2004－CVIM－145　(25)
2004／9／11

カーネル学習法とその画像認識への応用

栗田 多喜夫, 西田健次
産業技術総合研究所 脳神経情報研究部門

概要　サポートベクターマシンは、現在知られている多くのパターン認識手法の中でも認識性能の優れた手法であると考えられている。サポートベクターマシンがこのような優れた認識性能を発揮できるのは、カーネルトリックを用いて非線形の識別関数を構成できるようにし、しかも、未学習データに対しても高い認識性能（汎化性能）を得るための工夫を導入したためである。本論文では、サポートベクターマシンを中心に、カーネル学習法の考え方と汎化性能を向上させるための工夫について概説し、画像認識への応用例について紹介する。

Kernel Methods and their Application for Image Understanding

Takio Kurita and Kenji Nishida
Neurosceince Research Institute,
National Institute of Advanced Indastrial Science and Technology

Abstract　Support vector machine (SVM) has been extended to build up nonlinear classifier using the kernel trick. It is recognized as one of the best models for two class classification among the many methods currently known because it is devised to obtain high performance for unlearned data. This paper reviews kernel methods centering on the SVM and introduces some examples of applications for image understanding.

1　はじめに

サポートベクターマシン (Support Vector Machine, SVM)[1, 2, 3] は、現在知られている様々なパターン認識手法の中でも最も認識性能の優れた学習モデルの一つと考えられている。サポートベクターマシンを用いると、カーネルトリックにより非線形の識別関数が構成できる。しかも「マージン最大化」という基準を用いることで、未学習サンプルに対しても高い認識性能（汎化性能）を得ることができる。また、現在では、サポートベクターマシンのためのソフトウェアツールも手軽に利用できる。そのため、訓練用のデータを準備するだけで、誰でも、比較的簡単に、サポートベクターマシンを用いた非線形の識別器を実現することができる。画像認識の分野でも、顔の検出や歩行者の検出等の対象認識 [4, 5, 6]、文字認識 [7] 等に利用され、高い認識性能が得られることが報告されている。

カーネルトリックは、サポートベクターマシンだけでなく、多変量データ解析等の線形モデルで表される手法を非線形に拡張するためにも利用することができる [8, 9, 10]。すでに、カーネル主成分分析、カーネル判別分析、カーネル部分空間法、カーネル正準相関分析等が提案されている。

本稿では、サポートベクターマシンを中心に、カーネル学習法 [9] の考え方と汎化性能を向上させるためのの工夫について概説し、その画像認識への応用例について紹介する。

2　カーネル学習法

2.1　サポートベクターマシン

サポートベクターマシンは、1960 年代に Vapnik 等が考案した Optimal Separating Hyperplane を起源とし、1990 年代になってカーネル学習法と組み合わせた非線形の識別手法へと拡張された。カーネルトリックにより非線形の識別関数が構成できるように拡張したサポートベクターマシンは、現在知られている手法の中でも最もパターン認識性能の優秀な学習モデルの一つである。ただし、サポートベクターマシンは、基本的には 2 つのクラスを識別する識別器を構成するための学習法であり、文字認識などの多クラスの識別器を構成するためには、複数のサポートベクターマシンを組み合わせるなどの工夫が必要となる。ここでは、まず、

サポートベクターマシンを中心にカーネル学習法を用いて訓練サンプルから非線形の識別器を構成する方法について概説する。一般に、カーネル学習法を用いて学習された識別器が、訓練サンプルに含まれていない未学習データに対しても高い識別性能を発揮できるためには、汎化能力を向上させるための工夫が必要である。サポートベクターマシンでは、「マージン最大化」という基準を用いることでこれを実現している。これは、結果的には、不必要なパラメータが値を持たないように学習の評価基準にペナルティ項を追加する shrinkage 法の一種と考えることができる。

サポートベクターマシンは、単純パーセプトロン（ニューロンのモデルとして最も単純な線形しきい素子）を用いて、2 クラスのパターン識別器を構成する手法である。ここでは、サポートベクターマシンの一般的な定義に従って、線形しきい素子として、入力特徴ベクトルに対し、識別関数（線形識別関数）

$$y = \mathrm{sign}(\boldsymbol{w}^T\boldsymbol{x} - h) \tag{1}$$

により 2 値 (± 1) を出力するモデルを用いて説明する。ここで、$\boldsymbol{w}$ はシナプス荷重に対応するパラメータであり、h はしきい値である。また、関数 $\mathrm{sign}(u)$ は、$u > 0$ のとき 1 をとり、$u \leq 0$ のとき -1 をとる符号関数である。このモデルは、入力ベクトルとシナプス荷重の内積がしきい値を超えれば 1 を出力し、超えなければ -1 を出力する。これは、幾何学的には、識別平面により、入力特徴空間を 2 つに分けることに相当する。今、2 つのクラスを C_1,C_2 とし、各クラスのラベルを 1 と -1 に数値化しておくとする。また、訓練サンプル集合として、N 個の特徴ベクトル $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ と、それぞれのサンプルに対する正解のクラスラベル $t_1, \ldots, t_N$ が与えられているとする。また、この訓練サンプル集合は、線形分離可能であるとする。すなわち、線形しきい素子のパラメータをうまく調整することで、訓練サンプル集合を誤りなく分けることができると仮定する。

訓練サンプル集合が線形分離可能であるとしても、一般には、訓練サンプル集合を誤りなく分けるパラメータは一意には決まらない。サポートベクターマシンでは、訓練サンプルをすれすれに通るのではなく、なるべく余裕をもって分けるような識別平面が求められる。具体的には、最も近い訓練サンプルとの余裕をマージンと呼ばれる量で測り、マージンが最大となるような識別平面を求める。もし、訓練サンプル集合が線形分離可能なら、

$$t_i(\boldsymbol{w}^T\boldsymbol{x}_i - h) \geq 1, \quad i = 1, \ldots, N \tag{2}$$

を満たすようなパラメータが存在する。これは、H1: $\boldsymbol{w}^T\boldsymbol{x} - h = 1$ と H2: $\boldsymbol{w}^T\boldsymbol{x} - h = -1$ の 2 枚の超平面で訓練サンプルが完全に分離されており、2 枚の超平面の

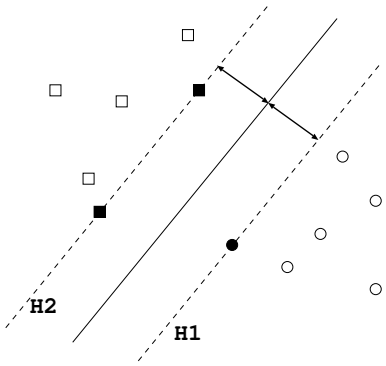

図 1: 線形しきい素子の分離超平面とマージン (◯がクラス 1 のサンプルで、□がクラス-1 のサンプルを示す。●と■はサポートベクターを示す。)

間にはサンプルがひとつも存在しないことを示している。線形識別関数の性質についての説明で触れたように、識別平面とこれらの超平面との距離（マージンの大きさ）は、$\frac{1}{||\boldsymbol{w}||}$ となる。したがって、マージンを最大とするパラメータ $\boldsymbol{w}$ と h を求める問題は、結局、制約条件

$$t_i(\boldsymbol{w}^T\boldsymbol{x}_i - h) \geq 1, \quad (i = 1, \ldots, N) \tag{3}$$

の下で、目的関数

$$L(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||^2 \tag{4}$$

を最小とするパラメータを求める問題と等価になる。この最適化問題は、数理計画法の分野で 2 次計画問題として知られており、さまざまな数値計算法が提案されている。ここでは、双対問題に帰着して解く方法を紹介する。まず、Lagrange 乗数 $\alpha_i(\geq 0), i = 1, \ldots, N$ を導入し、目的関数を

$$L(\boldsymbol{w}, h, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{N} \alpha_i\{t_i(\boldsymbol{w}^T\boldsymbol{x}_i - h) - 1\} \tag{5}$$

と書き換える。パラメータ $\boldsymbol{w}$ および h に関する偏微分から停留点では、

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i t_i \boldsymbol{x}_i \tag{6}$$

$$0 = \sum_{i=1}^{N} \alpha_i t_i \tag{7}$$

という関係が成り立つ。これらを上の目的関数の式に代入すると、制約条件、

$$\sum_{i=1}^{N} \alpha_i t_i = 0 \tag{8}$$

$$0 \leq \alpha_i, \quad i = 1, \ldots, N \tag{9}$$

の下で、目的関数

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j t_i t_j \boldsymbol{x}_i^T \boldsymbol{x}_j \qquad (10)$$

を最大とする双対問題が得られる。これは、Lagrange 乗数 α_i (≥ 0), $i = 1, \ldots, N$ に関する最適化問題となる。その解で α_i^* が 0 でない、すなわち、$\alpha_i^* > 0$ となる訓練サンプル $\boldsymbol{x}_i$ は、先の 2 つの超平面 $\boldsymbol{w}^T\boldsymbol{x} - h = 1$ か $\boldsymbol{w}^T\boldsymbol{x} - h = -1$ のどちらかにのっている。このことから、α_i^* が 0 でない訓練サンプル $\boldsymbol{x}_i$ のことを「サポートベクター」と呼んでいる。これが、サポートベクターマシンの名前の由来である。直感的に理解できるように、一般には、サポートベクターは、もとの訓練サンプル数に比べてかなり少ない。つまり、沢山の訓練サンプルの中から小数のサポートベクターを選び出し、それらのみを用いて線形しきい素子のパラメータが決定されることになる。

実際、双対問題の最適解 $\alpha_i^*(i \geq 0)$、および停留点での条件式から、最適なパラメータ $\boldsymbol{w}^*$ は、

$$\boldsymbol{w}^* = \sum_{i \in S} \alpha_i^* t_i \boldsymbol{x}_i \qquad (11)$$

となる。ここで、S はサポートベクターに対応する添え字の集合である。また、最適なしきい値 h^* は、2 つの超平面 $\boldsymbol{w}^T\boldsymbol{x} - h = 1$ か $\boldsymbol{w}^T\boldsymbol{x} - h = -1$ のどちらかにのっているという関係を利用して求めることができる。すなわち、任意のサポートベクター $\boldsymbol{x}_s, s \in S$ から

$$h^* = \boldsymbol{w}^{*T}\boldsymbol{x}_s - t_s \qquad (12)$$

により求まる。

また、最適な識別関数を双対問題の最適解 $\alpha_i^*(i \geq 0)$ を用いて表現すると

$$\begin{aligned} y &= \mathrm{sign}(\boldsymbol{w}^{*T}\boldsymbol{x} - h^*) \\ &= \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i \boldsymbol{x}_i^T \boldsymbol{x} - h^*) \end{aligned} \qquad (13)$$

となる。すなわち、$\alpha_i^* = 0$ となる多くの訓練サンプルを無視し、$\alpha_i^* > 0$ となる識別平面に近い少数の訓練サンプルのみを用いて識別関数が構成される。ここで、重要な点は、「マージン最大化」という基準から自動的に識別平面付近の少数の訓練サンプルのみが選択されたことであり、その結果として、未学習データに対してもある程度良い識別性能が維持できていると解釈できる。 すなわち、サポートベクターマシンは、マージン最大化という基準を用いて、訓練サンプルを撰択することで、モデルの自由度を抑制するようなモデル撰択が行われていると解釈できる。

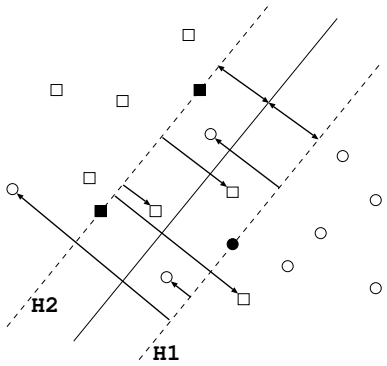

図 2: ソフトマージン (◯がクラス 1 のサンプルで、□がクラス-1 のサンプルを示す。●と■はサポートベクターを示す。)

2.1.1 ソフトマージン

上述のサポートベクターマシンは、訓練サンプルが線形分離可能な場合についての議論であるが、パターン認識の実問題で線形分離可能な場合は稀である。したがって、実際的な課題にサポートベクターマシンを使うには、さらなる工夫が必要である。まず考えられるのは、多少の識別誤りは許すように制約を緩める方法である。これは、「ソフトマージン」と呼ばれている。

ソフトマージン法では、マージン $\frac{1}{||\boldsymbol{w}||}$ を最大としながら、図 2 に示すように、幾つかのサンプルが超平面 H1 あるいは H2 を越えて反対側に入ってしまうことを許す。反対側にどれくらい入り込んだかの距離を、パラメータ $\xi_i(\geq 0)$ を用いて、$\frac{\xi_i}{||\boldsymbol{w}||}$ と表すとすると、その和

$$\sum_{i=1}^{N} \frac{\xi_i}{||\boldsymbol{w}||} \qquad (14)$$

はなるべく小さいことが望ましい。これらの条件から最適な識別面を求める問題は、制約条件

$$\xi_i \geq 0, \quad t_i(\boldsymbol{w}^T\boldsymbol{x}_i - h) \geq 1 - \xi_i, \quad (i = 1, \ldots, N) \qquad (15)$$

の下で、目的関数

$$L(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 + \gamma \sum_{i=1}^{N} \xi_i \qquad (16)$$

を最小とするパラメータを求める問題に帰着される。ここで、あらたに導入したパラメータ γ は、第 1 項のマージンの大きさと第 2 項のはみ出しの程度とのバランスを決める定数である。

この最適化問題の解法は、基本的には線形分離可能な場合と同様にふたつの制約条件に対して、Lagrange 乗数 α_i、および、ν_i を導入し、目的関数を

$$L(\boldsymbol{w}, h, \boldsymbol{\alpha}, \boldsymbol{\nu}) = \frac{1}{2}||\boldsymbol{w}||^2 + \gamma \sum_{i=1}^{N} \xi_i$$

$$-\sum_{i=1}^{N}\alpha_i\{t_i(\boldsymbol{w}^T\boldsymbol{x}_i-h)-(1-\xi_i)\}$$
$$-\sum_{i=1}^{N}\nu_i\xi_i \tag{17}$$

と書き換える。パラメータ $\boldsymbol{w}$、h、ν_i に関する偏微分を 0 とする停留点では、

$$\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i t_i \boldsymbol{x}_i \tag{18}$$
$$0 = \sum_{i=1}^{N}\alpha_i t_i \tag{19}$$
$$\alpha_i = \gamma-\nu_i \tag{20}$$

という関係が成り立つ。これらを目的関数の式に代入すると、制約条件

$$\sum_{i=1}^{N}\alpha_i t_i = 0 \tag{21}$$
$$0 \le \alpha_i \le \gamma, \quad i=1,\ldots,N \tag{22}$$

の下で、目的関数

$$L_D(\boldsymbol{\alpha})=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j t_i t_j \boldsymbol{x}_i^T\boldsymbol{x}_j \tag{23}$$

を最大とする双対問題が得られる。線形分離可能な場合には、最適解 α_i^* の値により、平面 H1 および H2 上の訓練サンプル (サポートベクター) とそれ以外のサンプルに分類されたが、ソフトマージンの場合には、さらに、H1 および H2 をはさんで反対側にはみ出すサンプルが存在する。それらは、同様に、最適解 α_i^* の値により区別することができる。具体的には、$\alpha_i^*=0$ なら、平面 H1 あるいは H2 の外側に存在し、学習された識別器によって正しく識別される。また、$0<\alpha_i^*<\gamma$ の場合には、対応するサンプルは、ちょうど平面 H1 あるいは H2 の上に存在するサポートベクターとなり、これも正しく識別される。$\alpha_i^*=\gamma$ の場合には、対応するサンプルはサポートベクターとなるが、$\xi_i\neq 0$ となり、平面 H1 あるいは H2 の内側に存在することになる。

2.2 カーネルトリック

本質的に非線形な問題に対応するための方法として、特徴ベクトルを非線形変換して、その空間で線形の識別を行う「カーネルトリック」と呼ばれている方法が知られている。この方法を用いることでサポートベクターマシンの性能が飛躍的に向上した。それがサポートベクターマシンを有名にした大きな要因であると考えられる。

一般に、線形分離可能性はサンプル数が大きくなればなるほど難しくなり、逆に、特徴空間ベクトルの次元が大きくなるほど易しくなる。例えば、特徴ベクトルの次元が訓練サンプルの数よりも大きいなら、どんなラベル付けに対しても線形分離可能である。しかし、高次元への写像を行うと、次元の増加に伴い汎化能力が落ちてしまう。また、難しい問題を線形分離可能にするためには、訓練サンプルと同程度の大きな次元に写像しなければならないので、結果的に膨大な計算量が必要となってしまう。

今、元の特徴ベクトル $\boldsymbol{x}$ を非線形の写像 $\phi(\boldsymbol{x})$ によって変換し、その空間で線形識別を行うことを考えてみよう。例えば、写像 ϕ として、入力特徴を 2 次の多項式に変換する写像を用いるとすると、写像した先で線形識別を行うことは、もとの空間で 2 次の識別関数を構成することに対応する。一般には、こうした非線形の写像によって変換した特徴空間の次元は大きくなりがちである。しかし、サポートベクターマシンの場合には、幸いにも、目的関数 L_D や識別関数が入力パターンの内積のみに依存した形になっており、内積が計算できれば最適な識別関数を構成することが可能である。つまり、もし非線形に写像した空間での二つの要素 $\phi(\boldsymbol{x}_1)$ と $\phi(\boldsymbol{x}_2)$ の内積が

$$\phi(\boldsymbol{x}_1)^T\phi(\boldsymbol{x}_2)=K(\boldsymbol{x}_1,\boldsymbol{x}_2) \tag{24}$$

のように、入力特徴 $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ のみから計算できるなら、非線形写像によって変換された特徴空間での特徴 $\phi(\boldsymbol{x}_1)$ や $\phi(\boldsymbol{x}_2)$ を陽に計算する代わりに、$K(\boldsymbol{x}_1,\boldsymbol{x}_2)$ から最適な非線形写像を構成できる。ここで、このような K のことをカーネルと呼んでいる。このように高次元に写像しながら、実際には写像された空間での特徴の計算を避けて、カーネルの計算のみで最適な識別関数を構成するテクニックのことを「カーネルトリック」と呼んでいる。

実用的には、K は計算が容易なものが望ましい。例えば、多項式カーネル

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=(1+\boldsymbol{x}_1^T\boldsymbol{x}_2)^p \tag{25}$$

Gauss カーネル

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\exp\left(\frac{-||\boldsymbol{x}_1-\boldsymbol{x}_2||^2}{2\sigma^2}\right) \tag{26}$$

シグモイドカーネル

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\tanh\left(a\boldsymbol{x}_1^T\boldsymbol{x}_2-b\right) \tag{27}$$

などが使われている。

2.2.1 カーネルサポートベクターマシン

式 (10) や式 (23) の目的関数 L_D は、

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j t_i t_j\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j t_i t_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad (28)$$

のように内積をカーネルで置き換えた形に書ける。また、式 (13) から最適な識別関数は、

$$\begin{aligned} y &= \mathrm{sign}(\boldsymbol{w}^{*T}\phi(\boldsymbol{x}) - h^*) \\ &= \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) - h^*) \\ &= \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i K(\boldsymbol{x}_i, \boldsymbol{x}) - h^*) \end{aligned} \quad (29)$$

のようにサポートベクターマシンの内積をカーネルで置き換えた形に書ける。ここで、この式にシグモイドカーネルを代入すると、いわゆる 3 層の多層パーセプトロンと同じ構造となる。また、Gauss カーネルを代入すると、Radial Basis Function (RBF) ネットワークと同じ構造になり、構造的には従来のニューラルネットワークと同じになる。しかし、カーネルトリックを用いて非線形に拡張したサポートベクターマシンでは、中間層から出力層への結合荷重のみが学習により決定され、前段の入力層から中間層への結合荷重は固定で、訓練データから機械的に求められる。また、中間層のユニット数が非常に大きく、訓練サンプル数と同じになる。つまり、カーネルトリックを用いて非線形に拡張したサポートベクターマシンでは、入力層から出力層への結合荷重を適応的に学習により求めない代わりにあらかじめ中間層に非常に多くのユニットを用意することで複雑な非線形写像を構成しようとする。

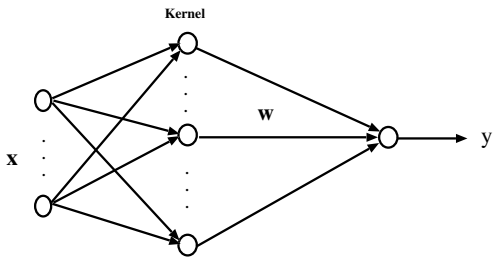

図 3: サポートベクターマシンの構造

カーネル学習法と組み合わせて非線形の識別関数が構成できるように拡張することで、カーネルサポートベクターマシンは、現在知られている多くのパターン認識手法の中でも最もパターン認識性能の良い学習モデルのひとつと考えられている。図 4 に非線形のサポートベクターマシを用いて構成した識別器の例を示す。

2.3 カーネル判別分析

カーネルトリックを用いると、多変量データ解析等の線形モデルで表される手法を非線形に拡張することができる [8, 9, 10]。すでに、カーネル主成分分析、カーネル判別分析、カーネル部分空間法、カーネル正準相関分析等が提案されている。ここでは、カーネル判別分析について簡単に紹介する。

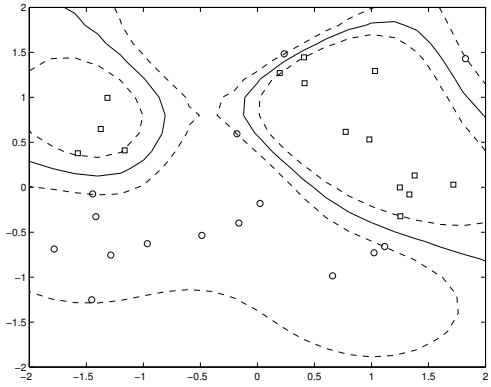

図 4: サポートベクターマシンによる識別例 (◯と□の識別課題であり、実線が識別境界である。点線上にあるサンプルはサポートベクターと呼ばれている。)

今、特徴ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_M)^T$ を K 個のクラスに識別する課題について考えよう。この時、線形判別分析は、特徴ベクトルの空間から線形判別写像

$$\boldsymbol{y} = A^T \boldsymbol{x} \quad (30)$$

により、写された空間でのクラス内の平均的な散らばりがなるべく小さく、クラス間のちらばりがなるべく大きくなるような係数行列 $A = [a_{ij}]$ を求める問題として定式化される。判別写像の良さの評価としては、判別基準が用いられる。判別基準は、いくつかの等価な基準が知られているが、ここでは、

$$J = \boldsymbol{tr}(\hat{\Sigma}_T^{-1} \hat{\Sigma}_B) \quad (31)$$

を用いるものとする。ここで、$\hat{\Sigma}_T$ および $\hat{\Sigma}_B$ は、それぞれ、新特徴 $\boldsymbol{y}$ 上で定義された分散共分散行列およびクラス間平均分散共分散行列である。

判別基準 (31) を最大とする最適な係数行列 A は、固有値問題

$$\Sigma_B A = \Sigma_T A \Lambda \quad (A^T \Sigma_T A = I) \quad (32)$$

の解として求まる。ここで、Λ は固有値を対角要素とする対角行列である。また、Σ_T および Σ_B は、それぞれ、入力特徴 $\boldsymbol{x}$ 上で定義された分散共分散行列およびクラス間平均分散共分散行列であり、

$$\begin{aligned} \Sigma_T &= \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^T \\ \Sigma_B &= \sum_{k=1}^{K} \omega_k (\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \end{aligned} \quad (33)$$

のように定義される。ここで、$\omega_k = N_k/N$, $\bar{\boldsymbol{x}}_k$, および $\bar{\boldsymbol{x}}_T$ は、それぞれ、クラス C_k の先見確率、$\boldsymbol{x}$ のクラス C_k の平均ベクトル、および、全平均ベクトルである。

線形判別分析では、式 (30) のように線形写像を構成したが、カーネル判別分析では、非線形の変換 $\Phi(\boldsymbol{x})$ により特徴を抽出し、それらの線形結合で判別写像を構成する。ここでは、簡単のため 1 次元の判別特徴を抽出する場合について考えよう。すなわち、

$$y = \boldsymbol{a}^T \Phi(\boldsymbol{x}) \tag{34}$$

のような変換を考える。この変換の結合重み $\boldsymbol{a}$ は訓練サンプルの線形結合によって、

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i) \tag{35}$$

のように書ける。これを上式に代入すると

$$\begin{aligned} y &= \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) \\ &= \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) \\ &= \boldsymbol{\alpha}^T \boldsymbol{k}_i(\boldsymbol{x}) \end{aligned} \tag{36}$$

となる。ただし、$\boldsymbol{k}_i(\boldsymbol{x}) = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_N, \boldsymbol{x}))^T$ は、カーネル特徴を並べたベクトル（カーネル特徴ベクトル）である。

これらの関係からわかるように、判別基準

$$J = \frac{\boldsymbol{\alpha}^T \Sigma_B^{(K)} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \Sigma_W^{(K)} \boldsymbol{\alpha}} \tag{37}$$

を最大とするパラメータ $\boldsymbol{\alpha}$ を求める問題は、カーネル特徴ベクトルに基づいて線形判別分析を行うことと等価となり、固有値問題

$$\Sigma_B^{(K)} \boldsymbol{\alpha} = \Sigma_W^{(K)} \boldsymbol{\alpha} \lambda \tag{38}$$

の解として求まる。ここで、$\Sigma_B^{(K)}$ および $\Sigma_W^{(K)}$ は、それぞれ、カーネル特徴ベクトルに関する平均クラス間分散共分散行列および平均クラス内分散共分散行列である。

判別分析は、識別に有効な低次元の特徴を抽出する手法であり、汎化性能は比較的良いが、カーネル判別分析の場合には、汎化性能を向上させる工夫が必要となることもある。最も簡単で良く知られている方法は、平均クラス間分散共分散行列の対角要素に適当な定数を加えて、

$$\tilde{\Sigma}_W^{(K)} = \Sigma_W^{(K)} + \alpha I \tag{39}$$

のようにする手法である。これは、各特徴に平均 0 の正規ノイズを加えるのと同様の効果があり、数値計算を安定化させる。

3 非線形識別器の画像認識への応用

3.1 Chamfer Distance に基づくカーネルを用いた歩行者検出

Gavrila は、あらかじめ複数枚のテンプレートを用意し、Chamfer Distance を用いてそれらのテンプレートとのマッチングにより歩行者を検出する手法 [11] を提案している。

ここでは、あらかじめテンプレートを用意せずにサポートベクターマシンによって、認識に有効なテンプレートを自動的に決定し、歩行者を検出する識別器を構成することを考えてみよう。

Gavrila は、テンプレートと入力画像のマッチングの判定に Chamfer Distance を用いているが、それと同様なマッチングを実現するために Chamfer Distance に Gauss 関数を適用したものをカーネル関数として使用し、カーネルサポートベクターマシンにより識別器を実現することを考えた。カーネル関数がサポートベクターマシンで通常用いられるものと異なるため、既存のサポートベクターマシンライブラリをそのまま用いることが出来ないが、Chamfer distance によるカーネルグラムマトリクスを経験的カーネルマップ (empirical kernel map)[9, 12] に変換すると、線形のサポートベクターマシン用のプログラムをそのまま利用して、カーネルサポートベクターマシンを実現することが可能になる。

Chamfer distance は、画像の類似性を「距離」として測るもので、以下の手順で求められる。まず、二枚の原画像 I, J(図 5(A)) からエッジ特徴 I_e, J_e を抽出する (図 5(B)。次に、エッジ画像を DT(distance transformation) 画像 I_d, J_d(図 5(C)) に変換する。DT 画像の各ピクセルは、最も近いエッジ特徴までの距離を表している。

画像 I から J への chamfer distance は、

$$D_{chamfer}(I, J) = \frac{1}{|I_e|} \sum_{i \in I_e} J_d(i) \tag{40}$$

によって計算される。ここで、$|I_e|$ は I のエッジ画像 I_e のピクセル数、$J_d(i)$ は画像 I のエッジ特徴から画像 J の最も近いエッジ特徴までの chamfer distance を示す。実際には、この値は、エッジ画像 I_e と DT 画像 J_d をピクセルごとに掛け合わせることによって求められる。ここでは、$D_{chamfer}(I, J)$ と $D_{chamfer}(J, I)$ は必ずしも等しい値とならないため、この値の小さい方を I と J の距離として用いることとした。

訓練サンプルに対するカーネル特徴をまとめた行列は、グラムマトリクスと呼ばれており、

$$K = [K_{ij}] = \left[\exp(-\frac{D_{chamfer}(i, j)}{2\sigma^2}) \right] \tag{41}$$

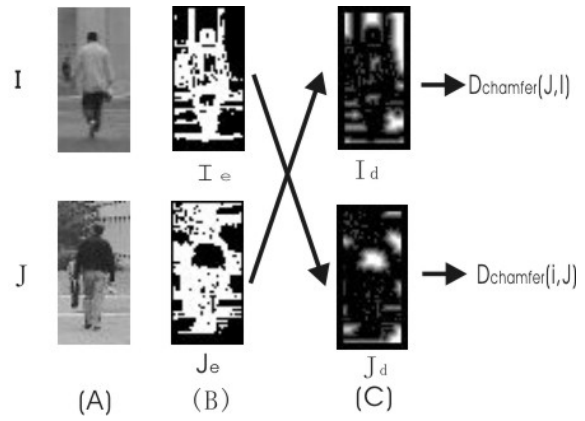

図 5: (A) 原画像,(B) エッジ画像,(C)DT 画像

で定義される。

任意のカーネル関数によるグラムマトリクスが与えられた場合、グラムマトリクスの主成分分析（カーネル主成分分析）を用いて、経験的カーネル特徴ベクトルを

$$\boldsymbol{k}_{emp}(\boldsymbol{x}) = K^{-\frac{1}{2}}\boldsymbol{k}(\boldsymbol{x}) \tag{42}$$

のように変換すると、この経験的カーネル特徴を新たな特徴とみなして線形のサポートベクターマシンを学習させれば、このカーネル関数を用いたカーネルサポートベクターマシンが構成できる。

実験では、歩行者の画像は MIT CBCL 画像データベースの 924 枚の画像、非歩行者の画像はランダムに選択した 2700 枚の画像を使用した。訓練データとして、歩行者画像 100 枚、非歩行者画像 300 枚を選び、Chamfer Distance に基づくカーネルとしてサポートベクターマシンを訓練した。続いて、テストデータとして、歩行者画像 300 枚、非歩行者画像 900 枚を与えたところ、総認識率 90.5 %、False Positive 率 10.8 %という結果を得た。False Positive 率の高さが総認識率をも引き下げているが、この原因は歩行者画像に含まれる背景が影響していると思われる。特徴点の選択 [13] により画像中の有効な領域を選択してカーネル特徴を計算することにより、認識率 (False Positive 率) の向上が期待できる。

3.2 カーネル判別分析による顔検出

サポートベクターマシンでは、「マージン最大化」による汎化性能を向上させたパーセプトロンとカーネル学習法を組み合わせることで汎化性能の高い非線形識別器を構成した。パーセプトロンを他のモデルに変更したり、他の非線形化法を利用すると、サポートベクターマシンと同等の性能を持つ非線形の識別器が構成できる。そこで重要となるのは、非線形化しても汎化性能が低下しないような工夫を識別器に組み込むことである。以下、著者等が行った非線形識別器の画像認識への応用例について簡単に紹介する。

線形判別分析をカーネルトリックを用いて、非線形に拡張するとカーネル判別分析が得られる。それを顔検出に適用する場合には、顔と顔以外のすべての対象とを識別する識別器を構成する必要があるが、顔以外のすべての対象をひとつのまとまったクラスと考えるのは難しく、2 クラスの判別分析をそのまま適用するのは無理がある。栗田等 [14] は、判別分析のための判別基準を検出課題用に変形したカーネル判別分析を顔検出に適用した。

具体的には、顔のクラスはなるべく集まり、顔以外のサンプルはクラスの重心からなるべく離れるような判別特徴を構成した。

今、顔クラスのサンプル集合と顔以外のサンプル集合をカーネル特徴ベクトルを用いて、

$$\begin{aligned} C &= \{\boldsymbol{k}(\boldsymbol{x}_i) \mid i = 1, \cdots, n_f\} \\ \bar{C} &= \{\boldsymbol{k}(\boldsymbol{x}^k) \mid k = 1, \cdots, n_{\bar{f}}\}, \end{aligned} \tag{43}$$

のように表すとする。ここで、n_f は訓練サンプル中の顔画像の枚数である、$n_{\bar{f}}$ は、顔以外の画像の枚数である。この時、顔クラスの分散共分散行列は、

$$\begin{aligned} \bar{\boldsymbol{k}}_f &= \frac{1}{n_f}\sum_{i=1}^{n_f}\boldsymbol{k}(\boldsymbol{x}_i), \\ \Sigma_f &= \frac{1}{n_f}\sum_{i=1}^{n_f}(\boldsymbol{k}(\boldsymbol{x}_i) - \bar{\boldsymbol{k}}_f)(\boldsymbol{k}(\boldsymbol{x_i}) - \bar{\boldsymbol{k}}_f)^T \end{aligned} \tag{44}$$

となる。また、全平均ベクトルは、

$$\bar{\boldsymbol{k}}_T = \omega_f\bar{\boldsymbol{k}} + \frac{n_{\bar{f}}}{N}\sum_{k=1}^{n_{\bar{f}}}\boldsymbol{k}(\boldsymbol{x}^k) \tag{45}$$

となる。ここで、$\omega_f = \frac{n_f}{N}$ であり、$N = n_f + n_{\bar{f}}$ は、全サンプル数である。これらから、平均クラス内分散共分散行列および平均クラス間共分散行列は、

$$\begin{aligned} \Sigma_W^{(f)} &= \omega_f\Sigma_f \\ \Sigma_B^{(f)} &= \omega_f(\bar{\boldsymbol{k}}_f - \bar{\boldsymbol{k}}_T)(\bar{\boldsymbol{k}}_f - \bar{\boldsymbol{k}}_T)^T \\ &+ \frac{1}{N}\Sigma_{k=1}^{n_{\bar{f}}}(\boldsymbol{k}(\boldsymbol{x}^k) - \bar{\boldsymbol{k}}_T)(\boldsymbol{k}(\boldsymbol{x}^k) - \bar{\boldsymbol{k}}_T)^T \end{aligned} \tag{46}$$

となる。

従って、判別基準

$$J = \boldsymbol{tr}(\hat{\Sigma}_W^{(f)-1}\hat{\Sigma}_B^{(f)}) \tag{47}$$

を最大とする新特徴ベクトル $\boldsymbol{y} = A^T\boldsymbol{k}(\boldsymbol{x})$ を求めるための最適な係数行列 A は、固有値問題

$$\Sigma_B^{(f)}A = \Sigma_W^{(f)}A\Lambda \quad (A^T\Sigma_W^{(f)}A = I). \tag{48}$$

を解くことで求まる。構成される、新特徴 $\boldsymbol{y}$ の次元は、$\min(n_{\bar{f}}, N) = n_{\bar{f}}$ となるが、2 クラスのカーネル判別分析のように新特徴の次元が 1 次元となることはなく、高次元の特徴が得られる。

汎化性能を向上させるためには、前述のように平均クラス内分散共分散行列を

$$\tilde{\Sigma}_W^{(f)} = \Sigma_W^{(f)} + \alpha I \tag{49}$$

のように修正して計算すればよい。

MIT 顔画像データベース、CMU テスト画像データベース、Web 上で収集した顔データ等を含む画像データベースに対して上述の手法とサポートベクターマシンの識別性能を評価した。100 枚の顔画像と 200 枚の顔以外の画像を用いて学習し、325 枚の顔画像と 1000 枚の顔以外の画像で認識率を計算した。表 1 にそれぞれの手法での認識率を示す。カーネル判別分析の判別基準を工夫することで、サポートベクターマシンよりも若干良い結果が得られている。

表 1: 認識性能の比較

	全サンプル	顔画像	顔以外の画像
カーネル判別	99.17 %	98.15 %	99.50 %
SVM	98.34 %	99.69 %	97.90 %

図 6 に、この方法で構成した識別器を用いて顔検出を行った結果を示す。

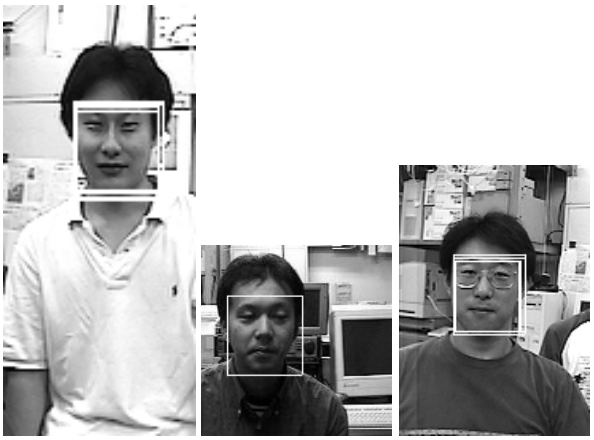

図 6: 顔検出例

4 おわりに

本稿では、サポートベクターマシンを中心に汎化性能の高い非線形の識別器を構成するための手法としてのカーネル学習の話題について紹介した。多様化するパターン認識技術への高い要求に答えるためには、問題に応じて、非線形性と高い汎化性能を両立させた識別器を構成することが重要であると考えらるれが、カーネル学習法はそれを実現するための選択肢のひとつとして重要な技術である。

参考文献

[1] V.N.Vapnik, *Statistical Learning Theory*, John Wiley & Sons (1998).

[2] 赤穂, 津田, “サポートベクターマシン—基本的仕組みと最近の発展—,” 数理科学, No.444, pp.52-58 (2000).

[3] 前田, “痛快！サポートベクトルマシン-古くて新しいパターン認識手法–,” 情報処理, Vol.42, No.7, pp.676-683 (2001).

[4] E.Osuna and R.Freund and F.Girosi, “Training Support Vector Machines: an Application to Face Detection,” Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp.130-136, 1997.

[5] C.P.Papageorgiou and M.Oren and T.Poggio, “A General Framework for Object Detection,” Proc. Fifth Int’l Conf. on Computer Vision, 1998.

[6] A.Mohan and C.P.Papageorgiou and T.Poggio, “Example-Based Object Detection in Images by Components,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vo.23, No.4, pp.349-361, 2001.

[7] T.Hastie, R.Tibshirani, J.Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer-Verlag, 2001.

[8] K.R.Muller, S.Mika, G.Ratsch, K.Tsuda, B.Scholkopf, “An introduction to kernel-based learning algorithms,” IEEE Trans. On Neural Networks, Vol.12, No.2, pp.181-201, 2001.

[9] B.Scholkpf and A.J.Smola, “Learning with Kernels,” The MIT Press, 2002.

[10] 麻生、津田、村田、” パターン認識と学習の統計学,” 岩波書店, 2003.

[11] D.M.Gavlira, “Pedestrian Detection from a Moving Vehicle”, *Proc. of Europian Conference on Computer Vision*, pp.37-49, 2000.

[12] B.Schölkopf, S.Mika, C.J.C,Burges, P.Knirsch, KR.Müller, G.Rätsch, and A.J.Smola, “Input Space Versus Fueature Space in Kernel-Based Methods”, *Trans. on Neural Networks*, val.10, No.5, pp.1000-1017, 1999.

[13] 堀田一弘、三島健稔、栗田多喜夫, “未知の画像に対する識別率を用いた顔検出のための特徴点の順序付け,” 電子情報通信学会論文誌, Vol.J84-D-II, No.8, pp.1781-1789, 2001.

[14] T.Kurita, T.Taguchi, “A modification of kernel-based Fisher discriminant analysis for face detection,” Proc. of the Fifth Inter. Conf. on Automatic face and Gesture Recognition, 20-21 May 2002, Washington, D.C., pp.300-305, 2002.