カーネル主成分分析を用いた学習機械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，カーネル主成分分析（KPCA）に基づいた新しい学習機械を提案する．KPCA は，パターン識別の前処理として用いられ，主成分分析を使う場合より良い認識性能が出せる場合も報告されている．KPCA と線形サポートベクトルマシンを合わせたカーネルプロジェクションマシン（KPM）は，モデル選択との併用により，少ない学習時間でサポートベクトルマシン（SVM）と同等の汎化性能が得られる利点があるが，その性能は SVM と同様カーネルパラメータに大きく依存する．本論文では，KPM に対し，KPCA の理論に基づいて最適なカーネルパラメータを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 in Learning with Kernel PCA

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

This paper proposes a new learning machine based on Kernel Principal Component Analysis (KPCA). KPCA is usually used as a pre-processing process preceding application of learning machines, thereby better performance is achieved than linear Principal Component Analysis in some cases. Kernel Projection Machine (KPM), which is proposed by Blanchard etc., applies linear SVM after KPCA with a model selection process. Although KPM can perform equally to Support vector Machine (SVM) with smaller execution time, its performance heavily depends on the kernel parameter. We propose a novel algorithm to determine the optimal kernel parameter in the learning process. The algorithm is obtained based on the theory of KPCA, and we show that the proposed learning method show a better performance than SVM in both generalization and computation time through computer experiments.

1. はじめに

パターン識別の分野では，主成分分析（PCA）やカーネル主成分分析（KPCA）はノイズなどが入った高次元データにたいして前処理として適用することで，識別性能を上げるために用いられることが多い．識別問題では，ノイズがパターン情報を上回るところで，次元圧縮をすることにより，分類精度を改善できる．残念ながら，KPCA を次元圧縮に，SVM を識別器に用いて手書き文字認識に応用した例では，次元圧縮が有効になされていないことが分かっている[4]．この理由が，KPCA と SVM による二重の正則化にあるとの観測に基づき，Blanchard ら[1),2)] は，カーネル射影学習機械（KPM）を提案した．KPM では識別器を線形 SVM とすることにより，二重の正則化を避けることができ，その結果 KPCA による正則化が有効となり，モデル選択により最適な学習機械を選択できる．

KPM は，二重の正則化を防ぐという意味において，新たな学習の枠組みを提案している．実際，識別性能は SVM と同等であり，次元圧縮によりさらにサイズの小さい，実働化に適したネットワークが得られる．

KPM においては，KPCA 処理において 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主として用いられるが，そのカーネルパラメータに対しては，自動決定する方法がなく，膨大な予備実験なしに KPM の能力を最大限引き出すことはできない．このような学習機械をパラメータを持つ学習機械といい，実用上経験的にパラメータを設定するのが一般的であり，パラメータを含めて自動的に学習により決定する方法が望まれる．

本論文では，KPM に対し，最適なカーネルパラメータを自動決定する方法を提案し，計算機実験により，その効果を検証する．一般に KPCA を適用する段階

†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

で，どこの固有値で次元圧縮するかは，モデル選択にゆだねられる．モデル選択については，クロスバリデーション法など，種々の手法があり，問題に適した方法で対応すればよい．カーネルパラメータ決定をモデル選択に含めてしまうと，その手間は膨大となり，できればこれを避けたい．このため，パラメータの自動決定は大事である．

パラメータを選択する基本的アイデアは，選択された固有値（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．評価関数は，基本的に選択された特徴に入る情報と選択されなかった特徴に入る情報の比が最大になるよう選ばれる．本論文では，特徴のそれぞれの領域での平均値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．

提案法の効果を検証するため，いくつかの問題について計算機実験を行い，その汎化性能と出力されたネットワークのサイズについて 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主成分分析について述べ，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理論を述べる．4 章で計算機実験とその考察を示す．

2. カーネル主成分分析

主成分分析（PCA）は，特徴量の相関関係を利用することによって次元圧縮された部分空間への射影を求める方法であり，パターン識別信号処理の基礎的手法である．PCA は線形変換であるため，本来，分布が線形性を持つデータに対して次元圧縮の効果を得ることができる（**図 1** 左）．しかし，図 1 中央のような，分布の非線形性が強いデータに対しては，非線形の特徴量を抽出することができず大きく情報を損失してしまうと考えられる．

Schölkopf ら[4] によって提案されたカーネル主成分分析（KPCA）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写像先で PCA を行うことにより，分布の非線形性の強いデータについて PCA を行うことを可能にする（図 1 右）．すなわち，非線形主成分分析を行うことになり，線形主成分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，サンプルを $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\}$ で表す．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，無限次元も含む高次元線形空間 F に非線形写像される．

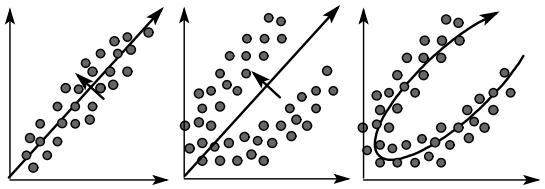

図 1 主成分分析とカーネル主成分分析の例
Fig. 1 Three examples for PCA and KPCA.

カーネル写像を用いれば，高次元写像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）

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

を使えばよい．

カーネル関数から入力ベクトル $\{\xi_1, ..., \xi_l\}$ によって作られる $(l \times l)$ 行列 $K = (K_{ij}) \equiv (K(\xi_i, \xi_j))$ をカーネル行列あるいはグラム（Gram）行列と呼ぶ．写像先の特徴行列を $\Phi = (\phi(\boldsymbol{\xi}_1), \ldots, \phi(\boldsymbol{\xi}_l))^t$ とすれば，$K = \Phi^t \Phi$ となり半正定値行列となる．

2.2 カーネル主成分分析による定式化

線形 PCA は基準点をデータの平均ベクトルにおくことによる次元圧縮によって，原空間と部分空間の誤差を最小にできることが分かっている．KPCA でも重心を基準点とすれば良い結果が期待できる．たとえば，中心から 3 方向に分布の「羽」が出ているような集合であれば，その中心を KPCA の基準点とするほうが効率的に次元圧縮ができる．

写像先で主成分分析を行うため，l 次元特徴空間から d 次元線形部分空間への変換行列を A とする．特徴空間座標で表現された d 次元部分空間を張る i 番目の正規直交基底を $\boldsymbol{a}_i$ とするとき $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ と表せる．特徴 $\phi(\boldsymbol{x})$ の変換後のベクトル $\boldsymbol{\psi}(\boldsymbol{x}) = (\psi_1(\boldsymbol{x}), \ldots, \psi_d(\boldsymbol{x}))^t$ は写像先の平均ベクトル

$$\bar{\boldsymbol{\phi}} = \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{\phi}(\boldsymbol{\xi}_i)$$

を用いて

$$\boldsymbol{\psi}(\boldsymbol{x}) = A^t \left(\boldsymbol{\phi}(\boldsymbol{x}) - \bar{\boldsymbol{\phi}} \right)$$

と表現できる．$\boldsymbol{a}_i = \Sigma_{j=1}^{l} b_j^i \left[\boldsymbol{\phi}(\boldsymbol{\xi}_j) - \bar{\boldsymbol{\phi}} \right]$ とすれば，

$$A = \Phi M B$$
$$M = \left(I_l - \frac{1}{l} \mathbf{1}\mathbf{1}^t \right)$$

が成り立つ．ここで M は中心化の行列であり，$B = (\boldsymbol{b}_1 \ldots \boldsymbol{b}_d)$，$\boldsymbol{b}_i = (b_i^1, \ldots, b_i^l)^t$ は変換行列，I_l は l 次の単位行列，$\mathbf{1}$ はすべての要素が 1 の l 行 1 列のベクトルである．

中心化グラム行列[8),9)] G を次のように定義する．

$$G = MKM = \left(I_l - \frac{1}{l}\mathbf{1}\mathbf{1}^t\right) K \left(I_l - \frac{1}{l}\mathbf{1}\mathbf{1}^t\right)$$

K が正定値行列 $\Leftrightarrow G$ も正定値行列である．

部分空間での分散の和を最大化する制約つき最適化問題により B を求める．特徴空間での共分散行列 $\Sigma_\phi = \frac{1}{l}\Phi M \Phi^t$ から変換後の共分散行列 Σ_ψ は

$$\Sigma_\psi = A^t \Sigma_\phi A = \frac{1}{l} B^t G B$$

となる．したがって，最適化問題

$$\begin{aligned} &\textbf{Maximize} : tr(\Sigma_\psi) \\ &\quad = \quad tr\left(\tfrac{1}{l} B^t G^2 B\right) \\ &\textbf{Subject to} : A^t A = I_d \\ &\quad \Longleftrightarrow \quad B^t G B = I_d \end{aligned}$$

を解けばよい．

ラグランジュの未定乗数法により，Λ をラグランジュ乗数として，ラグラジアンの B での導関数を 0 とおいて

$$\frac{1}{l} G^2 B = GB\Lambda \tag{1}$$

$$B^t G B = I_d \tag{2}$$

を得る．この条件式を満たす B を求める．

この連立方程式の解は，G の固有値，固有ベクトルを求めることで得られる．G の i 個目の固有値を μ_i，固有値行列を D とし，対応する固有ベクトル行列を V とする．ただし $V^t V = I$ に正規化されているものとし，d 番目の固有値までは 0 以上の値を持つものとする．すると，式 (1)，(2) の解は $B = VD^{\frac{1}{2}}$ となる．ここで

$$\Sigma_\psi = \frac{1}{l} D = \Lambda$$

が成立する．

したがって，λ_i が部分空間の各基底空間の分散となり，各基底間は無相関となる．また，変換ベクトル $\boldsymbol{a}_i$ に対応する分散は λ_i となっていることも理解できる．固有値 μ_i と分散 λ_i の間には $\lambda_i = \frac{\mu_i}{l}$ の関係があり，固有値が 0 となる固有ベクトルに対して分散は 0 である．

2.3 カーネル主成分分析の問題点

カーネル主成分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

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{z})\right)$$

に対しこの問題を考える．この場合のカーネルパラメータは

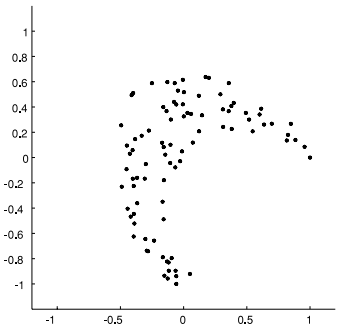

図 2 非線形データの例
Fig. 2 Example for non-linear dataset.

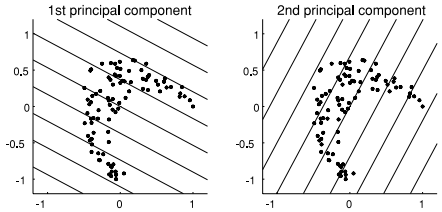

図 3 線形 PCA の第 1 主成分と第 2 主成分
Fig. 3 1st and 2nd principal components with linearPCA.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} \tag{3}$$

である．本手法は，パラメータを持つ一般的なカーネルも適用できる．

カーネル主成分分析の能力は，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を示す．まず線形主成分分析について考える．2 次元平面上に非線形成分を持つデータ（**図 2**）を用意し，線形主成分分析を行った結果を**図 3** に示す．このときのカーネルパラメータは $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ とする．このようなデータに対しては，線形主成分分析は 2 つの主成分軸しか得ることができないため，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 KPCA を行った結果を**図 4**，**図 5**，**図 6** に示す．図 4 のようにパラメータ σ が最適値よりも小さすぎる場合，データの近傍にしか特徴が現れず，未知データに対する情報を抽出できない．またガウス分布の分散

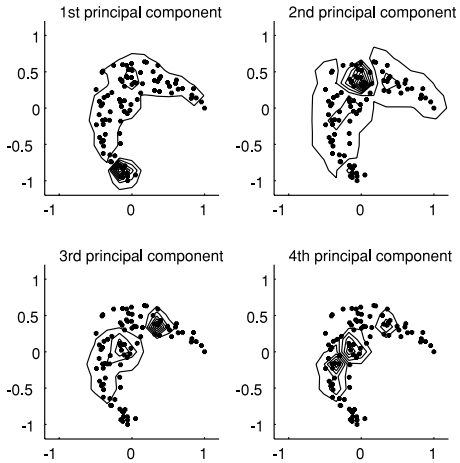

図 4 小さいカーネルパラメータでの KernelPCA の出力（$\sigma_{small} = 0.0625$）

Fig. 4 Outputs of KernelPCA with a small parameter value ($\sigma_{small} = 0.0625$).

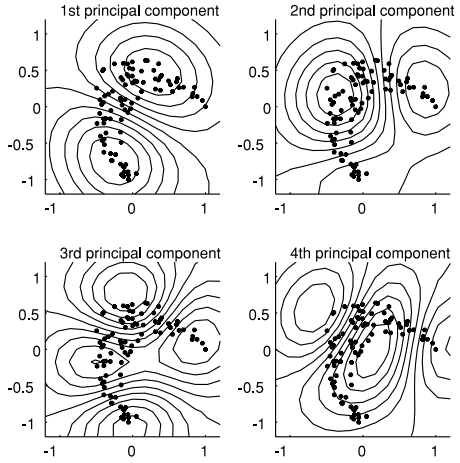

図 5 最適パラメータによる KernelPCA の出力（$\sigma_{opt} = 0.493$）

Fig. 5 Outputs of KernelPCA for the optimal parameter ($\sigma_{opt} = 0.493$).

が小さすぎるため，多くの点が自分自身の情報だけしか表現しておらず，情報圧縮に何ら関与しない状態になっている．逆に図 6 のように σ が最適値より大きすぎる場合は，出力が大まかになりすぎて細かい分布情報を抽出できなくなっている．この状態では，すべてのデータを元のデータに復元することは困難である．後述する手法で求めた最適パラメータで同様に計算した結果を図 5 の分布に示す．この図ではバランスよく第 4 主成分まで特徴をとらえており，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．

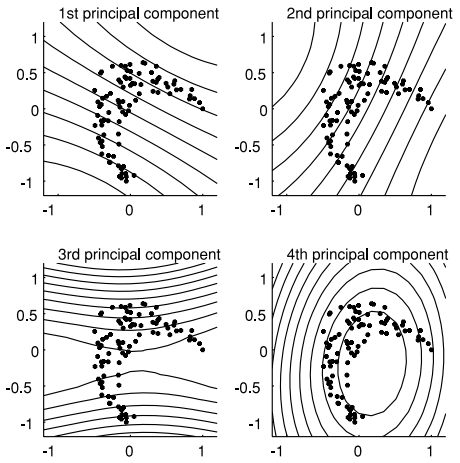

図 6 大きなパラメータでの KernelPCA の出力（$\sigma_{large} = 2.00$）

Fig. 6 Outputs of KernelPCA for a large parameter ($\sigma_{large} = 2.00$).

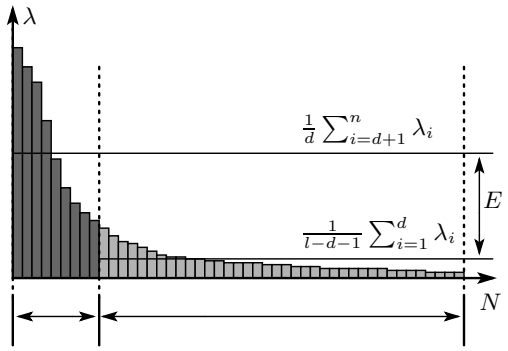

図 7 各主成分軸の分散と評価関数

Fig. 7 Eingenvalues and Cost function.

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まれることを述べた．パターン識別のための次元削除の観点から考えれば，残すべき主成分軸に分散が多く集まり，切り捨てるべき軸にはほぼ 0 に近い小さな分散が集中すればよい．なぜならば，このような写像空間であれば残すべき主成分が多くの圧縮された情報を持ち，切り捨てられる主成分によって損失する情報を最小限に抑えることができるからである．

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．本論文では，最適パラメータの評価関数として**図 7** のように，残す主成分軸（第 1 主成分から第 d 主成分）の分散の平均と，切り捨てる主成分軸（第 $d+1$ 主成分

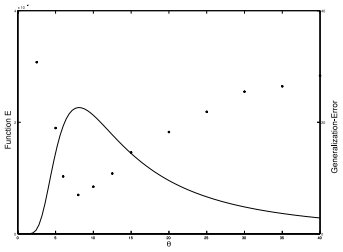

図 8 評価関数 $E(\theta)$
Fig. 8 Cost function $E(\theta)$.

から第 l 主成分）の分散の平均の差をとり，これを最大化させる．このとき評価関数は，カーネル関数のパラメータを $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_l]^t$ とし，第 i 主成分の分散を λ_i とすると

$$E(\boldsymbol{\theta}) = \frac{1}{d}\sum_{i=1}^{d}\lambda_i - \frac{1}{l-d-1}\sum_{i=d+1}^{l}\lambda_i \qquad (4)$$

となる．この関数を，後述する計算機実験で用いるデータ，また iris，wine，glass，vehicle などの自然界の識別データで計算し，評価関数のグラフを求めると典型的に**図 8** のような曲線が得られる．また，同図ではカーネルパラメータの値における，線形識別器の汎化誤差を点で表している（顔画像データ集合 “Gender” の結果を示す）．前述の自然界データに対し，評価関数が最大となるパラメータ σ のときに，汎化性能が最も良くなっていることが確認できる．

多くの問題では，このように $E(\boldsymbol{\theta})$ 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，このことを一般に示すことは難しい．

3.2 評価関数 $E(\boldsymbol{\theta})$ の微分

評価関数 $E(\boldsymbol{\theta})$ の $\boldsymbol{\theta}$ に関する偏導関数は

$$\begin{aligned}\frac{\partial E}{\partial \boldsymbol{\theta}} &= \frac{1}{d}\sum_{i=1}^{d}\frac{\partial \lambda_i}{\partial \boldsymbol{\theta}} - \frac{1}{l-d-1}\sum_{i=d+1}^{l}\frac{\partial \lambda_i}{\partial \boldsymbol{\theta}} \\ &= \frac{1}{dl}\sum_{i=1}^{d}\frac{\partial \mu_i}{\partial \boldsymbol{\theta}} - \frac{1}{l(l-d-1)}\sum_{i=d+1}^{l}\frac{\partial \mu_i}{\partial \boldsymbol{\theta}}\end{aligned}$$

となる．

μ_k は行列 $G(\boldsymbol{\theta})$ の固有値であり，固有方程式

$$\det(\mu_k I_l - G(\boldsymbol{\theta})) = 0 \qquad (5)$$

を満たす．μ_i は陽な形で表せないため直接微分することはできない．そこで陰関数定理を用いて評価する．ここで $f_k(\cdot)$ を式 (5) の解，すなわち $f_k(\boldsymbol{\theta}) = \mu_k$ とする．したがって $\det(\mu_k I_l - G)$ の転置余因子行列を C とすると，$\Sigma_{i=1}^{l} C_{ii} \neq 0$ ならば陰関数が微分可能で，

$$\begin{aligned}\frac{\partial f_k}{\partial \boldsymbol{\theta}} &= -\frac{\frac{\partial \det(\mu_k I_l - G)}{\partial \boldsymbol{\theta}}}{\frac{\partial \det(\mu_k I_l - G)}{\partial \mu_k}} \\ &= \frac{\Sigma_{i=1}^{l}\Sigma_{j=1}^{l} C_{ij}\frac{\partial G_{ij}}{\partial \boldsymbol{\theta}}}{\Sigma_{i=1}^{l} C_{ii}}\end{aligned}$$

と表すことができる．ここで G の $\boldsymbol{\theta}$ における偏微分は l 次正方行列となり

$$\frac{\partial G}{\partial \boldsymbol{\theta}} = \left(I_l - \frac{1}{l}\mathbf{1}\mathbf{1}^t\right)\frac{\partial K}{\partial \boldsymbol{\theta}}\left(I_l - \frac{1}{l}\mathbf{1}\mathbf{1}^t\right)$$

となる．ここで転置余因子行列 C の計算を簡略化するため，次の定理を用いる[8]．

Theorem3.1（$\mu_k I_l - G$ の転置余因子定理） 行列 G の固有値，固有ベクトルをそれぞれ μ_i，$\boldsymbol{v}_i$（$i = 1, \ldots, l$）とすれば，行列 $(\mu_k I_l - G)$ の転置余因子行列は

$$C = \left(\prod_{\substack{i=1\\ i\neq k}}^{l}(\mu_k - \mu_i)\right)\boldsymbol{v}_i\boldsymbol{v}_i^t$$

となる（証明は**付録**参照）．

定理 3.1 より

$$\begin{aligned}\frac{\partial f_k}{\partial \boldsymbol{\theta}} &= \frac{\left(\prod_{\substack{i=1\\ i\neq k}}^{l}(\mu_k - \mu_i)\right)\Sigma_{i=1}^{l}\Sigma_{j=1}^{l}\left[\boldsymbol{v}_k\boldsymbol{v}_k^t\right]_{ij}\frac{\partial G_{ij}}{\partial \boldsymbol{\theta}}}{\left(\prod_{\substack{i=1\\ i\neq k}}^{l}(\mu_k - \mu_i)\right)\Sigma_{i=1}^{l}[\boldsymbol{v}\boldsymbol{v}^t]_{ii}} \\ &= \frac{\left(\prod_{\substack{i=1\\ i\neq k}}^{l}(\mu_k - \mu_i)\right)\boldsymbol{v}_k^t\frac{\partial G_{ij}}{\partial \boldsymbol{\theta}}\boldsymbol{v}_k}{\prod_{\substack{i=1\\ i\neq k}}^{l}(\mu_k - \mu_i)} \\ &= \boldsymbol{v}_k^t\frac{\partial G}{\partial \boldsymbol{\theta}}\boldsymbol{v}_k\end{aligned}$$

これを $\frac{\partial \mu_k}{\partial \boldsymbol{\theta}}$ として用いる．

なお，Gaussian カーネルの微分係数は，

$$\frac{\partial K(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}} = (\boldsymbol{x} - \boldsymbol{z})^t(\boldsymbol{x} - \boldsymbol{z})K(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{\Sigma}^{-\frac{3}{2}}$$

より得られる．

評価関数 $E(\boldsymbol{\theta})$ を最大にする最適パラメータ $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ を求めるために，最急勾配法を適用する．カーネルパラメータが 1 次元（スカラー）の場合は，最急勾配法ではなく極値を探索する二分法（Bisection method）を適用する．この方法では $E(\boldsymbol{\theta})$ 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，パターン識別の性能比較を行う．

KPCA による前処理の後，ハードマージンを持つ

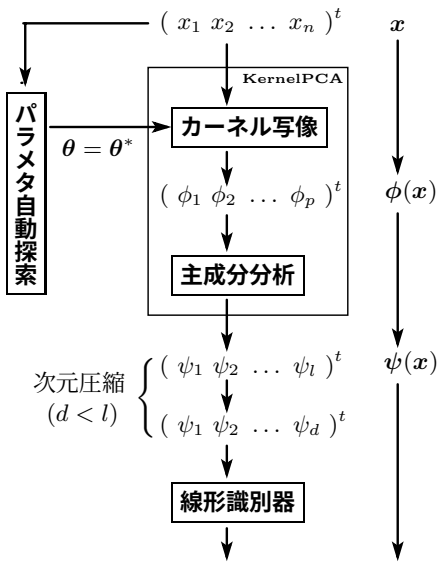

図 9 パターン識別への応用
Fig. 9 Application to Pattern recognition.

表 1 データセット “diabetes” の性能比較
Table 1 Result of the Dataset “diabetes.”

	最小汎化誤差 [%]	識別器入力次元数	実動学習時間 [s]
LSVM	38.19	8	223.6
NLSVM	26.37	684	1190
KPCA+LSVM（等方分散）	22.95	72	194.3
KPCA+LSVM（非等方分散）	22.02	72	217.1

表 2 データセット “heart” の性能比較
Table 2 Result of the Dataset “heart.”

	最小汎化誤差 [%]	識別器入力次元数	実動学習時間 [s]
LSVM	35.73	13	119.6
NLSVM	24.32	243	548.5
KPCA+LSVM（等方分散）	16.98	44	91.39
KPCA+LSVM（非等方分散）	16.21	44	103.0

線形サポートベクトルマシン（LSVM）を識別器として用いる．**図 9** にこの学習法を図示する．提案する学習法全体としてはモデル選択に依存する主成分数以外のパラメータを必要としない構成となる．

学習データは Artificial Intelligence and Computer Science Laboratory ⋆1 より “heart”，“diabetes” のデータを取得し，提案手法と原特徴に対する線形サポートベクトルマシン（LSVM），非線形のサポートベクトルマシン（NLSVM）との識別性能の比較を行った．

それぞれのデータに対しては，式 (3) のカーネルパラメータ $\boldsymbol{\Sigma}$ について，パラメータが 1 次元（スカラー）の等方分散（$\sigma_1=\sigma_2=\cdots=\sigma_n$）

$$\boldsymbol{\Sigma}=\sigma^2\boldsymbol{I}$$

の場合と，パラメータが多次元の非等方分散（実験では，クラス別に異なる分散であるとした）

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & & & 0\\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2\end{bmatrix}$$

のそれぞれの場合について計算機実験を行った．

使用プログラムとして，SVM and Kernel Methods Matlab Toolbox ⋆2 を用いた．演算においては，CPU：Pentium 4 3.20 GHz，メモリ：2.0 GB，OS：Windows の計算機を使用した．

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

4.1 提案手法の評価

パラメータ探索法における特徴選択や次元数の決定には，様々な議論が行われている[2]．本実験では，十分大きな $d=d_L$ に対し最適な σ の値が d に対し鈍感になることを利用し，初期値としてあらかじめ大きな d でカーネルパラメータ $\boldsymbol{\Sigma}_d$ を求めてから，そのパラメータにおいて評価関数 $E(\boldsymbol{\theta})$ が最大となる d_0 を決定する方法をとった．ここでは大きな d_L として，KPCA から出力された写像データの次元数の 50%から 75%の間でランダムに定めた．

LSVM と NLSVM についてはそれぞれ 10 回の学習を行い，その中での最も性能の良い値を選択した．特に NLSVM では，Gaussian カーネルのカーネルパラメータ σ とソフトマージンのパラメータである C を探索する必要があるため，$\sigma=[2,4,\ldots,10]$，$C=[2,4,\ldots,10]$ の 25 の組合せについてそれぞれランダムに 10 点のパラメータ探索を行った．どちらの学習機械に対しても，汎化誤差の計算には 10-fold cross-validation を 5 回行った．

データセット “heart”，“diabetes” に対する識別器の性能比較を**表 1**，**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実行時間を示している．

NLSVM は広い範囲で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．

パラメータ探索を含めた学習時間を考えるとき，提案手法は小さな計算規模で，SVM よりも高い汎化性能を実現することができる．

図 10 入力の顔画像データの例
Fig. 10 Example for input face data.

表 3 ジェンダー識別の性能比較
Table 3 Result of the Dataset "Gender."

	最小汎化誤差 [%]	識別器入力次元数	実動学習時間 [s]
NLSVM（50 点探索）	10.33	270	572.4
NLSVM（256 点探索）	5.602	270	2828
KPCA+LSVM	6.814	168	148.7
KPCA+LSVM（第 1 主成分削除）	6.340	168	157.6

また，多次元のカーネルパラメータ（非等分散）の場合についてはクラスごとに異なるカーネルパラメータを用いた．この方法では，対象データの性質をより詳しくとらえることができるので，等方パラメータの場合と比べて汎化性能が改善された．

4.2 ジェンダー識別

本手法を男女 2 クラスのジェンダー識別問題へ適用し，SVM との比較を行った．

入力データは，サイズ 46×44 の 300 枚の男女顔画像[⋆1]（**図 10**）を用意した．

この実験に対しては，NLSVM のパラメータである σ と C を

$$\sigma = [e^{-5}, e^{-4}, \ldots, e^{0}, \ldots, e^{9}, e^{10}]$$
$$C = [e^{-5}, e^{-4}, \ldots, e^{0}, \ldots, e^{9}, e^{10}]$$

の 256 点の組合せから，ランダムに 50 点でのパラメータ探索を行った場合と，すべての範囲で探索した場合の性能を出力し，その中でのそれぞれ最も性能の良い値を選択した．

汎化誤差の計算には 10-fold cross-validation を 5 回ずつ行った．

顔画像データセット "Gender" に対する識別器の性能比較を**表 3** に示す．

この問題に対しても，提案手法は，パラメータ探索回数が小さく抑えられた結果，小規模の計算量で良い結果を出すことができた．

⋆1 本論文に使用した顔画像データは財団法人ソフトピアジャパン研究開発部地域結集型共同研究推進室から使用許諾を受けたものであり，権利者に無断で複写，利用，配布などを行うことは禁じられています．

NLSVM に関してはパラメータ探索をすべての範囲で探索したとき，最小汎化誤差は提案手法よりも下げることができた．ただし，この結果を出すためには 2828 秒かかり事前実験の計算量が膨大となる．

またこの問題では，データの性質により第 1 主成分を削除して識別した方が，良い結果を得られることが分かる．これは顔画像データの第 1 主成分が，男女を問わず「顔」の特徴を表し，ジェンダー識別には影響を与えないためと考えられる（ただし，顔検出問題では第 1 主成分が重要だと考えられる）．逆に第 2 主成分まで削除して画像識別をした結果，識別性能は大幅に下がることも確認した．この事実は，第 2 主成分に「男女を決定する強い因子」が含まれていることを意味する．

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，カーネル主成分分析に基づいた識別器を提案した．LSVM と組み合わせることで，非線形の SVM よりも高速で同等の性能を持つ識別器を実現することができた．

提案手法のパラメータ自動決定法は他のパラメータを持つカーネルマシンについても応用できる．マルチクラス SVM については，各クラス間ごとに別のカーネルパラメータを使用して KPCA を行うことで，より精度の高いマルチクラス識別が可能となる．

また，今後は変分法を応用することで，カーネルそのものの選択などへの拡張が考えられる．

参考文献

1) Blanchard, G., Bousquetart, O. and Zwald, L.: Statistical Properties of Kernel Principal Component Analysis, *Proc. 17th Conf. on Learning Theory*, pp.594–608 (2004).
2) Blanchard, G., Massart, P., Vert, R. and Zwald, L.: Kernel Projection machine: A new tool for pattern recognition, *Proc. 18th Neural Information Processing System*, pp.1649–1656 (2004).
3) Zwald, L. and Blanchard, G.: On the convergence of eigenspaces in kernel principal components analysis, *Advances in Neural Inf. Proc. Systems*, Vol.18 pp.1649–1656 (2006).
4) Schölkopf, B., Smola, A. and Muller, K.: Kernel Principal Component Analysis, *Advances in Kernel Methods*, pp.327–353 (1998).
5) Shawe-Taylor, J., Williams, C., Cristianini, N. and Kandola, J.: On the eigenspectrum of the Gram matrix and the generalisation error of

Kernel PCA, *IEEE Trans. Information Theory*, Vol.7 No.51 pp.2510–2522 (2005).
6) Vapnik, N.: *Statistical Learning Theory*, Wiley (1998).
7) Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press (2000).
8) Nogayama, T., Takahashi, H. and Muramatsu, M.: Genetalization of kernel PCA and Automatic Paremeter Tuning, *The 8th Australian and New Zealand Intelligent Information Systems Conference*, Macquarie University, Sydney, Australia, pp.173–178 (2003).
9) 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).

付　　録

$\mu_i I - G$ の余因子行列

対称行列 $\mu_i I - G$ の余因子行列を求める．ただし μ_i は G の固有値である．通常対称行列 A の余因子行列 $\tilde{A}$ を求めるときは

$$\tilde{A} = \det(A)A^{-1}$$

を用いるが，$\mu_i I - G$ には逆行列が存在しないためこの方法は使えない．そこで $F(\mu) = \mu I - G$ を考え，μ を固有値に近づけることを考える．$F(\mu)$ は μ が G の固有値のときのみ行列式がゼロとなるため，$\mu = \mu_i + \epsilon$ のときは行列式は必ず非ゼロとなる．そのため $\tilde{F}(\mu)$ は計算可能である．いま G と固有値解析済みとし，その固有値を μ_i，固有ベクトルを $\boldsymbol{v}_i$ とする．すると $F(\mu_i)$ は

$$\begin{aligned} F(\mu) &= \mu I - G \\ &= \mu I - \sum_{i=1}^{l} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^t \\ &= \sum_{i=1}^{l} (\mu - \mu_i) \boldsymbol{v}_i \boldsymbol{v}_i^t \end{aligned}$$

となる．$\mu \neq \mu_i$ ならば逆行列は存在し，

$$F^{-1}(\mu) = \sum_{i=1}^{l} \frac{1}{(\mu - \mu_i)} \boldsymbol{v}_i \boldsymbol{v}_i^t$$

で表せる．固有値でない μ に対する $F(\mu)$ の余因子行列は

$$\begin{aligned} \tilde{F}(\mu) &= \det(F(\mu))F(\mu)^{-1} \\ &= \det(\mu I - G) \sum_{i=1}^{l} \frac{1}{(\mu - \mu_i)} \boldsymbol{v}_i \boldsymbol{v}_i^t \\ &= \prod_{j=1}^{l} (\mu - \mu_j) \sum_{i=1}^{l} \frac{1}{(\mu - \mu_i)} \boldsymbol{v}_i \boldsymbol{v}_i^t \\ &= \sum_{i=1}^{l} \prod_{\substack{j=1 \\ j \neq i}}^{l} (\mu - \mu_j) \boldsymbol{v}_i \boldsymbol{v}_i^t \end{aligned}$$

となる．μ を固有値に近づけ，余因子行列の解を得る．

$$\begin{aligned} \tilde{F}(\mu_i) &= \lim_{\mu \to \mu_i} \tilde{F}(\mu) \\ &= \prod_{\substack{j=1 \\ j \neq i}}^{l} (\mu - \mu_i) \boldsymbol{v}_i \boldsymbol{v}_i^t \end{aligned}$$

通常，サイズ $(l \times l)$ の行列の余因子行列を求めるにはサイズ $((l-1) \times (l-1))$ の行列式をサイズ $l \times l$ 通り求める必要がある．行列式はサイズ $O(l^3)$ なので，余因子行列の計算量は $O(l^5)$ となる．しかしこの結果を用いると l に関して余因子行列の計算量は $O(l)$ となり，大幅な計算量の削減が可能となる．

(平成 19 年 2 月 2 日受付)
(平成 19 年 8 月 4 日再受付)
(平成 19 年 10 月 27 日採録)

関口　涼平（学生会員）
昭和 56 年生．平成 19 年電気通信大学大学院情報通信工学専攻博士前期課程修了．同年 4 月より電気通信大学大学院情報通信工学専攻博士後期課程在籍．

高橋　治久
昭和 27 年生．昭和 50 年電気通信大学電気通信学部通信工学科卒業．昭和 52 年同大学大学院修士課程修了．昭和 55 年大阪大学大学院工学研究科博士後期課程修了．博士（工学）．同年豊橋技術科学大学助手．昭和 61 年電気通信大学講師を経て現在同教授．現在形式ニューラルネットワーク，学習等の研究に従事．昭和 59 年度電子情報通信学会学術奨励賞受賞，電子情報通信学会，国際ニューラルネットワーク学会各会員．

堀田 一弘（正会員）

昭和 50 年生．平成 9 年埼玉大学工学部情報工学科卒業．平成 11 年同大学大学院理工学研究科博士前期課程修了．平成 14 年同大学院理工学研究科博士後期課程修了．博士（工学）．平成 11〜14 年日本学術振興会特別研究員．平成 14〜19 年電気通信大学電気通信学部情報通信工学科助手．平成 19 年より同大学助教．平成 19 年より理化学研究所客員研究員．平成 19 年より東京大学人工物工学センター協力研究員．パターン認識，コンピュータビジョンの研究に従事．IEEE Computer Society，電子情報通信学会，日本顔学会各会員．