情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

大幾何マージン最小分類誤り学習法

Large Geometric Margin Minimum Error Classification

渡辺 秀行* 片桐 滋† 山田 幸太† マクダーモット エリック‡
Hideyuki Watanabe Shigeru Katagiri Kouta Yamada Erik McDermott
中村 篤‡ 渡部 晋治‡ 大崎 美穂†
Atsushi Nakamura Shinji Watanabe Miho Ohsaki

Abstract: The recent dramatic growth of computation power has resulted in increased interest in discriminative training methods for pattern recognition. Minimum Classification Error (MCE) training is especially attracting a great deal of attention, and it can be used to achieve minimum-error classification of various types of patterns. However, for increasing the robustness of classification, the conventional MCE framework has no practical optimization procedures like the geometric margin maximization in Support Vector Machine (SVM). To realize high robustness in a wide range of classification tasks, we derive the geometric margin for a general class of discriminant functions and develop a new MCE training method that increases the geometric margin value. We demonstrate the effectiveness of the new method by experiments using prototype-based classifiers and clarify relationships between the new method and such existing methods as SVM.

Keywords: Minimum Classification Error, MCE, margin, geometric margin, robustness

1 まえがき

統計的パターン認識における究極の分類器設計目標は，すべてのパターンに対する最小分類誤り確率状態の達成，すなわちベイズ誤りの推定である [1, 2, 3]．この実現を目指し，識別学習の研究が盛んに行われ [4, 5, 6, 7, 8]，識別学習は今やパターン認識研究の主流となるに至った．

中でも，音声などの時系列信号処理の分野で発展した最小分類誤り（MCE）学習 [4, 5] が，高い分類率を実現する学習法として幅広い分野で利用されている．MCE は，可変長パターンをも含む多様なパターンに対するベイズ誤り推定を直接的に追求する学習法であり，広範な判別関数に適用可能である．しかし分類頑健性の向上（汎化能力向上）に対しては，従来の MCE 学習は具体的な最適化手続きを持たず，損失関数や判別関数の平滑度の制御 [5, 9] という間接的で不十分な制御機構に頼っている．

頑健性を直接的に表現する測度として，分類決定境界とそれに最も近い学習標本との間の距離，すなわち幾何マージンが挙げられる [7]．実際，主に固定次元ベクトルの標準的な分類手法として知られるサポート・ベクター・マシン（SVM）[6, 7] は，この幾何マージンの最大化により頑健性を向上させている．しかし，SVM では幾何マージンの導出が線形判別関数に限られており，しかも SVM は，特に線形分離不可能という現実的な状況において，最小化標的となる損失関数などがベイズ誤り推定との一貫性を持たず，本来目指すべき分類誤り最小化の意味での学習の最適性が保証されない．

本稿では，線形判別関数に限定されない判別関数の一般形に対する幾何マージンを定式化し，この一般的な幾何マージンを増大させながら MCE 学習を行う新たなアプローチを提案する．これにより，幅広い分類タスクにおいて，未知標本をも含むすべての標本を高精度に分類

*情報通信研究機構 MASTAR プロジェクト 音声コミュニケーショングループ, 〒 619-0289 京都府相楽郡精華町光台 3-5, tel. 0774-98-6309, e-mail hideyuki.watanabe@nict.go.jp
Spoken Language Communication Group, MASTAR Project, National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289 Japan
†同志社大学大学院 工学研究科 情報工学専攻, 〒 610-0394 京都府京田辺市多々羅都谷 1-3
Graduate School of Engineering, Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe City, Kyoto 610-0394 Japan
‡日本電信電話株式会社 NTT コミュニケーション科学基礎研究所, 〒 619-0237 京都府相楽郡精華町光台 2-4
NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan

できることが大いに期待される．更に，汎用性の高い分類器であるプロトタイプ型分類器を用いた実験により提案法の有効性を示すとともに，提案法と SVM などの従来法との理論的関係を明らかにする．

2　最小分類誤り学習

2.1　決定則を包含する定形化 [4]

簡単のために，固定次元ベクトル・パターンの空間 $\mathcal{X}$ に属するパターン $\boldsymbol{x}$ を J 個のクラス（$C_j; j = 1, ..., J$）の 1 つに分類する問題を考える．分類器は，調整可能なパラメータ集合 Λ から構成され，この調整のために，それぞれの帰属クラスが既知である N 個の学習標本の集合 $\Omega_N = \{\boldsymbol{x}_n\}_{n=1}^N$ が与えられているものとする．

このとき目指すべき設計目標は，クラス C_j に対する真の事後確率が $p_\Lambda(C_j|\boldsymbol{x})$ の形で表現され，最大事後確率値を示すクラスに分類するベイズ決定則を利用することを前提として，次式の分類誤り数リスク

$$R(\Lambda) = \sum_{y=1}^{J} \int_{\mathcal{X}} p_\Lambda(C_y, \boldsymbol{x}) \, 1\Big(p_\Lambda(C_y|\boldsymbol{x}) \neq \max_j p_\Lambda(C_j|\boldsymbol{x})\Big) \, \mathrm{d}\boldsymbol{x} \tag{1}$$

を最小化する Λ の推定となる [2]．ここで $1(\mathcal{P})$ は，命題 $\mathcal{P}$ が真ならば 1，偽ならば 0 を返す指示関数である．しかし実際には，真の事後確率の関数形式はほとんど知り得ることはなく，式 (1) の推定は現実的ではない．事後確率の関数形式が未知であっても遂行し得る分類誤り数リスクの推定方式が望まれる．

MCE 学習は，上記の要請に対する現実的な解を以下の関係式，

$$R(\Lambda) \simeq \sum_{y=1}^{J} \int_{\mathcal{X}} p(C_y, \boldsymbol{x}) \ell\big(d_y(\boldsymbol{x}, \Lambda)\big) \, \mathrm{d}\boldsymbol{x} \tag{2}$$

に基づいて提供する [4]．ここで，$d_y(\boldsymbol{x}, \Lambda)$ は誤分類測度とよばれ，事後確率 $p_\Lambda(C_j|\boldsymbol{x})$ に限らない一般の判別関数 $g_j(\boldsymbol{x}, \Lambda)$ から，次式により構成される（$\psi > 0$）．

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

なお $\boldsymbol{x} \in C_y$ とし，また最大の $g_j(\boldsymbol{x}, \Lambda)$ の値を与えるクラスに $\boldsymbol{x}$ を分類する決定則を考える．式 (3) は，分類判断の正誤をスカラー量の符号によって簡潔に表現した

$$d_y(\boldsymbol{x}, \Lambda) = -g_y(\boldsymbol{x}, \Lambda) + \max_{j, j \neq y} g_j(\boldsymbol{x}, \Lambda) \tag{4}$$

の，Λ に関して微分可能な近似表現である（$\psi \to \infty$ で式 (3) が式 (4) に帰着）．$d_y(\boldsymbol{x}, \Lambda) < 0$ は正分類を，$d_y(\boldsymbol{x}, \Lambda) > 0$ は誤分類を表す．また，$\ell(\ \cdot\ )$ は，一般に

$$\ell\big(d_y(\boldsymbol{x}, \Lambda)\big) = \frac{1}{1 + \exp\big(-a d_y(\boldsymbol{x}, \Lambda)\big)} \quad (a > 0) \tag{5}$$

などで定義される，Λ に関して微分可能な分類誤り数損失であり，式 (1) の指示関数 $1(\ \cdot\ )$ の近似表現である．

式 (1) の式 (2) による近似の精度は，式 (3) と (5) の（ψ や a に基づく）関数平滑度によって制御され，原理的に MCE 学習は，Λ に関する微分可能性を維持しつつ，分類判断全体を関数形式に組み込んだ式 (2) によって理想的な分類誤りリスクの優れた近似を達成する．結果的に，求めるリスク最小状態に対応する Λ の状態は，Λ を変数とする式 (2) の勾配探索によって容易に計算することができる．

MCE 学習における関数平滑性は，上記の勾配法を利用するためには定式化上欠くことができないものである．しかしそれは，その定式化のための必要を超えて，実質的に学習標本を増やすことによって頑健性の向上の効果をもたらす [5].

2.2　誤分類測度の不十分性

式 (5) より，MCE 学習によって Λ の更新が進むにつれ，誤分類される標本に対する誤分類測度の値は，正の領域から 0 に近づき，やがて負の領域に変更される．正分類の標本に対しては，損失の勾配に応じて，負の領域における測度のより小さな値（絶対値のより大きな値）の実現に向けて変更される．ここで負値の誤分類測度の絶対値は，決定の正しさの確信度に対応している．直感的に，負の領域における誤分類測度が分類の正誤の節目に対する余裕すなわちマージンとして振舞うことが分かる．実際この考え方は，後述する関数マージンの考え方と共通する．ブースティング [8] や，最近多く提案されている大マージンに基づく手法 [10, 11, 12, 13] も，基本的にこの性質を利用している．

上記の分析より，MCE 学習は，分類誤り数の最小化のみならず，分類判断の確かさを表すマージンの増加をも追及しているように見受けられる．しかし，判別関数値のスケール変換により，判断の確かさに変化がなくとも誤分類測度値が変更される事実などから明らかなように，誤分類測度は判断の確かさ，結果的に頑健性の強度を表すには不十分である．より直接的に頑健性を表現するために，誤分類測度の改良が必要である．

3　2 種類のマージン

前節で現れたマージンという用語は，線形判別に関する初期の研究から広く用いられている．マージンには関

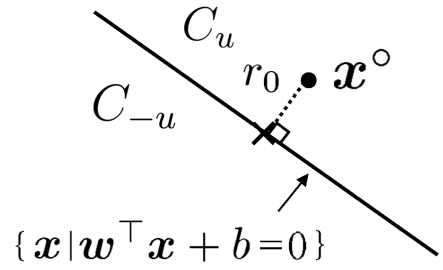

図 1: 2 クラス分類の線形判別関数に対する幾何マージン.

数マージンと幾何マージンの 2 種類がある [7]. ここでは, 単一判別関数 $f(\boldsymbol{x})$ による 2 クラス分類を考える. クラスは C_1 と C_{-1}, 分類決定則は $\boldsymbol{x} \mapsto C_j$ iff $j = \mathrm{sgn}\big(f(\boldsymbol{x})\big)$ である. ここで sgn は, 引数が正ならば 1, 負ならば -1 を返す符号関数である.

今, 学習標本 $\boldsymbol{x}$ とその教師クラス信号 C_u が与えられたとき[1], $z = uf(\boldsymbol{x})$ とすると, $z > 0$ は正分類を, $z < 0$ は誤分類を表し, 更に $|z|$ は分類決定の確信度を表す. この z は関数マージンとよばれ, 容易に分かるとおり, 2 クラスの場合において誤分類測度の正負を反転させたものに等しい. すなわち関数マージンと誤分類測度は, 正負の反転を除いて同一である. よって誤分類測度と同様, 関数マージンは分類判断の強さを表現はしているものの, 決定境界と標本分布との相対関係を反映すべき分類判断の確かさを必ずしも的確に表現していない.

分類判断の確かさを直接的に反映する合理的な測度を, 決定境界とその付近における標本との距離に求めることができる. 特に分類器の頑健性を議論する際に重要となるのは, 決定境界とそれに最も近い正分類の学習標本 $\boldsymbol{x}^\circ$ との距離, すなわち幾何マージンである. SVM では線形判別関数 $f(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x} + b$ による 2 クラス分類に対して幾何マージンを導出している[2] (図 1). ここで $\boldsymbol{w}$ は重みベクトル, b はバイアス項, $^\top$ は行列の転置を表す. $\boldsymbol{x}^\circ \in C_u$ として, 簡単な計算により, 幾何マージンは

$$r_0 = \frac{u(\boldsymbol{w}^\top\boldsymbol{x}^\circ + b)}{||\boldsymbol{w}||} \tag{6}$$

で与えられる. ただし $||\cdot||$ は L_2 ノルムである. 学習標本が母集団から適切に抽出されているならば, この幾何マージンを大きくとることは, 同じ母集団から将来 (学習標本の付近に) 出現する未知標本の正確な分類を約束

[1]教師クラス・インデクスに関して, 2 クラスの場合は記号 u を $(u = 1, -1)$, 多クラスの場合は記号 y を $(y = 1, ..., J)$ 用いる.

[2]カーネル関数を用いた非線形 SVM では, $\boldsymbol{x}$ の替わりにそれを変換した特徴ベクトル $\phi(\boldsymbol{x})$ を用い, 特徴空間上での線形判別関数に対する幾何マージンを考える.

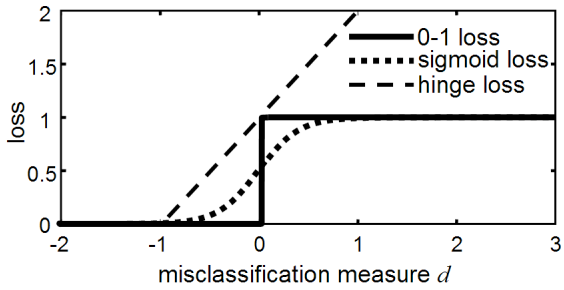

図 2: 各種の損失関数.

してくれるもののように思われる. すなわち, 幾何マージンは決定の頑健性を直接的に表している.

式 (6) から分かるとおり, 幾何マージンは, 関数マージンを重みベクトルのノルム $||\boldsymbol{w}||$ で正規化したものと等しい. だが先述のように, 関数マージンの変化は頑健性の増加に直結しない. それゆえ SVM では, 目標の関数マージンを一定値 (例えば 1) に保ちながら $||\boldsymbol{w}||$ を減少させる.

4 大幾何マージン最小分類誤り学習

4.1 判別関数の一般形に対する幾何マージン

SVM で導出されている幾何マージンは, 2 クラスの線形判別関数に限定されている. また線形分離不可能な現実的状況の場合, SVM では, 係数ノルムの最小化を経由する幾何マージン最小化と, 関数マージンに対するヒンジ損失 (図 2 の破線) の最小化が, 両最適化に対する組み合わせ係数をデータ依存で決めながら, 同時に行われる [3] (しばしばソフトマージン SVM とよばれる). ここで, MCE が多クラスの広範な判別関数に適用可能であることと, MCE で最小化される式 (5) の平滑化 0-1 損失 (図 2 の点線) が, ヒンジ損失と違い, 本来最小化されるべき 0-1 損失 (図 2 の実線) と強い一貫性を持っていることに注意しよう. したがって, 目指すべき新しい学習法は, 未知標本をも対象にした最小分類誤り状態との一貫性の実現を基礎として, 広範な判別関数に適用可能な MCE 学習に, 幾何マージン最大化による分類頑健性の向上の機構を組み込んだものとなることが理解できる. そこでまず我々は, 線形判別関数に限定されない判別関数の一般形に対する幾何マージンを導出する.

式 (6) と同様に, 一般的な幾何マージン r も, 決定境界とそれに最も近い正分類の学習標本 $\boldsymbol{x}^\circ$ とのユークリッド距離として定義する (図 3). なお $\boldsymbol{x}^\circ \in C_y$ とし, また簡単化のため, $g_j(\boldsymbol{x}, \Lambda)$ $(j = 1, ..., J)$ は $\boldsymbol{x}$ と Λ に関して微分可能とする. ここで重要なことに, 誤分類測

度が 0 となる点の集合である

$$\mathcal{B}_y(\Lambda) = \left\{ \boldsymbol{x} \middle| d_y(\boldsymbol{x}, \Lambda) = 0 \right\} \tag{7}$$

は，パターンが C_y に分類されるか否かの境界を表す．実際，$\psi \to \infty$ とおいた式 (4) の場合，$\mathcal{B}_y(\Lambda)$ がその境界を正確に表現することは容易に確かめられる．そして ψ を有限値に設定することで，$\mathcal{B}_y(\Lambda)$ は滑らかな（微分可能な）超曲面としてその境界を（任意の精度で）近似する．したがって r は，$\mathcal{B}_y(\Lambda)$ 上の点と $\boldsymbol{x}^\circ$ との距離の最小値として与えられ，以下の制約条件付き最小化問題を解くことにより求まる．

$$\text{minimize}_{\boldsymbol{x}} \ ||\boldsymbol{x} - \boldsymbol{x}^\circ||^2 \quad \text{subject to } d_y(\boldsymbol{x}, \Lambda) = 0 \tag{8}$$

上式の解を $\boldsymbol{x}^*$ として，$r = ||\boldsymbol{x}^* - \boldsymbol{x}^\circ||$ となる．Lagrange 乗数 λ を導入して，次式の評価関数を考える．

$$J(\boldsymbol{x}, \lambda) = ||\boldsymbol{x} - \boldsymbol{x}^\circ||^2 + \lambda d_y(\boldsymbol{x}, \Lambda) \tag{9}$$

乗数法により，$\boldsymbol{x}^*$ は次式を満たさなければならない．

$$2(\boldsymbol{x}^* - \boldsymbol{x}^\circ) + \lambda \nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda) = \boldsymbol{0} \tag{10}$$

$$d_y(\boldsymbol{x}^*, \Lambda) = 0 \tag{11}$$

なお $\nabla_{\boldsymbol{x}}$ は $\boldsymbol{x}$ に関する勾配演算子であり，点 $\boldsymbol{x}^*$ において $\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda) \neq \boldsymbol{0}$ とする．式 (10) より，$\boldsymbol{x}^\circ - \boldsymbol{x}^*$ は $\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)$ の $\lambda/2$ 倍となり，r は次式となる．

$$r = \frac{|\lambda|}{2} ||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)|| \tag{12}$$

ここで点 $\boldsymbol{x}^*$ を中心に $d_y(\boldsymbol{x}, \Lambda)$ を Taylor 展開する．

$$d_y(\boldsymbol{x}, \Lambda) = d_y(\boldsymbol{x}^*, \Lambda) + \nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)^\top (\boldsymbol{x} - \boldsymbol{x}^*) + o\big(||\boldsymbol{x} - \boldsymbol{x}^*||\big) \tag{13}$$

なお $o(...)$ は Landau の記号である．このとき式 (11) より $d_y(\boldsymbol{x}^*, \Lambda) = 0$，よって式 (13) に $\boldsymbol{x} = \boldsymbol{x}^\circ$ を代入して

$$\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)^\top (\boldsymbol{x}^\circ - \boldsymbol{x}^*) = d_y(\boldsymbol{x}^\circ, \Lambda) + o(r) \tag{14}$$

が得られ，更に式 (10) より次式が得られる．

$$\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)^\top (\boldsymbol{x}^\circ - \boldsymbol{x}^*) = \frac{\lambda}{2} ||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)||^2 \tag{15}$$

上 2 式を λ について解き，式 (12) に代入することにより，次式が導かれる．

$$r = \frac{\big|d_y(\boldsymbol{x}^\circ, \Lambda) + o(r)\big|}{||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^*, \Lambda)||} \tag{16}$$

$|d_y(\boldsymbol{x}^\circ, \Lambda)|$ は $\boldsymbol{x}^\circ$ に対する関数マージンである．結果的に幾何マージンは，$\boldsymbol{x}^\circ$ が境界に十分近い場合，それに最も近い境界上の点における関数マージン（あるいは誤分類測度）の勾配のノルムで関数マージンを正規化したものに等しい．

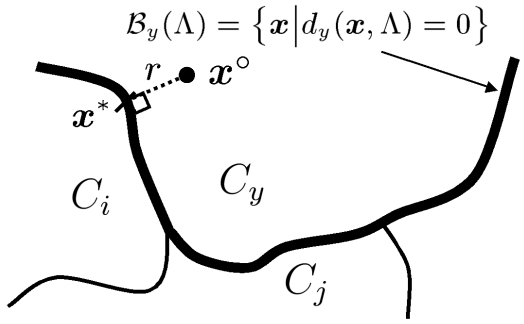

図 3: 判別関数の一般形に対する幾何マージン．

一般の判別関数の場合，$\boldsymbol{x}^\circ$ が境界に十分近いと仮定する．これにより，計算の難しい境界上の点 $\boldsymbol{x}^*$ を $\boldsymbol{x}^\circ$ に置き換えることができ，しかも式 (13) の展開における $o(...)$ を無視できる．この場合の幾何マージン（の近似値）は以下の扱いやすい形式で与えられる．

$$r = \frac{-d_y(\boldsymbol{x}^\circ, \Lambda)}{||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}^\circ, \Lambda)||} \tag{17}$$

なお線形判別関数型や後述の実験で扱うプロトタイプ型の分類器の場合，$\boldsymbol{x}$ に関する線形性ゆえに $o(...)$ 成分が消滅するとともに，$\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}, \Lambda)$ が $\boldsymbol{x}$ に無関係となるため，パターンと境界との十分な近さは要請されず，$\boldsymbol{x}^\circ$ が境界を離れても式 (17) が正確になりたつ[3] ．

4.2 MCE 学習への幾何マージン最大化機構の導入

一般に，決定境界に最も近い学習標本の探索も容易ではないので，式 (17) における $\boldsymbol{x}^\circ$ は，決定境界付近のいずれかの学習標本を表すとする．この式から分かるとおり，決定境界付近の関数マージン（右辺分子）の増加および関数マージンあるいは誤分類測度の勾配のノルム（右辺分母）の減少により，幾何マージンが増加する．だが 2.2 で述べたように，関数マージン（誤分類測度）の変化は頑健性の増加に直結しない．このことは，測度の定数倍により測度の勾配のノルムも同じ定数倍となり，両者の比である幾何マージンが変化しないことからも推測される．したがって幾何マージン（すなわち頑健性）を増加させるためには，関数マージンの増大のみならず，誤分類測度の勾配の大きさを拘束する（望むべくは小さくする）必要がある．誤分類測度の入力パターンに関する微分は，パターンの微小変化に対する測度の変化すな

[3] ただし誤分類測度として式 (4) を採用した場合．

わち分類決定の変動を表す．よってこの拘束により，決定境界付近の入力パターンの小さな変化に対する分類決定の変動が抑えられることとなり，この観点からも頑健性の向上が理解される．結局，“少なくとも決定境界付近における誤分類測度の勾配の大きさを拘束する MCE 学習” が，幾何マージンを大きく保つ MCE 学習であり，未知標本の高精度な分類を実現するであろうことが理論的に示された．

正則化に倣うならば，$\lambda > 0$ を正則化パラメータとして，$R(\Lambda) + \lambda \widetilde{R}(\Lambda)$ を Λ に関して最小化すればよいであろう．ここで $R(\Lambda)$ は式 (2) で近似される分類誤り数リスクであり，$\widetilde{R}(\Lambda)$ は，$\mathcal{X}_B$ を決定境界付近のパターン集合として

$$\widetilde{R}(\Lambda) = \sum_{y=1}^{J} \int_{\mathcal{X}_B} p(C_y, \boldsymbol{x}) ||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}, \Lambda)||^2 \, \mathrm{d}\boldsymbol{x} \tag{18}$$

で与えられる．実際，誤分類測度 $d_y(\boldsymbol{x}, \Lambda)$ が分類決定則をスカラーで表現していることを考えると，$\widetilde{R}(\Lambda)$ は分類決定則に対する Tikhonov 型の正則化項 [14] の形式を成している．しかしこのアプローチは，各最適化要素に対する重み付け λ の決定などが発見的な方法に頼らざるを得ず，ベイズ誤り推定との一貫性が失われる危険性が大きい．したがって，幾何マージン最大化を直接的に追求する MCE 学習手続きが望ましい．

そこで，正負を反転させた幾何マージンに対応する

$$D_y(\boldsymbol{x}, \Lambda) = \frac{d_y(\boldsymbol{x}, \Lambda)}{||\nabla_{\boldsymbol{x}} d_y(\boldsymbol{x}, \Lambda)||} \tag{19}$$

が，正分類のときに負値，誤分類のときに正値をとるという，誤分類測度として必要な性質を継承することを利用して，この $D_y(\boldsymbol{x}, \Lambda)$ を新たな誤分類測度として採用する MCE 学習を提案する．この新しい MCE 学習は，分類誤り数を減少させるのと同時に，損失が一定以上の傾きを持つ決定境界付近の標本に対して，$D_y(\boldsymbol{x}, \Lambda)$ を負方向に大きく更新させる．すなわち本手法は，境界付近において幾何マージンを直接的に増大させる．

5 比較実験

提案手法は元来，多様な判別関数に対して適用可能であるが，ここでは一例として，プロトタイプとのユークリッド距離を判別関数とする分類器を用いた実験により，提案手法の有効性を確認する．距離と確率との近縁性より，この分類器は汎用性が高く，音声認識などで多用される隠れマルコフモデル（HMM）などの確率測度型の判別関数に容易に適用可能である．

クラス C_j における判別関数は次式で与えられる．

$$g_j(\boldsymbol{x}, \Lambda) = -||\boldsymbol{x} - \boldsymbol{p}_j||^2 \tag{20}$$

ここで $\boldsymbol{p}_j$ は C_j に属するプロトタイプ・ベクトルであり，C_j にプロトタイプが複数ある場合は，$\boldsymbol{x}$ に最も近いプロトタイプとする．また Λ はすべてのプロトタイプの集合である．C_y に属する学習標本 $\boldsymbol{x}$ が与えられたとする．係数 ψ を ∞ にした式 (4) の誤分類測度は，$\boldsymbol{x}$ に対する best-incorrect クラスを C_i として

$$d_y(\boldsymbol{x}, \Lambda) = ||\boldsymbol{x} - \boldsymbol{p}_y||^2 - ||\boldsymbol{x} - \boldsymbol{p}_i||^2 \tag{21}$$

となり，更に式 (19) の幾何マージンに対応する誤分類測度は

$$D_y(\boldsymbol{x}, \Lambda) = \frac{||\boldsymbol{x} - \boldsymbol{p}_y||^2 - ||\boldsymbol{x} - \boldsymbol{p}_i||^2}{2||\boldsymbol{p}_y - \boldsymbol{p}_i||} \tag{22}$$

となる．損失関数は，$D_y(\boldsymbol{x}, \Lambda)$ を引数とした式 (5) の平滑化 0-1 損失であり，プロトタイプの更新式は，$j = y, i$ に対して

$$\boldsymbol{p}_j \leftarrow \boldsymbol{p}_j - \varepsilon \ell'\big(D_y(\boldsymbol{x}, \Lambda)\big) \nabla_{\boldsymbol{p}_j} D_y(\boldsymbol{x}, \Lambda) \tag{23}$$

で与えられ（$\varepsilon > 0$ は学習係数），上式における誤分類測度の微分式は，C_y，C_i それぞれに対して，

$$\nabla_{\boldsymbol{p}_y} D_y(\boldsymbol{x}, \Lambda) = \frac{\boldsymbol{p}_y - \boldsymbol{x}}{||\boldsymbol{p}_y - \boldsymbol{p}_i||} - \frac{d_y \cdot (\boldsymbol{p}_y - \boldsymbol{p}_i)}{2||\boldsymbol{p}_y - \boldsymbol{p}_i||^3} \tag{24}$$

$$\nabla_{\boldsymbol{p}_i} D_y(\boldsymbol{x}, \Lambda) = \frac{\boldsymbol{x} - \boldsymbol{p}_i}{||\boldsymbol{p}_y - \boldsymbol{p}_i||} - \frac{d_y \cdot (\boldsymbol{p}_i - \boldsymbol{p}_y)}{2||\boldsymbol{p}_y - \boldsymbol{p}_i||^3} \tag{25}$$

で与えられる．ただし d_y は式 (21) の従来型の誤分類測度である．

実験には UCI Machine Learning Repository [4] が提供する Glass Identification データセットを用いた．このデータセットは 6 クラス 214 個のガラス標本パターンで構成されており，各ガラス標本の中に含まれる 9 種類の酸化物の含有量が，特徴データとして与えられている．

データセットからある一つのパターンを認識対象として取り除き，残りのパターンを用いて分類器を学習した後に，取り除いたパターンを認識させるという処理を，214 個全てのパターンに対して行って認識率を計算した (Leave-One-Out 法)．また，取り除いた一つのパターンを認識対象とするオープン・データ・テストに加えて，学習に用いた 213 個のパターンを対象にして認識率を計算するクローズド・データ・テストも同時に行った．クローズド・テストの認識率計算は 214 回行われるので，それらを平均したものを最終的なクローズド・テストの認識率とした．

提案手法（ここでは幾何マージン法とよぶ）と従来の MCE 学習法（従来法）との比較実験結果を図 4 および

[4] http://archive.ics.uci.edu/ml/

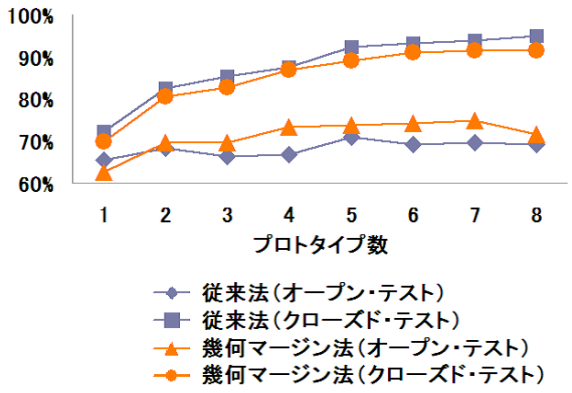

図 4: プロトタイプ数を変化させた場合の認識率.

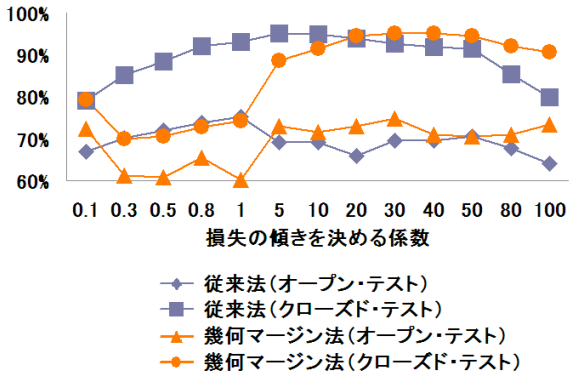

図 5: 損失の傾きを変化させた場合の認識率.

図 5 に示す．なお従来法は，誤分類測度として従来型の式 (21) を用いた MCE 学習法であり，式 (23) において $D_y(\boldsymbol{x}, \Lambda)$ を $d_y(\boldsymbol{x}, \Lambda)$ に置き換えたものである．図 4 は，式 (5) の損失の傾き係数 a を 10 に固定し，プロトタイプ数を変化させた場合の実験結果である．また図 5 は，プロトタイプ数を 8 個に固定し，係数 a の値を変化させた場合の実験結果である．なおプロトタイプ数はクラス間で同一とした．

図 4 では，幾何マージン法の性能が従来法と比較して，クローズド・テストでは低くなっている一方，オープン・テストでは（プロトタイプ数が 1 個のときを除いて）安定して高くなっている．これは幾何マージン法の未知標本に対する頑健性の高さを実証している．次に図 5 のオープン・テストでは，係数の値を 5 以上にしたときに幾何マージン法の性能は従来法に比べて高くなっている．また，オープン・テストとクローズド・テストの実験結果の相関係数を計算してみると，図 4 の実験結果に関しては，従来法の値は 0.80，幾何マージン法の値は 0.95 となった．図 5 の実験結果に関しては，従来法の値は 0.48，幾何マージン法の値は 0.86 となった（小数点以下第 3 位を四捨五入）．どちらの場合も，幾何マージン法におけるオープン・テストの認識率とクローズド・テストの認識率の相関係数値が最も高くなっていることがわかる．元来，実際の利用場面においてオープン・テストの評価結果を得ることは容易ではなく，それを補うためにもクローズド・テストがオープン・テストの優れた近似となっていることが望まれる．この意味でも，提案手法である幾何マージン法が大きな実用価値を持つことは明らかである．

6 従来手法との関連

まず線形判別関数 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ による 2 クラス分類（$C_u, u = -1, 1$）という限定された状況を考える．このとき先述のように，誤分類測度が $d_u(\boldsymbol{x}, \Lambda) = -u(\boldsymbol{w}^\top \boldsymbol{x} + b)$ で与えられる．そして学習において，最小化対象である式 (2) の右辺における平滑化 0-1 損失関数 $\ell(\ \cdot\ )$ をヒンジ型損失関数（図 2 の破線）に置き換え，更に幾何マージン制御機構として式 (18) の正則化を適用してみよう．この正則化項が（$\mathcal{X}_B$ が $\boldsymbol{w}$ に大きく影響されないならば）$||\boldsymbol{w}||^2$ の定数倍となることは容易に分かり，この学習はソフトマージン SVM における学習に帰着することがわかる．すなわち，提案手法と SVM の間には以下の違いが存在する．

- 損失関数：提案手法は平滑化 0-1 損失を，SVM はヒンジ型損失を採用.
- 幾何マージンの導出：提案手法は多クラス分類における一般の判別関数に対して，SVM は 2 クラス分類における線形判別関数（もしくは非線形特徴空間上の線形判別関数）に対して導出.
- 幾何マージンの制御方法：提案手法は幾何マージンを誤分類測度として用いて直接的に増大化，SVM は正則化手法を用いて増大化.

最小分類誤り確率状態との一貫性という観点から見ると，明らかに，SVM で使われるヒンジ損失よりも提案手法で使われる平滑化 0-1 損失の方が望ましい．そして提案手法は SVM と比べて，幾何マージンの適用範囲がより一般化されているのと同時に，その制御方法も，各最適化要素の重み付け決定などを発見的に行わなければならない正則化に頼ることなく，直接的に行えることがわかる．

我々が定式化した一般的な幾何マージンは，[10] の大マージン手法および [15] の改良型 LVQ それぞれで定義されているマージンと関係が深い．前者は次式で定めら

れる相対的マージン $\widetilde{r}$ の最大化を追求する．

$$\widetilde{r} = \min_{\boldsymbol{x} \in \mathcal{S}} \frac{-d_y(\boldsymbol{x}, \Lambda)}{|g_y(\boldsymbol{x}, \Lambda)|} \tag{26}$$

ここで $\mathcal{S}$ は決定境界付近の正しく分類される学習標本の集合である．この $\widetilde{r}$ は有界であり，その最大点を探索することが可能である [10]．また後者は，MCE 学習を通して，次式の正規化されたマージンを増大させる．

$$\mu = \frac{-d_y(\boldsymbol{x}, \Lambda)}{|g_y(\boldsymbol{x}, \Lambda)| + |\max_{j, j \neq y} g_j(\boldsymbol{x}, \Lambda)|} \tag{27}$$

μ を MCE 学習に組み入れることで，プロトタイプの発散が抑えられることが示されている [15]．なお $\widetilde{r}$，μ における $d_y(\boldsymbol{x}, \Lambda)$ として，式 (4) を採用する．ここで式 (17) と比較すれば分かるとおり，$\widetilde{r}$ および μ はいずれも，一般的幾何マージン r と分母が異なっている．つまり [10] および [15] は，判別関数の大きさを拘束することにより誤分類測度の勾配ノルムの拘束を図り，近似的に幾何マージンを制御していると考えることができる．

[11] の大マージン HMM 手法は，学習におけるパラメータの発散を抑えるために，HMM 平均ベクトルの差に対する制約を課している．平均ベクトルはプロトタイプと等価であり，しかもプロトタイプ型分類器の場合における幾何マージンに基づく誤分類測度は式 (22) となる．ここで式 (22) の分母が $2||\boldsymbol{p}_y - \boldsymbol{p}_i||$ であることから，幾何マージンを増加させるためにはこの値を減少させること，すなわちプロトタイプ同士の距離を近づけることが必要であるのがわかる．したがって [11] の手法も幾何マージンを制御していると考えられる．

7　おわりに

広範な分類課題における頑健性の向上のために，SVM の利点である幾何マージン最大化の概念を MCE 学習に適用することを目指した．そして，線形判別関数に限定されない判別関数の一般形に対する幾何マージンを定式化し，この一般的な幾何マージンを大きくするような MCE 学習を行う新たな学習アプローチを提案した．

MCE 学習にマージン最大化の機構を取り入れようとした研究は，これまでにもいくつか報告されている [10, 11, 12, 13]．しかしそれらの研究で行われている学習の仕組みは，実際には関数マージンの最大化を目指すものであった．また，分類決定の頑健性を高めるためには関数に何らかの制約を入れる必要があるとの指摘もあるが [9, 14, 16]，そこでは幾何マージン最大化という概念は明確に示されていない．

我々は，頑健性向上のためには関数マージンに替えて幾何マージンを最大化する必要があることを明確に示し，それを実現する分類器の学習方法を提案した上で，プロトタイプ型分類器への応用を例とした実験でその有効性を確かめた．今後は，様々な分類器における幾何マージン増大化の効果を実験により検証する予定である．

謝辞

研究の機会を与えて下さった，情報通信研究機構 MASTAR プロジェクト 中村 哲プロジェクト・リーダーに感謝します．本研究の一部は日本学術振興会 科学研究費補助金 基盤研究 (B)（課題番号：19300064）の援助により行われている．

参考文献

[1] 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，オーム社，東京，1998.

[2] R.O. Duda, P.E. Hart, and D.G. Stork（尾上守夫監訳）, パターン識別，新技術コミュニケーションズ，東京，2001.

[3] C.M. Bishop（元田 浩，栗田多喜夫，樋口知之，松本裕治，村田 昇 監訳）, パターン認識と機械学習 上・下，シュプリンガー・ジャパン，東京，2007.

[4] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” IEEE Trans. Signal Processing, vol.40, no.12, pp.3043-3054, Dec. 1992.

[5] E. McDermott and S. Katagiri, “A derivation of minimum classification error from the theoretical classification risk using Parzen estimation,” Computer Speech and Language, vol.18, pp.107-122, April 2004.

[6] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[7] N. Cristianini and J.Shawe-Taylor（大北 剛 訳）, サポートベクターマシン入門，共立出版，東京，2005.

[8] Y. Freund and R.E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Comput. Syst. Sci., vol.55, no.1, pp.119-139, 1997.

[9] 六井 淳，“認識誤り最小化基準に基づく高速な識別学習法,” 信学論 (D-II)，vol.J87-D-II，no.8，pp.1555-1564，Aug. 2004.

[10] C. Liu, H. Jiang, and X. Li, "Discriminative training of CDHMMs for maximum relative separation margin," Proc. ICASSP, pp.I-101-104, 2005.

[11] H. Jiang, X. Li, and C. Liu, "Large margin hidden Markov models for speech recognition," IEEE Trans. Audio, Speech, Lang., Process., vol.14, no.5, pp.1584-1595, Sept. 2006.

[12] J. Li, M. Yuan, and C.-H. Lee, "Approximate test risk bound minimization through soft margin estimation," IEEE Trans. Audio, Speech, Lang., Process., vol.15, no.8, pp.2393-2404, Nov. 2007.

[13] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training: a theoretical risk minimization perspective," Computer Speech and Language, vol.22, pp.415-429, Oct. 2008.

[14] C.M. Bishop, "Training with noise is equivalent to Tikhonov regularization," Neural Computation, vol.7, no.1, pp.108-116, 1995.

[15] 佐藤 敦, 山田敬嗣, "新しい誤分類尺度を用いた学習ベクトル量子化の定式化," 信学論 (D-II), vol.J82-D-II, no.4, pp.650-659, April 1999.

[16] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," Science, vol.247, pp.978-982, Feb. 1990.