Vol.2009-MPS-75 No.23
2009/9/11

カーネル判別分析を利用した多クラス識別のためのパラメータ自動決定法

関口涼平[†1]　高橋治久[†2]　堀田一弘[†2]

本論文では，カーネル判別分析 (KDA) に基づいた新しい多クラス識別器を提案する．KDA は主にパターン識別の前処理として用いられ，線形判別分析を使う場合に比べ良い識別性能が出せることが知られている．しかしながら，その性能は SVM と同様カーネルパラメータに大きく依存し，学習における最適なカーネルパラメータを導くには膨大な事前実験を必要とする．このため学習そのものよりも事前実験に要する計算量が膨大になり応用の障害になっている．本論文では，KDA に対し，分離度の理論に基づいて最適なカーネルパラメータを自動決定するアルゴリズムを提案し，計算機実験によりその性能を評価する．SVM との計算機実験による比較により，提案手法が少ない計算時間でより良い性能を達成できることを示す．

The Automatic Parameter Tuning for Multi-class Learning with KDA

Ryohei Sekiguchi,[†1] Haruhisa Takahashi[†2] and Kazuhiro Hotta[†2]

This paper proposes a new learning machine based on Kernel Discriminant Analysis (KDA). KDA is mainly used as a pre-processing process of learning machines, thereby better performance is achieved than linear discriminant analysis in some cases. Despite potential ability, KDA is greatly dependent on a kernel parameter, and thus to attain the better performance takes the more of huge preparing experiments for drawing the optimal kernel parameter. To alleviate this difficulty, we propose a novel algorithm to determine the optimal kernel parameter in a reasonable computational effort. The algorithm is obtained based on the theory of the discriminant creterion. We show that the proposed learning method outperforms SVM in both generalization and computation time through computer experiments.

1. はじめに

パターン認識の分野において，カーネル法は非線形識別問題に対する有効な手法であり，学習機械で用いられている．その中でも，カーネル判別分析法 (KDA) は線形判別法 (LDA) を非線形に拡張した特徴選択の手法である．LDA は各クラスの分離度を最大化するような部分空間を選択し，効率的な次元削減から分類精度を改善できるため，データの前処理として用いられることが多い．

また，判別分析は多クラス識別問題に対しても，各クラスのデータにおけるクラス内散布行列，各クラス間でのクラス間散布行列を用いることで，多クラス分類を容易にする空間を提供することが知られている．

しかし KDA においては，代表的な識別器であるサポートベクトルマシン (SVM) と同様の問題が生じる．すなわちカーネルとして，ガウシアンカーネルが広く用いられるが，そのカーネルのパラメータに対しては，自動決定する方法がない．従来は，経験的にパラメータを設定するのが一般的であり，クロスバリデーション (CV) 法などによりテスト性能を最大化するような最適パラメータを選択していた．したがって膨大な予備実験なしにその識別能力を最大限引き出すことはできない．学習機械として完全となるためには，パラメータが学習過程で，自動的に決定される方法が望まれる．

本論文では KDA に対し，最適なカーネルパラメータを自動決定する合理的な方法を提案し，計算機実験によりその効果を検証する．

パラメータを選択する基本的アイデアは，選択された特徴に対し，最適なパラメータを適切な評価関数を定めて求めることである．KDA における評価関数は，カーネル写像により高次元空間に写像された各クラス間の分離度とするのが合理的である．これを実現するため，特徴のクラス内分散とクラス間分散の比を評価関数として選び，最適なパラメータを自動決定するアルゴリズムを提案する．分離度を最大にする最適カーネルパラメータの選択と，分離度を最大にする部分空間の選択による，二重の分離度最大化を用いて識別性能の向上を目指す．

†1 電気通信大学大学院情報通信工学専攻
Department of Information and Communication Engineering, The University of Electro-Communications
†2 電気通信大学 情報通信工学科
The University of Electro-Communications

Vol.2009-MPS-75 No.23
2009/9/11

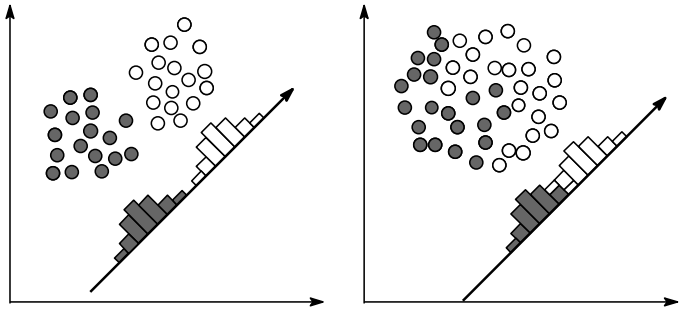

図 1 線形判別分析の例
Fig. 1 Two samples for LDA

提案法の効果を検証するため，多クラスの識別問題について計算機実験を行い，その汎化性能と計算時間についてマルチクラス SVM と比較して検討する．

2 章で本論文で必要となるカーネル判別分析について述べ，3 章ではカーネルの最適パラメータを自動的に決定する手法について述べる．4 章で計算機実験とその考察を示す．

2. カーネル判別分析

判別分析 (LDA) は，同クラス内のデータの分散を小さく，クラス間相互の分散を大きくすることによって次元圧縮された部分空間を求める方法であり，パターン識別における基礎的手法である．LDA は線形変換であるため，本来，分布が線形性をもつデータに対して次元を圧縮することができる（**図 1** 左）．しかし，図 1 右のような分布の非線形性が強いデータ，また複数のクラスが重なりを持つようなデータに対しては，非線形の特徴量を抽出することができず効率的な次元圧縮が難しいと考えられる．

Mika[1] らによって提案されたカーネル判別分析 (KDA) は，カーネル写像によって高次元空間に非線形写像を行い，線形判別分析を可能とした上で，写像先で LDA を適用することにより，分布の非線形性の強いデータについて判別分析を行うことを可能にする．すなわち，非線形判別分析を行うことになり，線形判別分析では抽出できないデータの特徴量を見つけることができる．

2.1 カーネル法による非線形写像

入力された n 次元特徴ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ とし，全 C クラスからなる訓練サンプルを $S = \{(\boldsymbol{x}_{ij}, l_{ij})_{j=1}^{N_c}\}_{i=1}^{C}$ とする．ここで， $\boldsymbol{x}_{ij}$ はクラス i の j 番目のベクトル，l_{ij} は $\boldsymbol{x}_{ij}$ のラベル， N_c は c 番目のクラスにおけるデータ数を表している．

カーネル写像によりベクトル $\boldsymbol{x}$ は $\boldsymbol{\phi}(\boldsymbol{x}) \in F$ によって，無限次元を含む高次元線形空間 F に非線形写像される．

カーネル写像を用いれば，高次元写像 $\boldsymbol{\phi}(\boldsymbol{x})$ を直接扱うことなく，その内積であるカーネル関数（スカラー関数）

$$K(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle = \boldsymbol{\phi}(\boldsymbol{x})^t \boldsymbol{\phi}(\boldsymbol{z})$$

を使うことにより，識別器を扱うことができる．

カーネル関数から入力ベクトル $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\}$ によって作られる $(p \times p)$ 行列 $K = (K_{ij}) \equiv (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ をカーネル行列あるいはグラム (Gram) 行列と呼ぶ．特徴行列を $\Phi = (\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_p))^t$ とすれば， $K = \Phi^t \Phi$ となり半正定値行列となる．

2.2 カーネル多重判別分析による定式化

C 個のクラスでの識別問題において，多重判別分析は $(C-1)$ 個の判別関数を持つ．入力特徴ベクトル $\boldsymbol{x}$ を用いた D 次元カーネル特徴を $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{w}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{w}_D, \boldsymbol{x}))^T$ とおくと， D 次元空間から $(C-1)$ 次元への射影は $(C-1)$ 個の判別関数により

$$y_k = \sum_{i=1}^{D} \alpha_{ki} K(\boldsymbol{w}_i, \boldsymbol{x}) \quad (k = 1, 2, \ldots, c-1) \tag{1}$$

と表すことができる．y_k を判別空間に写像されたデータ $\boldsymbol{y}$ の要素，係数ベクトル $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k1}, \ldots, \alpha_{kD})^T$ から構成される D 行 $(C-1)$ 列の係数行列を $A = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{C-1}]$ とすると，式 (1) は

$$\boldsymbol{y} = A^T \boldsymbol{k}(\boldsymbol{x})$$

と表せる．ここで K はカーネル関数， $\boldsymbol{w}_i$ は i 番目のカーネルの位置を表す．

各クラスの平均ベクトル $\bar{\boldsymbol{k}}_c$ と全平均ベクトル $\bar{\boldsymbol{k}}_T$ は以下のように表せる．

$$\bar{\boldsymbol{k}}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \boldsymbol{k}(\boldsymbol{x}_{ci})$$

$$\bar{\boldsymbol{k}}_T = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \boldsymbol{k}(\boldsymbol{x}_{ci})$$

ここで， N は全学習サンプルの数を表す．

クラス内散布行列 W とクラス間散布行列 B は以下のように与えられる．

$$W = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\boldsymbol{k}(\boldsymbol{x}_{ci}) - \bar{\boldsymbol{k}}_c)(\boldsymbol{k}(\boldsymbol{x}_{ci}) - \bar{\boldsymbol{k}}_c)^T$$

$$B = \frac{1}{N}\sum_{c=1}^{C} N_c(\bar{\boldsymbol{k}}_c - \bar{\boldsymbol{k}}_T)(\bar{\boldsymbol{k}}_c - \bar{\boldsymbol{k}}_T)^T$$

同様に，判別空間内でのクラス内散布行列とクラス間散布行列はそれぞれ A^TWA , A^TWA で表される．また，クラス内分散を最小にし，クラス間分散を最大にするような判別空間を構成するために，分離度と呼ばれる評価基準を設定する必要がある．多クラス判別分析における分離度の基準はいくつか存在し[2)]，例として以下のような基準が存在する．

$$J = tr\left((A^TWA)^{-1}(A^TBA)\right) \quad (2)$$

$$J = tr\left(W^{-1}B\right) \quad (3)$$

$$J = \frac{|A^TBA|}{|A^TWA|}$$

$$J = \ln|W^{-1}B|$$

本論文での判別空間の構成における基準は，式 (2) である

$$J = tr\{(A^TWA)^{-1}(A^TBA)\}$$

を分離度として設定する．これを最大にする A を求めることが KDA の主要ステップである．

入力データやカーネルの性能によってはクラス内散布行列 W が 0 に限りなく近づき，過学習を起こす可能性がある．そこで通常のクラス内散布行列 W に正則化項を加える．

$$\tilde{W} = W + \zeta I$$

ここで ζ は非負の正則化パラメータを表す．

分離度 J は A^TWA , A^TWA ともに二次形式になっているため， A を定数倍しても値は変化しない．そこで分母を単位行列 I と仮定し，分子であるクラス間散布行列を最大化する制約つき最適化問題により A を求める．

最適化問題

$$\textbf{Maximize} : tr(A^TBA)$$

$$\textbf{Subject to} : A^T\tilde{W}A = I$$

を， Λ を Lagrange 乗数とした Lagrange 関数は

$$L(A) = tr(A^TBA) - tr[(A^T\tilde{W}A - I)\Lambda]$$

となる．上式を A で編微分し 0 を取ることで，以下の固有値問題が得られる．

$$BA = (W + \zeta I)A\Lambda$$

これを解くことにより，分離度 J を最大にする最適な係数行列 A が求められる．ここで，$\Lambda = diag(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0)$ は固有値の対角行列を表している．

2.3 カーネル判別分析の問題点

カーネル判別分析を適用する場合，カーネルの種類によってカーネルパラメータの値をいかに設定するかが問題となる．本論文では最も頻繁に応用されている Gaussian カーネル

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{z}||^2}{2\sigma^2}\right) \quad (4)$$

を例としてこの問題を考える．この場合のカーネルパラメータは σ である．クラスごとに σ を変える場合には，多次元パラメータとなる．本手法は，パラメータを持つ一般的なカーネルにも適用できる．

カーネル判別分析の能力は，カーネルパラメータに非常に強く依存している．最適なパラメータを設定したとき，カーネル法の能力を最大限に引き出すことができるが，事前にこれを行う有効な手法は提案されていない．

カーネルパラメータによって大きく性能が左右される例として，3 クラスのサンプルを持つデータセットを用意し，主成分分析によって得られた最初の 2 次元について**図 2** に示す．また，図 2 のデータに対し，線形判別分析を行った結果を**図 3** に示す． このようなデータに対しては，線形判別分析は他クラスとのデータの重なりを取り除けないため，有効な分析とはいえない．

このデータに対して，Gaussian カーネルによる KDA を行った結果を**図 4** と**図 5** に示す． 図 4 の分布は，後述する手法で求めた最適パラメータで求めた判別空間を表している．この図では学習データは重なりがなくまとまっており，訓練サンプル以外の未知サンプルに対しても非線形な特徴を一般化することができる．一方，図 5 のようにパラメータ σ が最適値と大きく異なる場合，判別空間上では同クラス内のデータには散らばりが見られる．この状態で未知サンプルを入力すると大きく散らばった状態で判別空間に写像され，判別空間上での識別率が下がってしまう．

3. 最適パラメータの自動決定法

3.1 評 価 関 数

前節では，カーネルパラメータの値に依存して非線形写像後の次元圧縮の性能に大きな違いが生じることを述べた．パターン識別の観点から考えれば，同じクラスのデータ間の距離

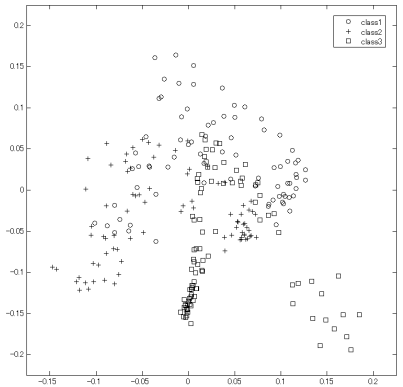

図 2 3 クラスデータセットの例
Fig. 2 Example for three-class dataset

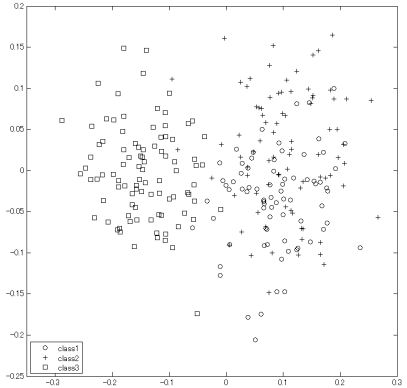

図 3 線形判別分析における判別空間
Fig. 3 Discriminant space by LDA

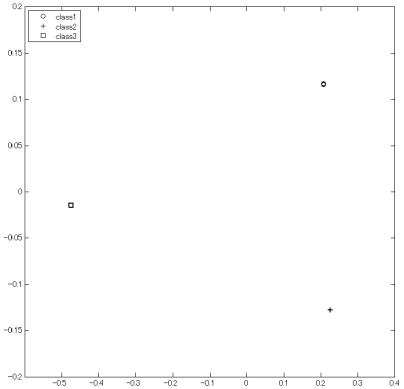

図 4 最適パラメータによる KDA の出力 ($\sigma_{opt} = 1.518$)
Fig. 4 Outputs of KDA for the optimal parameter ($\sigma_{opt} = 1.518$)

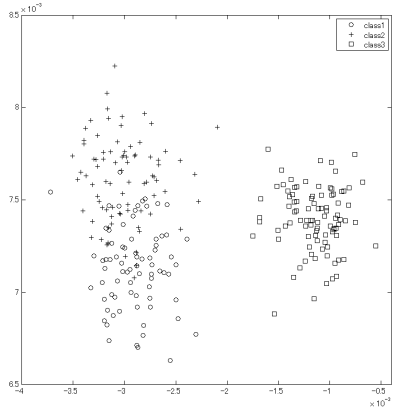

図 5 最適ではないパラメータによる KDA の出力 ($\sigma = 5.000$)
Fig. 5 Outputs of KDA for the not optimal parameter ($\sigma = 5.000$)

は小さく，異なるクラス間の距離は大きければ，判別分析の性能がよいと考えられる．なぜならば，写像空間において異なるクラスの重なりを発生させないようにパラメータを調節することが，多クラス識別におけるマージン最大化に繋がるからである．カーネル写像後の判別空間の構成（LDA の最適化）については，2.2 節で分離度が最大の写像行列を得られることが分かっている．そこで提案手法では，カーネル写像を行う際に，カーネルパラメータを調節して分離度 J を最大のものに導くこと目指す．

本論文では，カーネル写像における評価関数は式 (3) である

$$E(\theta) = tr(\tilde{W}^{-1}B)$$

を使用する．ここで， θ はカーネルパラメータである．最適パラメータは，この条件を満たすように構成された評価関数を最適化することにより得られる．

この関数を，後述する計算機実験で用いるデータで計算し，評価関数のグラフを求めると典型的に**図 6** のような曲線が得られる． 多くの問題では，このように $E(\theta)$ は局所解を持たないことが確認できる．しかし評価関数はデータに依存するため，局所解を持つかどうかを一般的に示すことは難しい．

3.2 評価関数 $E(\theta)$ の微分

評価関数 $E(\theta)$ の θ に関する偏導関数は式 (3) より

$$\frac{\partial E}{\partial \theta} = tr\left(-\tilde{W}^{-1}\frac{\partial \tilde{W}}{\partial \theta}\tilde{W}^{-1}B + \tilde{W}^{-1}\frac{\partial B}{\partial \theta}\right)$$

となる．正則化を行ったクラス内散布行列 $\tilde{W}$ とクラス間散布行列 B の偏導関数は

$$\begin{aligned}\frac{\partial \tilde{W}}{\partial \theta} &= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\left(\frac{\partial \boldsymbol{k}(\boldsymbol{x}_{ci})}{\partial \theta} - \frac{\partial \bar{\boldsymbol{k}}_c}{\partial \theta}\right)\left(\boldsymbol{k}(\boldsymbol{x}_{ci}) - \bar{\boldsymbol{k}}_c\right)^T \\ &\quad + \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\left(\boldsymbol{k}(\boldsymbol{x}_{ci}) - \bar{\boldsymbol{k}}_c\right)\left(\frac{\partial \boldsymbol{k}(\boldsymbol{x}_{ci})}{\partial \theta} - \frac{\partial \bar{\boldsymbol{k}}_c}{\partial \theta}\right)^T \\ \frac{\partial B}{\partial \theta} &= \frac{1}{N}\sum_{c=1}^{C} N_c\left(\frac{\partial \bar{\boldsymbol{k}}_c}{\partial \theta} - \frac{\partial \bar{\boldsymbol{k}}_T}{\partial \theta}\right)\left(\bar{\boldsymbol{k}}_c - \bar{\boldsymbol{k}}_T\right)^T \\ &\quad + \frac{1}{N}\sum_{c=1}^{C} N_c\left(\bar{\boldsymbol{k}}_c - \bar{\boldsymbol{k}}_T\right)\left(\frac{\partial \bar{\boldsymbol{k}}_c}{\partial \theta} - \frac{\partial \bar{\boldsymbol{k}}_T}{\partial \theta}\right)^T\end{aligned}$$

となる．

ここで，式 (4) である Gaussian カーネルの $\theta = \sigma^{-2}$ における微分係数は次式で与えら

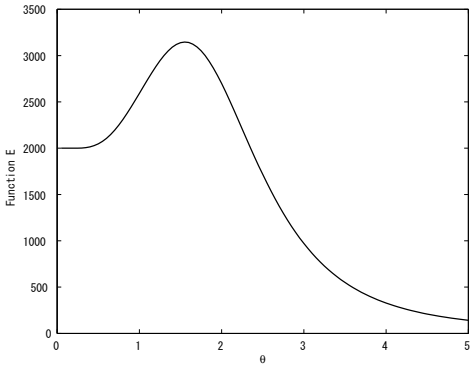

図 6 評価関数 $E(\theta)$
Fig. 6 Cost function $E(\theta)$

れる．

$$\frac{\partial K(\boldsymbol{x},\boldsymbol{z})}{\partial \sigma^{-2}} = -\frac{||\boldsymbol{x}-\boldsymbol{z}||^2}{2}\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{z}||^2}{2\sigma^2}\right)$$

評価関数 $E(\theta)$ を最大にする最適パラメータ $\theta = \theta^\star$ を求めるために，極値を探索する二分法 (Bisection method) を適用する．この方法では $E(\theta)$ の計算回数を少なくできる．また，初期条件への依存度が低く，毎回ほぼ同じ計算時間で結果が出力できるという利点がある．

4. 計算機実験

本研究では提案手法を用いて，多クラスパターン識別の性能比較を行う．

学習データについては Artificial Intelligence and Computer Science Laboratory[⋆1] より “iris” , “wine” , “glass” , “vowel” , “vehicle” の 5 つのデータを使用した（**表 1**）．

性能比較のための識別器として， SVM and Kernel Methods Matlab Toolbox[⋆2] のマルチクラス SVM (one-against-rest) を使用した．カーネルには Gaussian カーネルと多項式カーネルを選択した．Gaussian カーネルにおけるカーネルパラメータ σ , 多項式カーネル

⋆1 http://www.liacc.up.pt/ML/old/statlog/datasets.html
⋆2 http://asi.insa-rouen.fr/ arakotom/toolbox/index.html

表 1 実験に用いたデータセット
Table 1 Datasets for experiment

Database	クラス数	特徴数	データ数
iris	3	4	150
wine	3	13	178
glass	6	10	214
vowel	11	10	528
vehicle	4	18	846

における次数 d, そしてソフトマージンのパラメータである C については，

$$\sigma = [e^{-1}, e^0, \ldots, e^7, e^8]$$
$$d = [1, 2, \ldots, 9, 10]$$
$$C = [e^0, e^1, \ldots, e^8, e^9]$$

のそれぞれ 10 点を定め，各カーネルで 100 点の組み合わせに対し学習を行い，その中でのそれぞれ最も性能の良い値を選択した．

提案手法と比較手法についてはどちらも MATLAB で実装されたものを使用し，演算においては，CPU: 3.20GHz(Pentium4)，メモリ: 2.0GB，OS: Windows の計算機を使用した．

4.1 提案手法の評価

次に提案手法について**図 7** にこの学習法を図示する．

提案手法については，与えられたデータをランダムにパラメータ決定用，学習用，テスト用の 3 種類に分けた．これは，パラメータ決定用データによって最適な判別空間に構成し，その後に未入力である学習用データを射影し学習させることで，未知データに対するより複雑な識別面を学習しようという目的がある．最適カーネルパラメータ σ と判別空間への写像行列 A によって写像された学習用データは，サンプル $\hat{S} = \{(\boldsymbol{y}_{ij}, l_{ij})_{j=1}^{N_c}\}_{i=1}^{C}$ として多クラス識別器に入力される．この学習によって構成された判別空間での識別面を使って，テスト用データを評価する．

3 種類のデータの割合については，全データ数の 50% をパラメータ決定用，40% を学習用データ，そして 10% をテスト用データに設定した．今回の対象データについては，パラメータ決定用データを全データ数の 50% 以上とすると，最適パラメータの値が毎回一定に定まった．

学習用データとテスト用データについては 5-fold cross-validation で評価を行い，これを 20 回繰り返し平均を取った．なお，本実験では KDA で使用する正則化係数は $\zeta = 0.005$ で固定した．また，最適値ではないカーネルパラメータについても同様の実験を行ったとこ

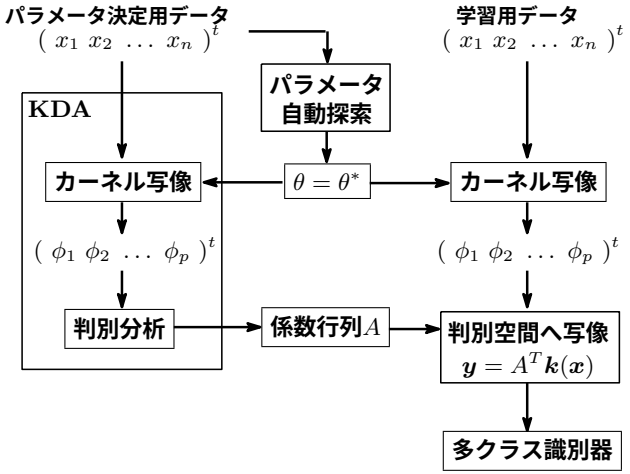

図 7 パターン識別への応用
Fig. 7 Application to Pattern recognition

ろ，図 6 のような評価関数 $E(\theta)$ の極大値で最も誤識別率が低いことを確認した．

本論文では，図 7 の多クラス識別器として one-against-rest 型のブースティング識別器である AdaboostMlt を用いた．AdaboostMlt は試行回数を設定する必要がないため，提案した多クラス識別器全体としては判別分析における正則化係数以外のパラメータを必要としない構成となる．

4.2 実験結果

各データセットに対する識別器の性能比較を**表 2** に示す．実働学習時間は，両方の学習機械に対し，パラメータ探索の時間を含む全実行時間を示している．

実験の結果，提案手法はパラメータ探索を含めた学習時間を考えるとき，探索回数が大きく抑えられ，小規模の計算量で良い結果を出すことができた．また，誤識別率に関しても分離度によるパラメータ決定方法が有効であったことを示している．

一方，多クラス SVM は広い範囲で離散的にパラメータ探索をする必要があるため，大きな実働時間を必要とする．パラメータ探索範囲をより細かく設定すれば提案手法よりも誤識別率を下げることも可能と考えられるが，その結果を出すためには事前実験の計算量が膨大となると考えられる．

表 2 対象データの汎化誤差 [%] と実働時間 [sec.]
Table 2 Generalization error and Implementation time of datasets

	Generalization error [%]			Implementation time [*sec.*]		
	KDA+ Adaboost	SVM (gaussian)	SVM (polynomial)	KDA+ Adaboost	SVM (gaussian)	SVM (polynomial)
iris	4.876	6.832	6.118	20.33	95.12	102.7
wine	2.583	4.983	3.569	48.57	240.3	233.9
glass	27.77	28.35	29.66	108.4	423.5	429.0
vowel	2.409	3.790	5.735	265.1	1326	1401
vehicle	14.80	17.74	18.52	324.8	1175	1221

本論文ではカーネルパラメータを 1 次元の変数として取り扱ったが，多次元パラメータの場合でも同様に提案手法を適用できる．例えば，Gaussian カーネルに関してはクラス c ごとにパラメータ σ_c を設定し，それぞれを調節することが可能である．しかし，これはクラス内分散のばらつきがクラスごとで極端に異なる場合などで効果的であるが，本研究で用いた例ではパラメータ探索の計算時間が増えるのみで，性能の改善は見られなかった．

5. おわりに

本論文ではカーネルのパラメータを自動的に決定することで，カーネル判別分析に基づいた多クラス識別器を提案した．各クラスの分離度を最大にする最適カーネルパラメータを求めることで，多クラスの識別に適した判別空間を構成することができた．また多クラス Adaboost と組み合わせることで，非線形の多クラス SVM よりも高速で同等以上の性能を持つ識別器を実現することができた．

計算機実験では最大 11 クラスの多クラス識別を行ったが，更に大規模なクラスを持つ識別問題についても，近似クラスにクラスタリングを行い，段階的に多クラス識別を行うことで対応できると思われる．

参考文献

1) Mika, S., Ratsch, G., and Muller, K.R.: A mathematical programming approach to the Kernel Fisher algorithm, *In Advances in Neural Inf. Proc. Systems*, Vol.13 pp.591–597 (2001).
2) Fukumizu, K., Bach, F.R. and Jordan, M.I.: Kernel Dimensionality Reduction for Supervised Learning with reproducing kernel Hilbert space, *JMLR*, pp.286:531–537 (1999).