C-04

幾何マージン増大型学習におけるデータ分割法について

On Data Partitioning for Large-Geometric-Margin-Based Training

白石 裕之†　竹中 一馬†　渡辺 秀行‡　片桐 滋†　大崎 美穂†　堀 智織‡
Hiroyuki Shiraishi Kazuma Takenaka Hideyuki Watanabe Shigeru Katagiri Miho Ohsaki Chiori Hori

1. 概要

パターン認識における認識器（分類器）学習の究極の目標は，（無限に発生し得る）未知パターン標本を正確に分類できる分類器パラメータ（クラスモデル）を実現することである．しかし，学習に用いることができる学習用標本は有限であり，学習の後に分類器が扱わなくてはならない未知標本を完全に学習に反映することは極めて困難である．また，クラスモデルの表現能力も一般には有限であり，未知標本データが持つであろうクラス境界を正確に表現することは，やはり困難である．

こうしたデータやクラスモデルの有限性の問題を少しでも軽減するため，多くのパターン認識の学習においては，利用可能な手持ちのデータを分割して有効利用することが行われている．分割には，Leave-One-Out法やBootstrap法などの様々な方法があるが，ここではその基本である，学習用と評価用への2分割法と，学習用と検証用，評価用への3分割法を考える．2分割法の場合，分類器は学習用標本をできる限り正確に分類できるように学習される．学習データが少ないとき，学習は，いわゆる過学習状態に陥りやすく，学習用標本には高い分類精度を達成するものの，評価用標本に対しは低い精度に留まることが多い．即ち，未知標本耐性が低い状態に陥る．2分割法においてこの問題を軽減するためには，少しでも学習用標本を増やすしかない．しかしそのとき，当然のことながら評価用標本は減少し，未知標本を代行するだけの安定な評価は難しくなる．学習用データへの過学習を回避するため，評価用標本を模するデータとして検証用標本を準備することが考えられる．これが3分割法である．学習用標本への過剰適応を抑えるため，学習に伴うハイパーパラメータ（例えばSupport Vector Machine（SVM）におけるカーネル幅など）の調整を，検証用標本を用いて行う．検証用標本が評価用標本を適切に模しているのであれば，この調整は有効なように思われる．しかし，分割数が増えることから明らかに，それぞれのデータセット中の標本数は少なくなり，学習用標本への過学習が生じやすくなり，また検証用標本を用いるハイパーパラメータの調整の効果も評価用標本を用いる評価の信頼性も低下し易くなる．こうして考えてみると，有効なように見受けられる検証用データの確保は必ずしも望ましいものではなく，不可避に違いない学習用標本と評価用標本の2分割にとどめ，しかもできるだけ多くの学習用標本を確保することに務めるべきであることがわかる．

過学習の回避を目指すとき，手持ちデータの分割とは異なるもう一つの重要なアプローチとして，学習対象であるパラメータが学習用データに過剰適応しないようにする正則化法も広く用いられてきた．伝統的な様々な正則化にも，比較的新しいSVM法のような手法においても，損失関数のような最小化の直接的な対象となるもの以外の要素を最適化の枠組みに取り込み，クラスモデル等の学習対象パラメータが学習データに適応し過ぎることの回避を目指す．言い換えれば，学習対象パラメータに制約を与えることで過学習を避けるアプローチととらえることもできる．

一方，損失関数の最小化と大きな幾何マージンの確保の双方を目指す学習法が注目されるようになってきた．ヒンジ損失の最小化を行いつつ線形識別関数による幾何マージンの最大化を目指すSVM法は，この手法の代表であろう．また，著者らによって提案された大幾何マージン型最小分類誤り（LGM-MCE: Large Geometric Margin Minimum Classification Error）学習法もまた，この手法の一つである．ここでは，分類誤り数損失の最小化と同時に誤分類尺度にも対応づけられる幾何マージンの増加が目指される．幾何マージンは，推定クラス境界とその最近傍パターン標本とのユークリッド距離であり，それが大きいということは，未知標本がその最近傍の学習標本とは多少ずれた位置に出現しても学習標本と同様の分類が可能であることを意味する．従って，これらの学習法が真に幾何マージンの増大を実現するのであれば，その学習は過学習状態に陥ることはなく（あるいは稀で），前述したような検証用標本の確保や，正則化によって学習対象パラメータの表現力に制限をかける必要もなくなることが期待される．

以上で総括したように，過学習の回避には様々な要因が複合的に関係していることが推察される．過学習の悪影響を適切に低減して未知標本に対する高い分類精度を達成するためには，まずこれらの種々の要因の関係あるいは影響を明らかにすることが大切である．本稿は，このような問題意識に基づいて，著者らが提案したLGM-MCE法を中心とした識別学習法を用いて，手持ちデータの分割が学習結果に及ぼす影響や，その学習における検証用標本の役割，幾何マージン増大化の役割等を実験的に調査するものである．

2. 調査に用いる学習法

2.1 最小分類誤り学習法

最小分類誤り（MCE; Minimum Classification Error）学習法は，分類の誤り程度を示す誤分類尺度から，平滑化した0-1損失（平滑化分類誤り数）を求め，それを最小化することを目指す．入力標本 $\mathbf{x}_n$ $(n=1,2,\cdots,N)$ が与えられた時，$\mathbf{x}_n$ が所属するクラス C_j $(j=1,2,\cdots,J)$ を判定する分類問題を考える．この時，$\mathbf{x}_n$ がクラス C_i に

† 同志社大学　Doshisha University
‡ 情報通信研究機構　NICT

所属する帰属度を表す識別関数 $g_i(\mathbf{x}_n;\Lambda)$ $(i=1,2,\cdots,J)$ 中で最も大きな値を示すクラスを $\mathbf{x}_n$ の所属クラスとする．そのとき，分類に用いる決定側は以下の式であるものとする．

$$C(\mathbf{x}_n)=C_j \qquad \text{iff} \qquad j=\arg\max_i g_i(\mathbf{x}_n;\Lambda) \qquad \cdots\cdots (2\cdot1)$$

Λ は分類器パラメータであり，学習によって決定される．

続いて $\mathbf{x}_n$ が C_y に所属していると判断した時の誤り程度を表す誤分類尺度を

$$d_y(x_n;\Lambda)=-g_y(x_n;\Lambda)+\log\left[\frac{1}{J-1}\sum_{i,i\neq y}\exp\left(\eta g_i(x_n;\Lambda)\right)\right]^{\frac{1}{\eta}} \qquad \cdots\cdots (2\cdot2)$$

のように，正解クラスである C_y に対する識別関数と不正解クラスである C_j に対する識別関数とで定義する．さらに，式（2・2）の正の定数 η を ∞ として，誤分類尺度の式は以下のように簡略化することができる．

$$d_y(\mathbf{x}_n;\Lambda)=-g_y(\mathbf{x}_n;\Lambda)+\max_{i,i\neq y} g_y(\mathbf{x}_n;\Lambda) \qquad \cdots\cdots (2\cdot3)$$

この式からわかるように，誤分類尺度は，$d_y(\mathbf{x}_n;\Lambda)>0$ の時に誤分類を，$d_y(\mathbf{x}_n;\Lambda)<0$ の時に正分類を表していることがわかる．

MCE学習は，さらにこの $d_y(\mathbf{x}_n;\Lambda)$ を用いて損失を定義する．様々な選択肢がある中で，多くの場合，平滑化分類誤り数損失は，以下のようにシグモイド関数で定義する．

$$\ell(\mathbf{x}_n;\Lambda)=\frac{1}{1+\exp\{-\alpha d_y(\mathbf{x}_n;\Lambda)\}} \qquad \cdots\cdots (2\cdot4)$$

式（2・3）の誤分類尺度はその値の正負により分類が間違っていたかどうかを表すため，これに0-1損失関数を適用することで，MCE法は損失を得ることができる．また，損失関数として式（2・4）のシグモイド関数を扱う理由としてはMCE法では分類誤り数を最小化するという本質から，誤分類をした場合に1，正分類をした場合に0，というように損失を得ることができる0-1関数を扱うことを前提としており，加えて一般的にMCE法では勾配法による分類器パラメータの最適化を考えるため，0-1関数を微分可能な関数で近似できるシグモイド関数を損失関数としている．なお，α の値は関数の傾きを表し，値を小さくすれば損失はより平滑になり，値を大きくすれば損失は0-1関数に近づく．

そして，全標本に対する損失を平均した経験的平均損失

$$L(\Lambda)=\frac{1}{N}\sum_{n=1}^{N}\ell(\mathbf{x}_n;\Lambda) \qquad \cdots\cdots (2\cdot5)$$

を最小化することで，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 Λ の状態の実現を目指す．また，識別関数 $g_y(\mathbf{x}_n;\Lambda)$ は Λ に関して微分可能であるので，式（2・2）と式（2・4），さらに前述したように式（2・5）は Λ に関して微分可能である．したがって，MCE法では勾配法による最適化が可能となり，パラメータの更新は以下のように実行される．

$$\Lambda^{(t+1)}=\Lambda^{(t)}-\varepsilon_t\frac{1}{N}\sum_{n=1}^{N}\nabla_\Lambda\ell(d_{yn}(\mathbf{x};\Lambda^{(t)})) \qquad (\varepsilon_t>0) \qquad \cdots\cdots (2\cdot6)$$

のような式で表すことができ，この操作を繰り返すことで Λ を求めることができる．ここでの t は繰り返し番号を表し，学習の繰り返しにおけるステップ回数を意味する．また，∇_Λ は Λ の偏微分を示し，$\Lambda^{(t)}$ は t 回目の学習によって求められたパラメータであり，初期値 $\Lambda(0)$ は初期化されているものとする．ε_t は学習係数である．

見て分かるように（2・6）式は全ての学習標本を得た状態で行うバッチ的手法である．これに対して確率的降下法によって学習標本を得るごとに Λ を逐次的に調節することもできる．その更新式は以下のようになる．

$$\Lambda^{(t+1)}=\Lambda^{(t)}-\varepsilon_t\nabla_\Lambda\ell(d_{yn}(\mathbf{x};\Lambda^{(t)})) \qquad (\varepsilon_t>0) \qquad \cdots\cdots (2\cdot7)$$

ちなみに，ε_t が以下の2つの条件を満たしている時，経験的平均損失 $L(\Lambda)$ が確率的には局所的最小解に達し得ることが証明されている．

$$\sum_{t=0}^{\infty}\varepsilon_t=\infty \qquad \cdots\cdots (2\cdot8)$$

$$\sum_{t=0}^{\infty}\varepsilon_t{}^2<\infty \qquad \cdots\cdots (2\cdot9)$$

2.2 大幾何マージン最小分類誤り学習法

前節で説明したように，MCE法では学習によって最小分類誤り状態を目指すのが狙いであった．また，誤分類尺度 $d_y(\mathbf{x}_n;\Lambda)$ が負値の時（正分類の時）の絶対値 $\left|d_y(\mathbf{x}_n;\Lambda)\right|$ は分類の確かさを表し，その大きさは分類の節目に対する余裕，すなわちマージンを意味し，学習によって未知標本に対する耐性を向上すると考えられていた．しかし，例えば（2・3）式からもわかるように，全ての識別関数を定数倍するだけでもこの誤分類尺度は大きくなる．この定数倍が分類境界を変えないことは明ら

かであり，こうした誤分類尺度は必ずしも分類の正しさの程度を表すものではないことがわかる．

この問題を解決するために大幾何マージン最小分類誤りLGM-MCE学習法が提案された．LGM-MCE法は誤分類尺度を境界と標本との幾何学的距離（ユークリッド距離），すなわち幾何マージンとして表現することにより，最小分類誤り状態を目指すとともに直接的に幾何マージンの増大をも目指す．結果的に，この学習による未知標本耐性の向上，あるいは未知標本に対する高い分類精度の実現が期待されることになる．

前小節と同様の分類問題を考える．この時，幾何マージン r は先に定義した誤分類尺度の絶対値をその勾配のノルムで正規化したもので表すことができる．

$$r \approx \frac{-d_y(\mathbf{x}_n;\Lambda)}{\left\|\nabla d_y(\mathbf{x}_n;\Lambda)\right\|} \qquad \cdots\cdots (2\cdot 10)$$

ちなみに $\left\|\nabla d_y(\mathbf{x}_n;\Lambda)\right\|$ は $d_y(\mathbf{x}_n;\Lambda)$ に対する $\mathbf{x}_n$ に関する微分値のノルムである．そして，この幾何マージン r の正負反転に対する以下の $D_y(\mathbf{x}_n;\Lambda)$ をLGM-MCE学習法での新たな誤分類尺度とし，以下のように示す．

$$D_y(\mathbf{x}_n;\Lambda) = \frac{d_y(\mathbf{x}_n;\Lambda)}{\left\|\nabla d_y(\mathbf{x}_n;\Lambda)\right\|} \approx -r \qquad \cdots\cdots (2\cdot 11)$$

なお，混乱を防ぐため，先に紹介したMCE学習法は，以下では関数マージン最小分類誤り（FM-MCE: Functional Margin Minimum Classification Error）学習法と呼ぶこととする．

2.3 多クラスサポートベクタマシン

本稿では，広く使われているSVM法に代えて，多クラス分類を直接扱うことができるMulti-class Support Vector Machine（MSVM）法を用いる．SVM法と同様に，通常，MSVM法の識別関数は以下のように非線形写像関数 $\phi(\cdot)$ を伴う形で用いられる．

$$g_j(\mathbf{x}_n;\Lambda) = \mathbf{w}_j^T \phi(\mathbf{x}_n) \qquad (j = 1,\ldots J) \qquad \cdots\cdots (2\cdot 12)$$

ここで，$\{\mathbf{w}_j\}_{j=1}^{J}$ はクラス C_j に対する重みベクトルであり，Λ に対応する．また T は転置を表す．

MSVM の学習も，MCE 学習法あるいは LGM-MCE 学習法と同様に，学習標本に対する損失の最小化を通して，その最小状態に対応する Λ の状態を求める．

MSVM 法における幾何マージンは以下の右項のように重みベクトルのノルムの逆数で与えられ，学習はその最大化を目指す．

$$\max_{\{\mathbf{w}_j\}} : \quad \frac{1}{\sum_{j=1}^{J}\left\|\mathbf{w}_j\right\|^2} \qquad \cdots\cdots (2\cdot 13)$$

またさらに，下の図のように（スラック変数 $\{\xi_n\}_{n=1}^{n=N}$ によるシフトを伴う）ヒンジ型損失関数を導入し，先の幾何マージンの最大化と，このヒンジ型損失の最小化を同時に目指す．

結果的に，この同時最適化は次に示す不等式制約条件付き最小化問題として定式化される．

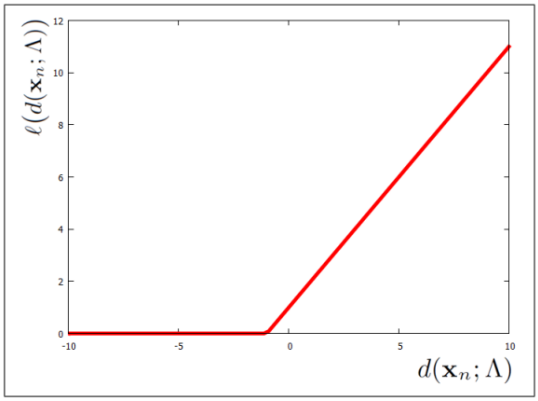

図１　ヒンジ型損失関数．

$$\min_{\{\mathbf{w}_j\},\{\xi_n\}} : \quad \sum_{n=1}^{N}\xi_n + \frac{\beta}{2}\sum_{j=1}^{J}\left\|\mathbf{w}_j\right\|^2 \qquad (\beta > 0)$$
$$\text{subject to}: \quad \forall_n : \xi_n \geq \max\{0, 1 - d(\mathbf{x}_n,\Lambda)\} \qquad \cdots\cdots (2\cdot 14)$$

ここで β は幾何マージンの重視度を表すパラメータである．式（2・14）の制約条件付き最適化問題は，以上のままでは必ずしも容易に解くことができるとは言えない．そこで MSVM 法は，まずラグランジュの未定乗数法を用いて最適化問題を以下のように置き換える．

$$\min_{\{\mathbf{w}_j\},\{\xi_n\},\{\eta_{n,j}\}} : \quad L = \sum_{n=1}^{N}\xi_n + \frac{\beta}{2}\sum_{j=1}^{J}\left\|\mathbf{w}_j\right\|^2 + \sum_{n=1}^{N}\sum_{j=1}^{J}\eta_{n,j}\{g_j(\mathbf{x}_n;\Lambda) - g_{y_n}(\mathbf{x}_n;\Lambda) - \xi_n - \delta_{y_{n,j}}\}$$
$$\text{subject to}: \quad \forall_n, \forall_j : \eta_{n,j} \geq 0 \qquad \cdots\cdots (2\cdot 15)$$

$\{\eta_{n,j}\}_{n=1\,j=1}^{N\;J}$ はラグランジュ乗数であり，$\delta_{i,j}$ は $i = j$ の時に１，それ以外の時は０になる，クロネッカーのデルタ関数である．さらに KTT 条件により，主関数 $\{\mathbf{w}_j\}$，$\{\xi_n\}$ に関して最小化し，そしてラグランジュ乗数 $\{\eta_{n,j}\}$ に関して最大となるような鞍点の探索により，$\{\eta_{n,j}\}_{n=1\,j=1}^{N\;J}$ のみの関数 D を得ることができる．そして以下の $\{\eta_{n,j}\}_{n=1\,j=1}^{N\;J}$ に関する制約条件付き最大化問題（双対問題）に帰着することとなる．

$$\max_{\{\eta_{n,j}\}}: \quad D=-\frac{\beta^{-1}}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) \times\sum_{j=1}^{J}(\delta_{y_n,j}-\eta_{n,j})(\delta_{y_m,j}-\eta_{m,j})-\sum_{n=1}^{N}\boldsymbol{\eta}_n^T\mathbf{1}_{y_n}$$

$$\text{subject to}: \quad \forall_n:\boldsymbol{\eta}_n\geq\mathbf{0} \qquad \text{and} \qquad \boldsymbol{\eta}_n^T\mathbf{1}=1 \qquad \cdots\cdots (2\cdot16)$$

$\{\eta_n\}_{n=1}^N$ は j 番目の要素が $\eta_{n,j}$ であるベクトルであり，$\mathbf{1}_j$ は j 番目の成分だけが 1 となったベクトルである．また，$\mathbf{1}$ は全ての要素が 1 であるベクトルで，$\mathbf{0}$ は全ての要素が 0 であるベクトルである．さらに J 次元ベクトル $\{\boldsymbol{\tau}_n\}_{n=1}^N$ を

$$\tau_{n,j}=\delta_{y_n,j}-\eta_{n,j} \quad (n=1,...,N;\ j=1,...,J)$$
$$\boldsymbol{\tau}_n=\left[\tau_{n,1},...,\tau_{n,J}\right]^T \quad (n=1,...,N)$$

のように，新たなベクトルとして定義する．すると，$\{\mathbf{w}_j\}$ と D は式変形により

$$\mathbf{w}_j=\beta^{-1}\sum_{n=1}^{N}\tau_{n,j}\phi(\mathbf{x}_n) \quad (j=1,...,J)$$

$$\max_{\{\tau_n\}}: \quad D=-\frac{\beta^{-1}}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m)\boldsymbol{\tau}_n^T\boldsymbol{\tau}_m-\sum_{n=1}^{N}\boldsymbol{\tau}_n^T\mathbf{1}_{y_n}$$
$$\text{subject to}: \quad \forall_n:\boldsymbol{\tau}_n\leq\mathbf{1}_{y_n} \qquad \text{and} \qquad \boldsymbol{\tau}_n^T\mathbf{1}=0$$

と表すことができ，最終的に，識別関数は以下の関数に置き換えることができる．

$$g_j(\mathbf{x}_n;\Lambda)=\beta^{-1}\sum_{m=1}^{N}\tau_{m,j}\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) \quad (j=1,...,J) \qquad \cdots\cdots (2\cdot17)$$

ここで，$\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m)$ は高次元空間における内積を表し，$K(\mathbf{x}_n,\mathbf{x}_m)$ として正定値カーネルを用いることにより，この内積は現実的に計算可能となり，は以下のようになる

$$g_j(\mathbf{x}_n;\Lambda)=\beta^{-1}\sum_{m=1}^{N}\tau_{m,j}K(\mathbf{x}_n,\mathbf{x}_m)$$

$$\max_{\{\tau_n\}}: D=-\frac{\beta^{-1}}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}K(\mathbf{x}_n,\mathbf{x}_m)\boldsymbol{\tau}_n^T\boldsymbol{\tau}_m-\sum_{n=1}^{N}\boldsymbol{\tau}_n^T\mathbf{1}_{y_n}$$
$$\text{subject to}: \forall_n:\boldsymbol{\tau}_n\leq\mathbf{1}_{y_n} \qquad \text{and} \qquad \boldsymbol{\tau}_n^T\mathbf{1}=0 \qquad \cdots\cdots (2\cdot18)$$

以上のように，$\{\eta_{n,j}\}_{n=1\,j=1}^{N\ \ J}$ はカーネル関数にかかる重みとなり，MSVM の学習では式（2・18）の最大化を目指して，$\{\tau_{n,j}\}$ の更新を行う．学習の結果，$\boldsymbol{\tau}_i$ の全ての要素が 0 となった時は，$\mathbf{x}_i$ に関するカーネル関数 $K(\mathbf{x}_n,\mathbf{x}_i)$ は全く使われなくなる．従って，$\mathbf{x}_i$ は分類器モデルとして記憶する必要性が無くなる．また逆に，$\boldsymbol{\tau}_i$ が 0 でない要素が一つでもある学習標本 $\mathbf{x}_i$ はサポートベクタとして記憶され，識別関数の計算に用いられる．このサポートベクタの数が MSVM 型分類器のサイズを表す．

3. 実験と考察

3.1 実験条件

扱った分類用標本データセットはUCI Machine Learning Repositoryが提供するAbaloneデータセットである．データ数は4177個で，手持ちの（即ち学習や検証に用いることができる）データを2611個とし，未知標本模す役割を持つ評価用標本を1566個とした．2種類の実験は行った．1つ目は手持ちのデータを学習用標本と検証用標本に分割し，もう1つは手持ちのデータを全て学習用標本として行った．

評価には，幾何マージン増大型学習法であるLGM-MCE学習法とSVM法，そして非幾何マージン増大型学習であるFM-MCE法を用いた．ただし，検証用標本を用いない実験にはSVM法の評価実験を行わなかった．これはSVM法が，未知標本耐性を制御するためには，非線形写像関数に伴うカーネルの幅を制御するために不可避的に検証用標本を必要とするためである．

3.2 実験

3.2.1 実験１：学習用標本と検証用標本を用いる学習

ここでの実験は手持ちのデータ2611個を学習用標本であるTraining Dataと検証用標本であるValidation Dataを以下の6パターンに分割して実験を行った．また，FM-MCE法とLGM-MCE法については，いずれもマルチプロトタイプ型の識別関数を用い，各ハイパーパラメータは以下のように設定した．α は5ずつ，μ_0 は10倍ずつ増やしていく．プロトタイプの個数は1，3，5，7，10の値で実験を行った．MSVMについてはそのパラメータを表のように設定し，σ は α と同じく5ずつ，β は μ_0 と同じく10倍ずつ増やした．また，ε は0.00001に固定した．そして，全ての学習法において学習回数は10000回とした．

Training Data	Validation Data
433	2178
869	1742
1303	1308
1739	872
2175	436

表1　学習用標本と検証用標本の分割表.

	α	μ_0
調整範囲	0.0001－1.0	0.0001－1.0

表2　MCE法とLGM-MCE法におけるハイパーパラメータの調整範囲.

	β	σ
調整範囲	0.0001－1.0	1－100

表3　MSVM法におけるハイパーパラメータの調査範囲.

FM-MCE法とLGM-MCE法で用いた識別関数は次式のプロトタイプと標本との負のユークリッド距離を用いた.

$$g_j(\mathbf{x}_n;\Lambda) = -\left\|\mathbf{x}_n - \mathbf{p}_j\right\|^2$$

また，MSVMの識別関数は次式のガウシアンカーネルに基づく識別関数を用いた.

$$g_j(\mathbf{x}_n;\Lambda) = \sum_{n=1}^{N} \tau_{j,n} K(\mathbf{x},\mathbf{x}_n)$$

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\left\|\mathbf{x}-\mathbf{x}'\right\|^2}{\sigma}\right)$$

（σ の値は分散を表す）

実験結果を以下に示す．図2が検証用標本の中で最適なハイパーパラメータ設定した分類器のテスト標本に対する認識率比較である．いずれの手法においても，ばらつきはあるものの，基本的に学習用標本が多いほど，認識率は高くなっていることが分かる.

続いて図3を見てみる．こちらは学習用標本の中で最適なパラメータを設定した分類器のテスト標本に対する認識率の比較である．よって，分類器の設計において検証用標本は全く使用していないことに注意してほしい．先ほどと同様に，各手法でみても，学習用標本の数が多くなるにつれ認識率は高くなっている．ここで，幾何マージン増大型学習法のLGM-MCE法の結果を見てみると，学習回数の増加に伴い，認識率が単調に増加していることが分かる．このことから，検証用標本を用いず，手持ちのデータを全て学習用標本として使えばさらに高い効果を得られるのではないか，という期待が生じる.

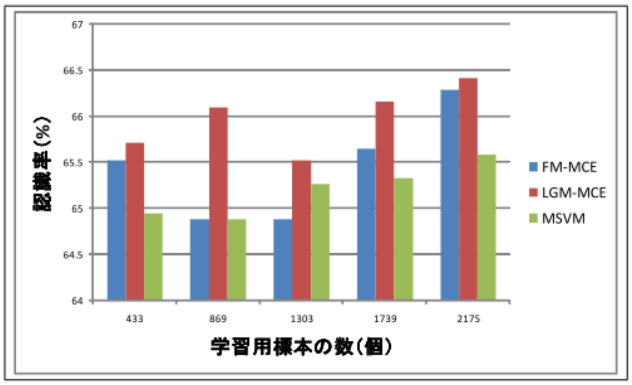

図2　検証用標本で最も高いハイパーパラメータを用いた評価用標本に対する認識率.

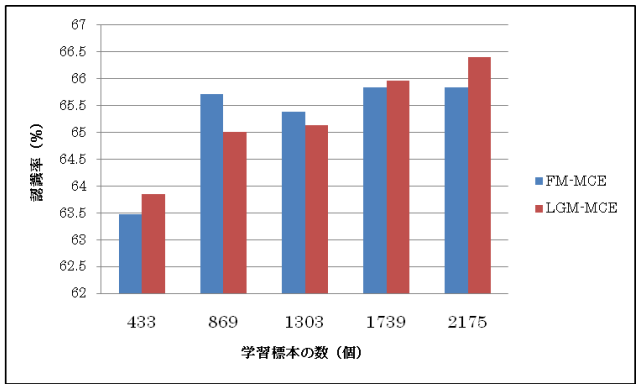

図3　学習用標本で最も高いハイパーパラメータを用いた評価用標本に対する認識率.

3.2.2　実験２：学習用標本のみを用いる学習

ここでは手持ちのデータ2611個を全て学習用データとして，実験を行った．ただし前述した通り，学習において検証用標本が必要となる性質を持つSVM法では本実験は行わず，ここではLGM-MCE法とFM-MCE法のみの実験結果となっている．学習における更新回数は各手法とも10000回，20000回，30000回の3種を設定して比較した．また，FM-MCE法，LGM-MCE法において α と μ_0 は表のような調査範囲とした．表からわかるように両手法ともの調査範囲は同じである.

本実験でのプロトタイプ数は，予備実験の結果十分に高い表現能力を持つ30とした．これは，プロトタイプ数が少ない，言い換えれば学習対象パラメータが少ない場合，クラス境界を表現する能力が低下し，結果的に学習の学習標本への過適応が抑制，言い換えれば未知標本耐性が向上されることになり，幾何マージン増大効果による未知標本耐性の向上なのか，あるいはこの低い表現力の副次的効果なのかの区別がつかなくなることを回避するためである．実際に前の小節において，幾何マージンを増大するLGM-MCE法がFM-MCE法よりも高い分類精度を達成してはいるが，それが分類器の表現能力が低いが故のことであるという可能性も考えられるため，正確に議論ができなかったという背景があった．さらに，分類器の表現能力を十分に引き出すために，学習回数（パラメータの更新回数）も十分に大きくとり，最大エポック数を10000回と20000回，30000回の3つのケースを試してみた.

以下が学習用標本の中で最適なハイパーパラメータで設計した分類器における学習用標本に対しての認識率（closed）と評価用標本（open）に対する認識率である.

	closed	open
FM-MCE	**81.50134**	63.85696
LGM-MCE	81.30984	**64.30396**

表2　FM-MCE法とLGM-MCE法における学習回数10000回の結果.

	closed	open
FM-MCE	**81.92264**	**64.55939**
LGM-MCE	81.38644	62.38825

表3　FM-MCE法とLGM-MCE法における学習回数20000回の結果.

	closed	open
FM-MCE	**81.99923**	63.7931
LGM-MCE	81.65454	**63.98467**

表4　FM-MCE法とLGM-MCE法における学習回数30000回の結果.

3つの表からわかるように，closedの認識率はFM-MCE法の方が高く，openの認識率はLGM-MCE法の方が高い傾向にある．これは学習用標本に対して，最小分類誤り状態を目指せる代わりに，学習回数に伴い学習用標本に過剰に適合してしまうFM-MCE法の性質と最小分類誤り状態を目指すと共に幾何マージンを増大させ過学習を防ぐLGM-MCE法の性質の違いから現れる結果であると考えられる．

では実際にFM-MCE法とLGM-MCE法の幾何マージンのグラフを見ていく．以下に示すのが学習用標本に対する学習内での各学習回数における幾何マージンの値である．

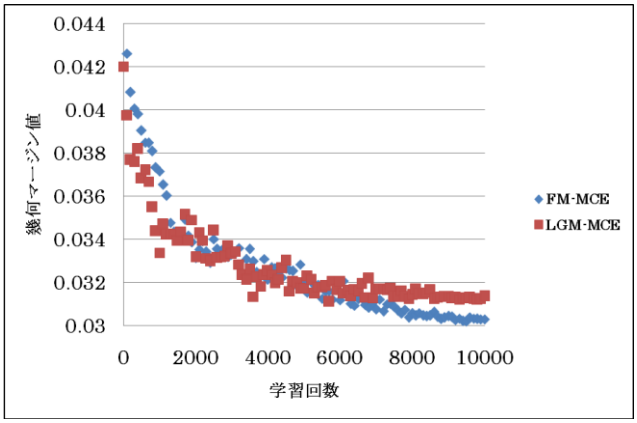

図3　学習回数10000回の学習での各学習回数における幾何マージンの値.

学習回数が10000回の時点でLGM-MCE法の幾何マージンの値がFM-MCE法よりも高くなっていることから，LGM-MCE法が幾何マージンを増大するよう学習したことが分かり，先ほど述べた傾向が幾何マージンの増大によるものであることが再確認できた．学習回数が0回から4000回辺りまで見てみると両手法とも幾何マージンの値が小さくなっていることが分かる．幾何マージンの概念が組み込まれていないFM-MCE法はともかく幾何マージンを増大させるLGM-MCE法でなぜこのようなことが起こるのか．それはLGM-MCE法が学習の初期段階では幾何マージンを増大させるよりも最小分類誤り状態を目指す働きが強いという性質があるからである．以下のグラフを見てみる．

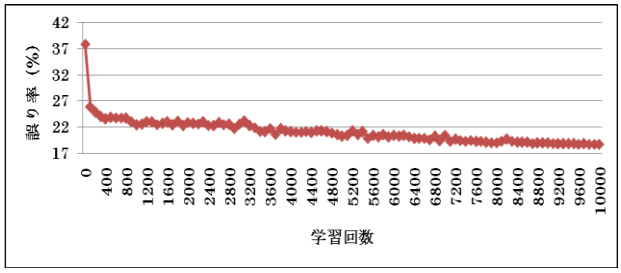

図4　LGM-MCE法における各学習回数においての学習用標本に対しての誤り率.

学習回数が1200回までは誤り率が単調に減少していることがわかる．これは最小分類誤り状態を目指している効果である．その反面，先ほどの幾何マージンの値の表のように学習回数が1000回辺りまでの幾何マージンの急激な減少が見られる．そして，学習がある程度収束すると，幾何マージンを増加させる挙動が見られるようになる．また図4の学習回数が2000回辺りのグラフを拡大した図5を見れば，同じような傾向があることが確認できる．

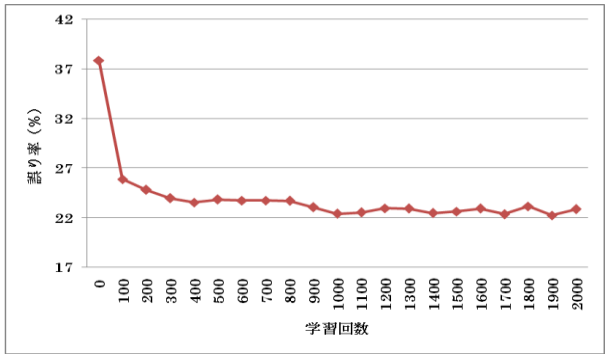

図5　図4における学習回数2000回までの拡大図.

以上のようにLGM-MCE法は最小分類誤り状態を目指すと共に幾何マージンを増大させるが，同時に実現するのではなく，学習の段階に応じて優先が変わるという事が分かった．

ちなみに学習回数が20000回と30000回の時の幾何マージン値の比較を以下の図に示す．先ほどの表と照らし合わせて見ると学習回数が30000回のものは10000回のものと全く同じ傾向が現れていることがわかる．一方で，学習回数が20000回の時のものはFM-MCE法の幾何マージンの値がLGM-MCE法よりも大きくなっている．この要因としては，表からわかるようにFM-MCE法の方がopenの結果が良い（LGM-MCE法と比べて，closedの結果も良いものの差は小さい）ことから，未知標本耐性が向上した，すなわち幾何マージンが増大したことがうかがえる．これは，幾何マージンの増大を保証はしないものの，増大させる可能性があるというFM-MCE法の性質が示されたものと考えられる．しかしながら，openの結果が2%以上の差があるにも関わらず，図より幾何マージンの値の差はほとんど

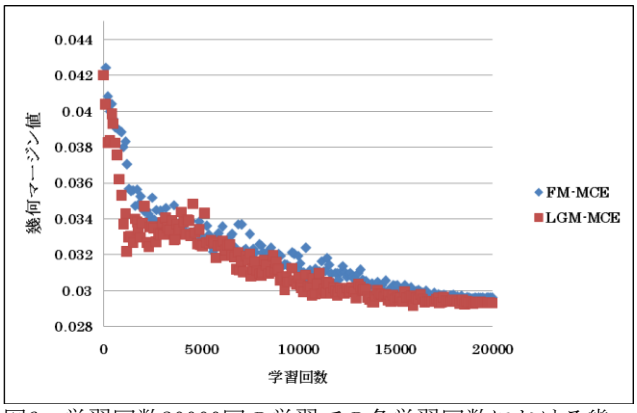

図6　学習回数20000回の学習での各学習回数における幾何マージンの値.

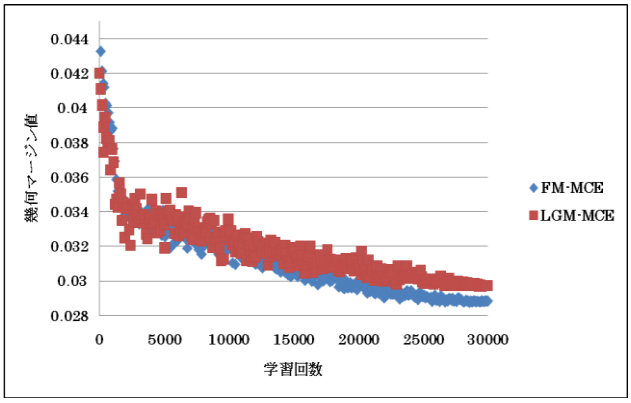

図7　学習回数30000回の学習での各学習回数における幾何マージンの値.

無いと言える．これは，結果的に分類精度では劣ったものの，LGM-MCE学習法がここでも理論通りに幾何マージンの増大に関する学習は行っていたことを示唆する．またこれは，先ほどのFM-MCE法が幾何マージンを増大しないとは限らないという事を示す裏付けともなる．学習回数が20000回と30000回における各学習回数の認識率は10000回のものと全く同様の結果が出たためここでは割愛する．

ここで，LGM-MCE法の3.2.1節の結果と3.2.2節の結果を比較する．ただし，比較前に確認すべきことがある．それは今回の3.2.1と3.2.2の実験の条件の違いについて確認したい．3.2.1はプロトタイプ数を1，3，5，7，10と設定し，検証用標本での認識率を指標とし，その中から最適なモデルを選択している．一方で，3.2.2はプロトタイプ数が30のみであり，モデルの最適化が行われていない．さらに今回はLGM-MCE法とFM-MCE法の性質の違いを明確にする目的を含んでいたため，分類器の表現能力を従来よりも十分大きいと思われる値に設定した．このことから，幾何マージン増大型学習法といえども，分類器自体の表現能力が高すぎるが故に過学習が起きてしまう可能性もあり，両者の正確な比較をするのは難しいと言える．ただし，本稿の動機となった学習に内在する種々の要因が交錯する可能性を示唆するのに重要な要素と考え，以上のことを踏まえたうえで比較を行う．

以下の表は3.2.1節で行った，LGM-MCE法における学習用標本数毎の認識率（青）と本節で行った検証用標本を用いない実験での認識率（赤）である．3.2.1節の結果は

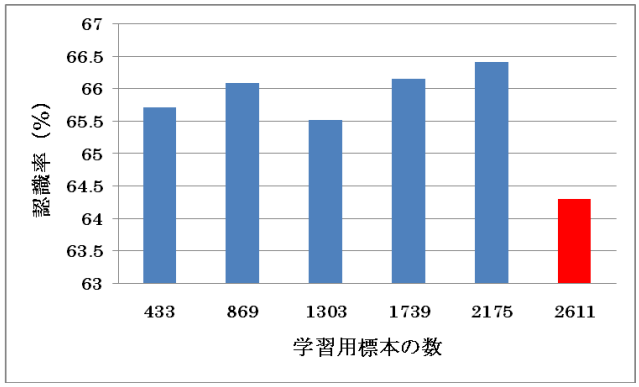

図8　LGM-MCE法における検証用標本を用いた学習用標本数毎の認識率（青）と検証用標本を用いない認識率（赤）の比較（認識率は評価用標本に対するものである）

検証用標本の中で最適なハイパーパラメータで設計した分類器における評価用標本に対する認識率，本節の結果は学習用標本の中で最適なハイパーパラメータで設計した分類器における評価用標本に対する認識率である．表を見てみると，本節の結果は3.2.1節の結果の，どの学習用標本数に対しても認識率が下回ることが確認できる．これは，先ほど述べたように今回はまだ正当に比較できる段階では無く，かつ3.2.2はモデルの最適化をおこなっておらず，また分類器の表現能力が高いことにより過学習を引き起こす可能性があったため，想定通りの結果と言える．以下に各学習標本の数に対するプロトタイプ数の表を示す.

学習用標本数	433	869	1303	1739	2175	2611
プロタイプ数	7	7	3	7	9	30

表5　図8の結果における各学習用標本数に対するプロトタイプ数の数.

見て分かるように検証用標本を用いている，433から2175までは検証用標本による過学習を抑える作用が働いたためプロタイプ数が少なくなっている．以上のように今回の比較はまだ十分なものであるとは言い難いが3.2.1の学習用標本で最適なハイパーパラメータを設定した場合のように，学習用標本が多いほど学習の効果が期待できることも示唆されている.

3.3　考察

今回行った実験全体の考察を述べる．まず，3.2.1節の実験ではどの手法においても検証用標本の数が少ないほど，すなわち，学習用標本の数が多いほど評価用標本に対する認識率が高いことが分かった．これは，学習に用いる数が多いほど，学習における効果が大きくなるという事実に基づいていると考えることができる.

次に3.2.2節の結果についてである．ここではFM-MCE法とLGM-MCE法の比較を行った．結果としてはLGM-MCE法における学習用標本に対する認識率はFM-MCE法のそれよりも低く，またテスト標本に対する認識率は逆になる結果があった．これは，実際に幾何マージンの値で見たよう

にLGM-MCE法が幾何マージン値をFM-MCE法よりも増大させていることから，未知標本耐性を向上させていたが故に，以上のような結果になったと考えられる．このことから，分類器の表現能力が非常に高い場合においても，LGM-MCE法は過学習を防ぎ，未知標本耐性を向上させることができるという事が確認できた．

5. おわりに

有限の手持ちのデータを用いて高い未知標本耐性を達成しなければならないという極めて困難かつ本質的な問題に対して，幾何マージン増大型という魅力ある学習法が登場した．もしこの幾何マージン増大の学習が，検証用標本の影響などから解放されて純粋に学習用標本のみを用いて効果的に実行され得るものであるならば，手持ちのデータは全て学習用に回し，未知標本に対してもより正確な分類を実現することができるように思われる．しかしながら，未知標本耐性の向上をもたらすように見受けられる要因には，学習対象パラメータそのものが持つ表現能力の大きさや，その表現能力を十分に発揮させるためのハイパーパラメータの適切な調整など，幾何マージンを直接的に増大させる学習メカニズムとは異なる様々なものが考えられる．本稿では，これらの種々の可能性を整理することを目指して，幾何マージンを大きくする背景を実験的に精査した．実験の結果，LGM-MCE学習法は，その幾何マージン増大の機構が有効に働き，学習対象パラメータの表現能力が適切に選ばれている場合でも，あるいはその能力が十分に準備されている場合でも，FM-MCE学習法と比べると安定的に幾何マージンを大きくし，結果的に高い未知標本耐性を達成できることが明らかとなった．また，不可避的に検証用標本を必要とするMSVM法とは対照的に，LGM-MCE学習法は学習用データのみを用いても未知標本耐性を高め得る可能性を把握することもできた．しかしその一方で，学習対象パラメータが持つ表現能力の制約は，実効的には未知標本耐性の向上に強く関連していることも示された．この表現能力そのものの制御は，正則化などのアプローチにおける古くからの重要な研究課題であるが，今後一層の調査研究が望まれるように思われる．

[謝辞] 本研究の一部は，平成26年度科学研究費助成事業・基盤研究 (B) 「高識別的特徴空間とその探索法の最小分類誤り基準に基づく統一的実現」に支援して行われたものである．

[参考文献]

渡辺　秀行　「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」

田中　秀明　「非線形カーネル写像を伴う分類器のための最小分類誤り学習法に関する研究」

石井　健一郎，上田修功，前田英作，村瀬　洋　「わかりやすいパターン認識」