boldsymbol{r}_i^{h,s} \right), \qquad (16)
\end{aligned}$$

$$\delta(n) = \begin{cases} 1 & \text{if} \quad n = 0 \\ 0 & \text{if} \quad n \neq 0 \end{cases}. \qquad (17)$$

学習標本が最も誤り易いクラス C_{y^*} に対する $\nabla_{\boldsymbol{r}_i^{h,s}} g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda})$ は，式 (16) と同様にして求めることができる．ただし，そこで用いられる DTW の最適経路は，そのクラス用に選択されたものである．

・LGM-MCE 学習法におけるパラメータ更新量

まず，式 (8) を鎖則で表現しやすいよう次のように表現しておく．

$$\begin{aligned}
D_y(\boldsymbol{X}; \boldsymbol{\Lambda}) &= \frac{\sqrt{T} d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{2\sqrt{\sum_{t=1}^{T} \left\| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t}, \theta_{y,t}, t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t}, \theta_{y^*,t}, t)} \right\|^2}} \qquad (18)\\
&= \frac{\sqrt{T} d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{N(\boldsymbol{\Lambda})}.
\end{aligned}$$

FM-MCE 学習法と同様に鎖則を用いると，$\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ は以下のようになる．

$$\begin{aligned}
&\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})\\
&= \frac{\partial \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial D_y(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \left\{ \frac{\sqrt{T}}{N(\boldsymbol{\Lambda})} \cdot \left(\frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_y(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_y(\boldsymbol{X}; \boldsymbol{\Lambda}) \right.\right.\\
&\qquad \left. + \frac{\partial d_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_{y^*}(\boldsymbol{X}; \boldsymbol{\Lambda}) \right)\\
&\qquad \left. - \frac{\sqrt{T} d_y(\boldsymbol{x}_1^T; \boldsymbol{\Lambda})}{N(\boldsymbol{\Lambda})^2} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} N(\boldsymbol{\Lambda}) \right\}. \qquad (19)
\end{aligned}$$

また，新しく式 (19) に登場した $\frac{\partial \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial D_y(\boldsymbol{X}; \boldsymbol{\Lambda})}$ と $\nabla_{\boldsymbol{r}_i^{h,s}} N(\boldsymbol{\Lambda}^{(m)})$ は，それぞれ以下のように表すことができる．

$$\frac{\partial \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})}{\partial D_y(\boldsymbol{X}; \boldsymbol{\Lambda})} = \alpha \ell_y(D_y(\boldsymbol{X}; \boldsymbol{\Lambda}))\{1 - \ell_y(D_y(\boldsymbol{X}; \boldsymbol{\Lambda}))\}, \qquad (20)$$

$$\begin{aligned}
&\nabla_{\boldsymbol{r}_i^{h,s}} N(\boldsymbol{\Lambda}) =\\
&\frac{2 \sum_{t=1}^{T} (Q)}{\sqrt{\sum_{t=1}^{T} \left\| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t}, \theta_{y,t}, t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t}, \theta_{y^*,t}, t)} \right\|^2}}, \qquad (21)
\end{aligned}$$

ここで，Q は次式で表すものとする．

$$\begin{aligned}
Q = \{&\delta(\varphi_{y,t} - h)\delta(\theta_{y,t} - s)\delta(i(\varphi_{y,t}, \theta_{y,t}, t) - i)\\
&- \delta(\varphi_{y^*,t} - h)\delta(\theta_{y^*,t} - s)\delta(i(\varphi_{y^*,t}, \theta_{y^*,t}, t) - i)\}\\
&\cdot \left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t}, \theta_{y,t}, t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t}, \theta_{y^*,t}, t)} \right). \qquad (22)
\end{aligned}$$

3.2　RPROP 法

3.2.1　概要

RPROP 法は，以下のように勾配の符号のみを用いて，予め設定しておいたパラメータ更新量を最適な値に向けて変化させつつ，損失の超曲面を下るようにしてその局所的最小状態の発見を目指す [7,8]．また，PD 法とは異なり，RPROP 法ではパラメータベクトルの要素毎に更新量が変化する．

$$r_{i\ (\mathrm{dim})}^{h,s\,(m+1)} = r_{i\ (\mathrm{dim})}^{h,s\,(m)} - \mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})^{\ (m)}}{\partial r_{i\ (\mathrm{dim})}^{h,s}} \right) \cdot \triangle_{i\ (\mathrm{dim})}^{h,s\,(m)}. \qquad (23)$$

ここで $\triangle_i^{h,s}{}^{(m)}_{(\mathrm{dim})}$ は，更新繰り返しの m ステップにおける h 番目の音素の s 番目の状態の i 番目のプロトタイプの dim 次元のパラメータに対する更新量を表している．この更新量は，例えば現在 m ステップの更新であるとした場合，それに先行する $(m-1)$ ステップの勾配の符合と m ステップの勾配の符合が等しいときはさらに大きく曲面を下ることができるとみなして更新量を増やし，符号が異なる場合は局小解を飛び越えたとみなして $(m-1)$ ステップのパラメータに戻し，さらに更新量も減らす処理を行うことで更新量を調整する．

RPROP 法には複数の版があり，本研究では特に RPROP$^+$ と呼ばれる RPROP 法を用いた [7,8]．RPROP$^+$ の具体的な更新手続きは，式 (23) に基づく次の疑似コードによって表すことができる．

Algorithm 1 RPROP$^+$ algorithm

for each $r_i^{h,s}{}_{(\mathrm{dim})}$ **do**

if $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m-1)} \cdot \frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} > 0$ **then**

$$\triangle_i^{h,s}{}^{(m)}_{(\mathrm{dim})} = \min\left(\triangle_i^{h,s}{}^{(m-1)}_{(\mathrm{dim})} \cdot \eta^+, \triangle_{max}\right)$$

$$\triangle r_i^{h,s}{}^{(m)}_{(\mathrm{dim})} = -\mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}\right) \cdot \triangle_i^{h,s}{}^{(m)}_{(\mathrm{dim})}$$

$$r_i^{h,s}{}^{(m+1)}_{(\mathrm{dim})} = r_i^{h,s}{}^{(m)}_{(\mathrm{dim})} + \triangle r_i^{h,s}{}^{(m)}_{(\mathrm{dim})}$$

else if $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m-1)} \cdot \frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} < 0$ **then**

$$\triangle_i^{h,s}{}^{(m)}_{(\mathrm{dim})} = \max\left(\triangle_i^{h,s}{}^{(m-1)}_{(\mathrm{dim})} \cdot \eta^-, \triangle_{min}\right)$$

$$r_i^{h,s}{}^{(m+1)}_{(\mathrm{dim})} = r_i^{h,s}{}^{(m)}_{(\mathrm{dim})} - \triangle r_i^{h,s}{}^{(m-1)}_{(\mathrm{dim})}$$

$$\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} = 0$$

else

$$\triangle r_i^{h,s}{}^{(m)}_{(\mathrm{dim})} = -\mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}\right) \cdot \triangle_i^{h,s}{}^{(m)}_{(\mathrm{dim})}$$

$$r_i^{h,s}{}^{(m+1)}_{(\mathrm{dim})} = r_i^{h,s}{}^{(m)}_{(\mathrm{dim})} + \triangle r_i^{h,s}{}^{(m)}_{(\mathrm{dim})}$$

end if

end for

疑似コード中の $\triangle_{max}$ は更新量の最大値であり，$\triangle_{min}$ は更新量の最小値である．また，$0 < \eta^- < 1 < \eta^+$ である．$\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}$ は m ステップにおける経験的平均損失を $r_i^{h,s}{}_{(\mathrm{dim})}$ で微分した値を表す．

3.2.2 パラメータ更新量

疑似コードの手順に従い，分類器パラメータの更新を行う．疑似コード中の $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}$ は鎖則によって以下のように表すことができる．

$$\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}} = \frac{1}{N}\sum_{n=1}^{N} \nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}). \tag{24}$$

ここでパラメータ更新量 $\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})$ は，前節の式 (13) や式 (19) において，パラメータがベクトル $\boldsymbol{r}_i^{h,s}$ であるときの dim 次元目の要素に関するものである．FM-MCE 学習法と LGM-MCE 学習法に対し，パラメータ更新量の具体的な定義はそれぞれ次のよう与えられる．

・FM-MCE 学習法の場合

$$\begin{aligned}
&\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}) = \\
&\quad \alpha \ell_{y_n}(d_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}))\{1-\ell_{y_n}(d_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}))\} \\
&\quad \cdot \Bigg(1 \cdot \Bigg\{-\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y,t}-h)\delta(\theta_{y,t}-s) \cdot \\
&\qquad \delta(i(\varphi_{y,t},\theta_{y,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right)\Bigg\} \\
&\qquad + (-1) \cdot \Bigg\{-\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y^*,t}-h)\delta(\theta_{y^*,t}-s) \\
&\qquad \cdot\delta(i(\varphi_{y^*,t},\theta_{y^*,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right)\Bigg\}\Bigg).
\end{aligned} \tag{25}$$

・LGM-MCE 学習法の場合

$$\begin{aligned}
&\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}) = \\
&\quad \alpha \ell_{y_n}(D_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}))\{1-\ell_{y_n}(D_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}))\} \\
&\quad \cdot \Bigg\{\frac{\sqrt{T}}{N(\boldsymbol{\Lambda})} \cdot \Bigg(1 \cdot \Bigg\{-\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y_n,t}-h)\delta(\theta_{y_n,t}-s) \\
&\qquad \cdot\delta(i(\varphi_{y_n,t},\theta_{y_n,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right)\Bigg\} \\
&\qquad + (-1) \cdot \Bigg\{-\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y_n{}^*,t}-h)\delta(\theta_{y_n{}^*,t}-s) \\
&\qquad \cdot\delta(i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right)\Bigg\}\Bigg) \\
&\qquad -\frac{\sqrt{T} d_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})}{N(\boldsymbol{\Lambda})^2} \\
&\quad \cdot \frac{2\sum_{t=1}^{T}(Q')}{\sqrt{\sum_{t=1}^{T} \| r^{\varphi_{y_n,t}\ \theta_{y_n,t}}_{i(\varphi_{y_n,t},\theta_{y_n,t},t)(\mathrm{dim})} - r^{\varphi_{y_n{}^*,t}\ \theta_{y_n{}^*,t}}_{i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)(\mathrm{dim})} \|^2}}\Bigg\},
\end{aligned} \tag{26}$$

ここで，Q' は次式で表すものとする．

$$\begin{aligned}
Q' = &\{\delta(\varphi_{y,t}-h)\delta(\theta_{y,t}-s)\delta(i(\varphi_{y,t},\theta_{y,t},t)-i) \\
&- \delta(\varphi_{y^*,t}-h)\delta(\theta_{y^*,t}-s)\delta(i(\varphi_{y^*,t},\theta_{y^*,t},t)-i)\} \\
&\cdot \left(r^{\varphi_{y_n,t}\ \theta_{y_n,t}}_{i(\varphi_{y_n,t},\theta_{y_n,t},t)(\mathrm{dim})} - r^{\varphi_{y_n{}^*,t}\ \theta_{y_n{}^*,t}}_{i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)(\mathrm{dim})}\right).
\end{aligned} \tag{27}$$

4 評価実験

4.1 実験概要

評価は，孤立単語音声パターンの分類タスクにおいて行った．用いた音声データは，国立情報学研究所が提供している ETL-WD-I&Ⅱ データセットである [1]．このデータセットは，男女各 10 名の計 20 名による単語リスト 1542 語の読み上げ音声で構成されている．しかし，このうち，女性の 558 単語データが収録されておらず，実験には，男性データと女性データがともに存在する 984 単語の音声データを用いた．従って，タスクは，1 クラスあたり 20 のパターン標本を持つ 984 クラス問題（標本総数は 19680）であった．

各時間窓位置における音響特徴量には，12 次元のメル周波数ケプストラム係数（MFCC: Mel-Frequency Cepstrum Coefficient）とその窓内の音声信号のパワーとからなる 13 次元の変数ベクトルと，さらにそれら各変数の 1 次微分値からなる 13 次元ベクトルを合わせた，合計 26 次元ベクトルのベクトルを用いた．

本実験に用いた各音素の状態遷移モデルは，それぞれ 3 状態から成り，各状態に 3 つのプロトタイプを配置した．ただし，無音区間を表す"sil"については，1 状態モデルを用い，その状態におけるプロトタイプ数も 1 とした．各状態遷移モデルのプロトタイプの初期化には segmental K 平均法を用いた [9]．

[1] http://research.nii.ac.jp/src/ETL-WD.html

評価方法としては，男女各 3 名ずつ，計 6 名分の単語を学習用標本集合とし，残りの男女各 7 名ずつ，計 14 名分の単語を未知標本集合とした Hold-Out 法を用いた．学習標本集合の標本数は 5904 であり，未知標本集合の標本数は 13776 であった．これら 2 つの標本集合に対する分類精度を用いて手法の有効性等の検証を行った．便利のため，学習標本を用いた評価を “Closed Test” と，未知標本集合を用いた評価のことを “Open Test” と呼ぶこととする．

実験に用いたハイパーパラメータを表 1 に示す．表 1 における損失関数の傾きは，値を大きくすると式 (5) の 0-1 損失に近づき，値を小さくすると 0-1 損失から離れるものの，仮想的に学習標本数を増やす効果が大きくなる．そのため，最小分類誤り確率状態の優れた推定に結びつく学習を行うためには，この傾きを適切に設定する必要がある．また，前述したように，学習係数や更新量の初期値も，分類器パラメータの更新量に，ひいては達成する分類精度にも影響する．これらの点を考慮し，実験では,（予備実験を通して設定した）一定の範囲内でこれらのハイパーパラメータ値を変化させ，それぞれのハイパーパラメータ値に対する Open Test と Closed Test の分類精度を調査した．

なお，FM-MCE 学習と LGM-MCE 学習のいずれにおいても，分類器パラメータの更新は，5904 個の学習標本を繰り返し用いて行った．その学習手続きにおいて,（5904 個の）全学習標本を一度ずつ用いるパラメータ更新の繰り返しをエポック（epoch）と呼ぶ．エポックの繰り返しに伴う経験的平均損失の減少（学習の収束）の様子を予備実験によって調査し，十分に減少を確認できた 50 エポックと，やや過剰とも考えられる 100 エポックとの学習を検証した．100 エポックの学習を行ったときのハイパーパラメータの設定法は，表 2 に示す通りとした．

表 1　ハイパーパラメータ設定 (最大エポック数: 50)

更新方法	ハイパーパラメータ	FM-MCE	LGM-MCE
PD	損失関数の傾き (α)	1.0 〜 8.0 (0.5 刻み)	10 〜 80 (10 刻み)
	学習係数 (ϵ)	0.5 〜 4.0 (0.5 刻み)	0.1 〜 1.2 (0.1 刻み)
RPROP	損失関数の傾き (α)	1.0 〜 8.0 (0.5 刻み)	10 〜 80 (10 刻み)
	更新量の初期値 (△)	2^n (n=-4,-3,...,1)	2^n (n=-4,-3,...,1)

表 2　ハイパーパラメータ設定 (最大エポック数: 100)

更新方法	ハイパーパラメータ	FM-MCE	LGM-MCE
PD	損失関数の傾き (α)	0.5 〜 8.0 (0.5 刻み)	5 〜 50 (5 刻み)
	学習係数 (ϵ)	0.5 〜 4.0 (0.5 刻み)	0.4 〜 2.0 (0.2 刻み)
RPROP	損失関数の傾き (α)	0.5 〜 8.0 (0.5 刻み)	5 〜 50 (5 刻み)
	更新量の初期値 (△)	0.2 〜 0.5 (0.1 刻み)	0.2 〜 0.5 (0.1 刻み)

4.2　結果と考察

2 つの異なる最大エポック数を用いて行った評価実験の結果を表 3 と表 4 に示す．ここで，表中の（PD 法か RPORP 法かの選択も含む）各学習法に対する分類精度は，いずれも，Closed Test において最も高い分類精度を達成した学習結果に基づくものである．学習が最小分類誤り確率状態の優れた推定を行った場合，学習標本に対する過学習を避けて達成された分類器は，Closed Test の分類精度を必要以上に高めることなく，むしろ未知標本に対する（Open Test の）分類精度を高めることが期待される．従って，上記のように Closed Test の結果を基準に選んだ表中の Open Test の分類精度を通して，学習の目標である最小分類誤り確率状態の達成程度の観点から，それぞれの学習手続きを評価検討することができる．

表 3　epoch 数 50 における各認識率 (%)

	FM-MCE (PD)	LGM-MCE (PD)	FM-MCE (RPROP)	LGM-MCE (RPROP)
Closed Test	99.93	99.95	99.86	99.85
Open Test	92.06	93.46	90.80	92.61

表 4　epoch 数 100 における各認識率 (%)

	FM-MCE (PD)	LGM-MCE (PD)	FM-MCE (RPROP)	LGM-MCE (RPROP)
Closed Test	99.98	99.98	99.92	99.93
Open Test	90.53	92.00	92.61	92.15

表 3 は，PD 法と RPROP 法のいずれにおいても，LGM-MCE 学習法が競合する FM-MCE 学習法を上回る分類精度を達成していることを示している．一方，表 4 は，RPROP 法の結果が FM-MCE 学習法において PD 法の結果を若干上回ってはいるものの，より忠実に損失の最小化を行っている PD 法を用いた場合は，やはり LGM-MCE 学習法が FM-MCE 学習法に勝ったことを示している．これらの結果は，LGM-MCE 学習法における幾何マージンの存在が，学習が目指す最小分類誤り確率状態（およびそれに対応する分類器パラメータの状態）の優れた推定に貢献したからと考えられる．

また表 3 と表 4 から，PD 法がより忠実に損失最小化を行っていることを読み取ることができる．2 つの表中において，PD 法は Closed Test においてより高い分類精度を達成し，その一方で Open Test においては精度をやや低くしている．これと比べると，損失勾配情報の一部（符号のみ）しか使わない RPROP 法は，学習標本に対する損失最小化を，ややおろそかにしているともみなすことができる．その副作用として，RPROP 法を用いた学習は，学習標本に対する学習を抑制した分，未知標本に対する分類精度を高めることができたと解釈することもできる．

最大エポック数が異なる 2 つの表の結果を比べると，LGM-MCE 学習法の分類精度が FM-MCE 学習法のそれに比べて（エポック数の違いによらず）安定していることがわかる．学習の繰り返し回数は，基本的にデータセットに依存する．従って，様々なデータに対して行われる学習の現実においては，学習の繰り返し回数が表 3 のように適切かやや少なめな状態になってしまったり，表 4 のようにやや多過ぎる状態になってしまったりすることが考えられる．LGM-MCE 学習法は，学習回数の多少にあまり影響されずに，安定的に高い分類精度をもたらし得ることを期待できる．この LGM-MCE 学習法の特質は，やはり，幾何マージン型誤分類尺度を用いることによって得られた，最小分類誤り確率状態の優れた推定力に拠るものと考えられる．

RPROP 法を用いた場合の Open Test に対する結果が，損失勾配情報の利用を簡略化した副作用である可能性がある以上，優れた最小分類誤り確率状態とそれに対応する分類器パラメータの推定に RPROP 法を利用することには慎重であるべきと考える．しかし，その一方，バッチ処理型である RPROP 法は，手続きの並列化による学習の高速化に明らかに適している．この点では，同じバッチ処理型である最急降下法も同様である．しかし，最急降下法も，PD 法と同様に学習係数の大きさを適切に設定する必要があり，その設定には手間がかかる．大量の学習標本を用いた，より高い信頼性を持つ分類器パラメータの学習の実現を目指すとき，並列化された RPROP 法による LGM-MCE 学習法が有望な学習法となり得ることが期待できる．

5　おわりに

可変長の音声パターン分類においては必ずしも十分に評価されていなかった LGM-MCE 学習法について，単語音声パターンの分類実験を通して，実験的な評価を行った．実験の結果，FM-MCE 学習法と比べ，LGM-MCE 学習法は，未知標本に対して安定的に高い分類精度を達成し得ることが確認された．

また，損失最小化に利用した PD 法と RPROP 法の比較からは，勾配情報利用に関する両者の違いが実験結果にもほぼ忠実に現れていた．定形化された学習手続きに忠実に学習を進める場合，勾配情報を忠実に利用する PD 法の方が手続きに沿った結果をもたらすように思われる．その一方で，RPROP 法は，学習標本に対する過学習を抑え，未知標本に対して高い分類精度をもたらすことがあった．しかし，この高い分類精度が，RPROP

法の学習手続きの副作用である可能性が大きく，精度向上を期待する利用はあまり勧められないように考える．

謝辞

本研究の一部は，科研費（番号：26280063）及び私学研究基盤形成支援事業「ドライバ・イン・ザ・ループ」の支援を受けて行われた．

参考文献

1) Biing-Hwang Juang and Shigeru Katagiri. Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing, Vol. 40, No. 12, pp. 3043-3054, 1992.
2) Katagiri, Shigeru and Juang, Biing-Hwang and Lee, Chin-Hui. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. Proceedings of the IEEE, Vol. 86, No. 11, pp. 2345-2373, 1998.
3) 渡辺秀行, 片桐滋, 山田幸太, 中村篤, 渡部晋治, 大崎美穂ほか. 幾何マージンに基づく誤分類尺度を用いた最小分類誤り学習法. 電子情報通信学会論文誌 D, Vol. 94, No. 10, pp. 1664-1675, 2011.
4) Erik McDermott and Shigeru Katagiri. Prototype-based mce/gpd training for word spotting and connected word recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, pp. 291-294, 1993.
5) 橋本哲也, 北岡見生代, 渡辺秀行, 片桐滋, ルシュガン, 堀智織, 大崎美穂. 大幾何マージン最小分類誤り学習法を用いた音声パターン認識. 日本音響学会春季研究発表会, 2015.
6) Mikiyo Kitaoka, Tetsuya Hashimoto, Tsubasa Ochiai, Shigeru Katagiri, Miho Ohsaki, Hideyuki Watanabe, Xugang Lu, and Hisashi Kawai. Speech pattern classification using large geometric margin minimum classification error training. IEEE Region 10 Conference in TENCON 2015, pp. 1-6, 2015.
7) Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In Neural Networks, 1993., IEEE International Conference on, pp. 586-591. IEEE, 1993.
8) Christian Igel and Michael Husken. Improving the RPROP learning algorithm. In Proceedings of the second international ICSC symposium on neural computation (NC 2000), Vol. 2000, pp. 115-121. Citeseer, 2000.
9) Biing-Hwang Juang and Lawrence R Rabiner. The segmental k-means algorithm for estimating parameters of hidden markov models. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 9, pp. 1639-1641, 1990.