前田 英作 maeda@cslab.kecl.ntt.co.jp
NTT コミュニケーション科学基礎研究所

解説

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−

いま巷ではやりの技術, サポートベクトルマシン (SVM)

最近，サポートベクトルマシン（**Support vector machine, SVM**）という新しいパターン認識手法に関する論文や紹介記事をいろいろなところで見かけるようになった．かつてのニューラルネットワークほどではないにしろ，ちょっとしたブームである．では，SVMという手法，どこが新しいのか，そして本当に良い方法なのか？ あちらこちらで手に入るパッケージを利用してちょっとした実験をしてみるのは比較的簡単である．しかしながら，実用システムの中でSVMを使う場合には，SVMという手法の本質を正しく理解し，SVMを使うことの価値をあらかじめ見極めておく必要があろう．SVMの属する**パターン認識**という研究分野には古い歴史がある．ここでは，SVMの技術ポイントを解説するとともに，「古きを訪ねて新しきを知る」という視点からみたSVMについても触れてみたい．このSVM，ちょっと使ってみる分にはなかなか面白いのである．できればその面白さも伝えたいと思う．

パターン認識の基礎知識が生死を分かつ

まず初めに，パターン認識になじみのない読者のために，パターン認識とはどんな技術なのか，何の役に立つのかについて簡単に述べておこう．パターン認識というとその主要な応用例である文字認識，音声認識がまず思い浮かぶ．このとき注意すべきことは，文字や音声の認識を実現しているこれらの技術の中には，**特徴抽出**☆1など対象に強く依存した部分と，識別・学習など対象依存性が比較的少ない部分とがあるという点である．特定の応用を念頭におくと前者の重要性がより強調される傾向がある．認識性能を高めるためには両者とも重要な技術であるが，「パターン認識」の本質的な部分はむしろ後者にある．この識別・学習理論は古くから重要な研究領域の1つであり，「文字認識」の本質が文字特徴抽出技術にあるとすれば，「パターン認識」の本質は識別・学習技術にあるといってよい．実際，パターン認識技術は，ベイズ理論，パラメータ推定，多変量解析，ノンパラメトリック手法，機械学習など多種多様な技術と密接に関連しており[1]，その適用領域は認識という言葉の意味を超えて広範である．パターン認識技術とは学習あるいは最適化の手法であるともいえ，システムに学習機能を組み込んだり，システムの最適パラメータを求めたりする際に必要となる技術なのである．

ここでは，パターン認識の具体例として毒キノコを見分ける方法を取り上げよう．今ここに老齢のきのこ名人がいて今年収穫されたキノコのうち，100本の毒キノコと100本の毒のないキノコを選び出したまま旅に出てしまった．さて村人たちは毒の有無が不明なまま残された大量のキノコをかかえてこの冬の飢えをどのようにしてしのぐのか？ 毒の有無が分かっている200本のキノコ（学習用キノコ）を利用して，毒キノコを見分ける最良の方法を導き出すこと，これがパターン認識の扱う課題である．

各キノコの特徴のうち，柄の長さ，柄の太さ，笠の面積，重さ，色スペクトル，アミノ酸含有率など定量的に扱えるものをキノコの**特徴量**と呼ぶ．今，d個のこうした特徴量に着目したとすると，あるキノコの特徴は**特徴空間**と呼ばれるd次元空間の1点$\mathbf{x}$として表現することができる．このd次元ベクトル$\mathbf{x}$を**パターン**あるいは**特徴ベ**

☆1 識別に有効な特徴量を認識対象から算出すること．

クトルと呼び，特に学習用キノコから得られるパターンを**学習パターン**と呼ぶ．毒のないキノコの集合をクラス1（χ_1），毒キノコの集合をクラス2（χ_2）とする．解決すべき問題は，属するクラスが不明なパターン$\mathbf{x}$（これを**未知パターン**と呼ぶ）のクラスを判定する**識別規則**を学習パターンを利用して求めることであり，これは特徴空間上での2クラスの境界を決定することに相当する（図-1）．この操作を**学習**と呼ぶ．

識別規則は，たとえば，

$$f_{\mathbf{w}}(\mathbf{x}) \equiv \mathrm{sign}\,(g_{\mathbf{w}}(\mathbf{x})) = \begin{cases} 1 & \mathbf{x} \in \chi_1 \\ -1 & \mathbf{x} \in \chi_2 \end{cases} \qquad (1)$$

と記述できる．ここで，$f_{\mathbf{w}}(\mathbf{x})$は$\mathbf{x}$を引数とする関数であり**識別関数**と呼ばれる．$\mathbf{w}$は関数fのパラメータを表すベクトルである．特徴空間上の識別境界は$g_{\mathbf{w}}(\mathbf{x})=0$で与えられる．もし，各クラスの出現確率$P_i$と特徴空間上における各クラスのパターンの確率密度$p(\mathbf{x}|\mathbf{x} \in \chi_i)$とが明らかであれば，パターン$\mathbf{x}$がクラス$i$に属する確率$P(\mathbf{x} \in \chi_i|\mathbf{x})$，すなわち**事後確率**を計算することにより最適な識別境界を決定できる．しかし，実際に扱う問題がこのような条件を満たすことはきわめて稀である．また，学習パターンをすべて正しく識別できるような識別規則が未知パターンを正しく識別できるとは限らない．したがって，限られた学習パターンからいかに汎用的な識別規則を見出すかが鍵となる．

パターン認識に関する基礎知識の有無が場合によっては生死を分かつという毒キノコの例は，やや極端かもしれない．だが，パターン認識が，すでにその善し悪しが分かっているデータ（事例）を用いて判断規則の作成やシステムの設計を行い，未知のデータに対する最善の策を講ずるための技術であることはお分かりいただけると思う．そして，人間という学習機械はこの作業を常に行っているのである．

パターン認識手法としてのSVM

前章で述べたように，学習，すなわち，学習パターンを使って最適識別境界を決定することによりパターン認識の問題は解決される．したがって，パターン認識手法は適切な学習アルゴリズムがあって初めて意味を持つことになる．本章では，まず，パターン認識課題がパターンの分布から2つに大別できること，パターン認識手法が3つの観点から分類できることを述べる．その上で，パターン認識手法としてのSVMが持つ特徴をまとめてみたい．

特徴空間上における2クラスのパターン分布には，超平面によって2クラスを分けることができる**線形分離可能**な場合（図-2）と分けることができない**線形分離不可能**な場合（図-1）とが存在する．線形分離可能となる特徴の組を見つけることができれば，識別境界の決定は比較的容易である．しかし，実際の問題は線形分離不可能であることが多く，そのような場合にも適用可能なパターン認識手法が必要となる．たとえば，パターン認識研究が発展する1つの契機となった**パーセプトロン**と呼ばれるパターン認識手法は，線形分離不可能な場合に有効な学習アルゴリズムがない点が欠点とされた．

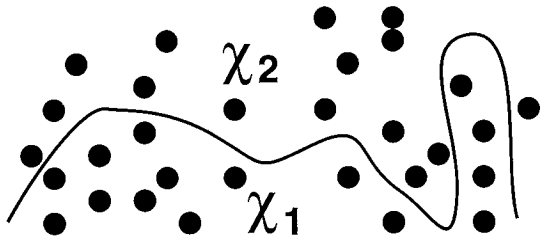

図-1　2種類のキノコの特徴ベクトル（青丸および赤丸）の分布と毒キノコ（赤丸）を見分けるための識別境界（黒曲線）

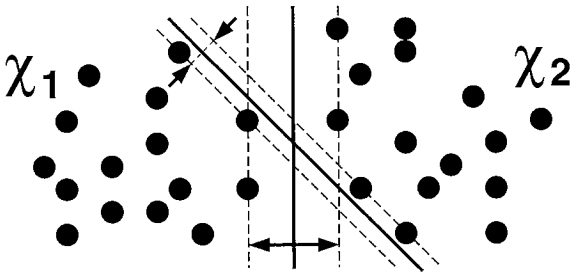

図-2　線形分離可能な2クラスのパターン（青丸および赤丸）とマージンを最大にする線形識別境界（黒実線）

一方，以下のような観点からパターン認識手法を分類することができる．(1) 識別関数としてどのようなクラスの関数を使うのか．線形関数がそのクラスの一例である．識別関数として線形関数を用いた場合には識別境界は必ず超平面になる．このように関数のクラスによって識別境界の記述能力が異なる．(2) 関数のパラメータをどのような基準で最適化するのか．学習パターンに対する誤識別最少が最適化基準の一例である．(3) 最適化のためのアルゴリズムは何を使うのか，あるいは何が使えるのか．たとえば，ニューラルネットワークでは最適化の手段として最急降下法が用いられる．

SVMは，線形分離可能な場合，不可能な場合の両方に適用可能なパターン認識手法である．識別関数が線形関数となる**線形SVM**と，識別関数が非線形関数となる**非線形SVM**とがある．ただし，非線形関数のクラスは後に述べる条件（式（35））によって制約を受ける．さらに，SVMのその他の特徴として，マージン最大化基準によって関数を最適化する点，最適化が凸2次計画法を解くことにより実現できる点が挙げられる．次章以降では，SVMの詳細について数学的な定式化を行いながら説明

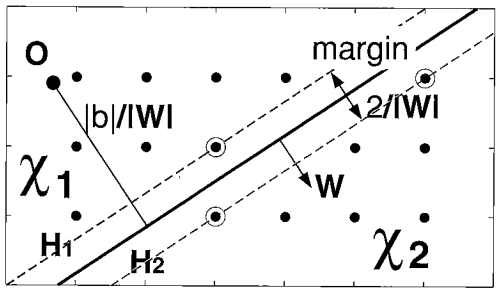

図-3　線形分離可能な場合の線形SVM

する．

線形SVMその1：線形分離可能な場合

ここではまず最も簡単な例として，線形分離可能な場合の線形SVMについて述べよう．

線形分離可能な場合には，一般に学習パターンをすべて正しく識別する超平面が複数存在する（図-2）．識別境界候補となるそうした超平面（黒実線）の周囲には学習パターンが存在しないマージン領域（破線で示した2つの超平面に挟まれた領域）が存在する．SVMでは，複数の識別境界候補の中でマージン（矢印）を最大にする超平面を最良と見なす．このマージン最大化がSVMの大きな特徴である．線形SVMの学習，すなわち学習パターンからマージンを最大にする最適識別関数を求める過程は以下のように数学的に定式化できる．

線形識別関数$f(\mathbf{x})$を，

$$f(\mathbf{x}) \equiv \mathrm{sign}\,(g(\mathbf{x})) = \mathrm{sign}\,(\mathbf{w}^t\mathbf{x} + b) \tag{2}$$

とおく．ここでn個の学習パターン$\mathbf{x}_i(i=1,...,n)$の満たすべき条件を，

$$\forall i,\quad g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}_i + b \begin{cases} \geqq 1 & \mathbf{x}_i \in \chi_1 \\ \leqq -1 & \mathbf{x}_i \in \chi_2 \end{cases} \tag{3}$$

とする．点$\mathbf{x}_i$から平面$g(\mathbf{x})=0$までの距離は$|g(\mathbf{x}_i)|/\|\mathbf{w}\|$であるから，式（3）は，識別境界$(g(\mathbf{x})=0)$から距離$1/\|\mathbf{w}\|$の範囲内，すなわち，平面$g(\mathbf{x})=\pm 1$（図-3の$H_1$，$H_2$）の間に学習パターンが存在しないことを意味する．ここで，$\mathbf{x}_i$の属するクラスを変数y_iで表し，

$$y_i = \begin{cases} 1 & \mathbf{x}_i \in \chi_1 \\ -1 & \mathbf{x}_i \in \chi_2 \end{cases} \tag{4}$$

と定義し，$\mathbf{x}_i$の教師信号と呼ぶ．また，$\mathbf{y}=(y_1,...,y_n)^t$とする．このとき，式（3）は$y_i$を用いて，

$$\forall i,\quad y_i \cdot (\mathbf{w}^t\mathbf{x}_i + b) - 1 \geqq 0 \tag{5}$$

と書ける．平面H_1，H_2間の距離（マージン）は$2/\|\mathbf{w}\|$であり，これを最大にする$f(\mathbf{x})$は，扱いやすくするために$\|\mathbf{w}\|^2$を考えて，

$$\underset{\mathbf{w},b}{\mathrm{Minimize}}\quad G(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{6}$$
$$\text{s.t.}\quad \forall i,\quad y_i \cdot (\mathbf{w}^t\mathbf{x}_i + b) - 1 \geqq 0$$

の解$\mathbf{w}^*$，b^*により決まる．

この最小化問題はラグランジュの未定乗数法によって解くことができる．λ_iを正の乗数，$\boldsymbol{\lambda}$を$\boldsymbol{\lambda}=(\lambda_1,...,\lambda_n)^t$と定義するとラグランジュ関数$L_p$は，

$$L_p(\mathbf{w}, b, \boldsymbol{\lambda}) \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\lambda_i[y_i \cdot (\mathbf{w}^t\mathbf{x}_i + b) - 1] \tag{7}$$

となる．L_pを偏微分して0とおくことにより，

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i = 0 \tag{8}$$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial b} = \sum_{i=1}^{n}\lambda_i y_i = \boldsymbol{\lambda}^t\mathbf{y} = 0 \tag{9}$$

となり，式（8）から，

$$\mathbf{w}^* = \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i \tag{10}$$

となる．$L_p(\mathbf{w}^*, b^*, \boldsymbol{\lambda})$を$F(\boldsymbol{\lambda})$とおき，$\mathbf{D}$をその$(i, j)$成分が$y_i y_j \mathbf{x}_i^t\mathbf{x}_j$である$(n, n)$行列とすると，式（8），式（9），式（10）より，

$$F(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\|\mathbf{w}^*\|^2$$
$$= \boldsymbol{\lambda}^t\mathbf{1} - \frac{1}{2}\boldsymbol{\lambda}^t\mathbf{D}\boldsymbol{\lambda} \tag{11}$$

が得られる．$\mathbf{1}$はそのすべての成分に1を持つ縦ベクトルを表す．

こうして，式（6）の最小化問題は$F(\boldsymbol{\lambda})$の最大化問題，

$$\underset{\boldsymbol{\lambda}}{\mathrm{Maximize}}\quad F(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\mathbf{1} - \frac{1}{2}\boldsymbol{\lambda}^t\boldsymbol{D}\boldsymbol{\lambda} \tag{12}$$
$$\text{s.t.}\quad \boldsymbol{\lambda}^t\boldsymbol{y} = 0$$
$$\boldsymbol{\lambda} \geqq 0$$

に帰着される．これは2次計画法において式（6）の双対問題☆2と呼ばれる．式（12）の解をλ_i^*とする．式（10）より正の（0でない）λ_i^*に対応する$\mathbf{x}_i$のみから$\mathbf{w}^*$が決まる．この$\mathbf{x}_i$はサポートベクトル（以降，SVと記す）と呼ばれる．一方，非線形計画法において相補性条件☆3と呼ばれる，

$$\forall i,\quad \lambda_i^*[y_i \cdot (\mathbf{w}^{*t}\mathbf{x}_i + b^*) - 1] = 0 \tag{13}$$

が成立するので，b^*は任意のSV，$\mathbf{x}_i\,(\lambda_i^* > 0)$を用いて，

$$b^* = y_i - \mathbf{w}^{*t}\mathbf{x}_i \tag{14}$$

と求まる．最終的に，線形SVMによる識別関数$f(\mathbf{x})$は式（2），式（10）より，

☆2 双対問題，相補性条件については（非）線形計画法に関する成書（たとえば文献2））を参考のこと．
☆3 Kuhn-Tucker条件の1つ．☆2参照．

$$f(\mathbf{x}) = \text{sign}\,(\mathbf{w}^{*t}\mathbf{x} + b^*) \tag{15}$$

$$= \text{sign}\left(\sum_{i=1}^{n}\lambda_i^* y_i \cdot (\mathbf{x}_i^t\mathbf{x}) + b^*\right) \tag{16}$$

となる．

ここで，SVの意味についてみておこう．SVの定義と式(16)から確認できるように，SVとなる$\mathbf{x}_i$のみによって識別関数すなわち識別境界が決まり，SVでない$\mathbf{x}_i$はその決定に寄与しない．図-4は，3種類のパターン分布に対して線形SVMを適用し識別境界を求めた例である．二重丸はそのパターンがSVであることを示す．Aでは各クラスはそれぞれ9個のパターンからなり，BではAにおいてSVとなったパターンのうちの3個を取り除いて学習パターンとし，さらにCではBにおける非SVパターン12個を取り除いて学習パターンとした．A, B, Cで得られた識別境界を比較して分かるように，SVを取り除くとSVMによって決まる最適識別境界が変わるが(A, B)，SVでないパターンをいくら除いても識別境界には影響しない(B, C)．これがサポートベクトルと呼ばれるゆえんである．

線形SVMその2：線形分離不可能な場合

線形分離不可能な場合，式(3)を満たす$\mathbf{w}$は存在しない．そこで正変数ξ_i $(i=1,\ldots,n)$を導入して条件を緩め，

$$\forall i,\quad \mathbf{w}^t\mathbf{x}_i + b \begin{cases} \geqq 1-\xi_i & \mathbf{x}_i \in \chi_1 \\ \leqq -1+\xi_i & \mathbf{x}_i \in \chi_2 \end{cases} \tag{17}$$

とする．そして最小化問題を

$$\underset{\mathbf{w},b,\boldsymbol{\xi}}{\text{Minimize}}\quad G(\mathbf{w},\boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + c\sum_{i=1}^{n}\xi_i \tag{18}$$

$$\text{s.t.}\quad \forall i,\quad y_i \cdot (\mathbf{w}^t\mathbf{x}_i + b) - (1-\xi_i) \geqq 0$$

$$\forall i,\quad \xi_i \geqq 0$$

と定める．図-5を用いて式(18)の意味を考えてみよう．右辺第1項は線形分離可能な場合と同様にマージンを大きくとるためのものであり，一方第2項はマージンからはみ出したパターンに対するペナルティ項である．たとえば図のクラス2に属するパターンのうち平面H_2から識別境界側にあるパターンは，式(17)の制約から$\xi_i>0$となる．図中の$\mathbf{x}_i$がその例であり，$\mathbf{x}_i$からH_2までの距離は$\xi_i/\|\mathbf{w}\|$となる．$\mathbf{x}_i$が誤識別されている時には$\xi_i>1$となるので，ξ_iの和は誤識別学習パターン数の上限を与える．係数cは第1項と第2項のバランスを決める定数である．cが小さければ$g(\mathbf{x})=\pm 1$(破線)間の距離は大きくなる．最適なcは通常実験により決めなければならない．

以下，最小化問題(18)の解法は基本的に前章と同様であるので結果のみを記す．λ_i, $\gamma_i$$(i=1,\ldots,n)$をラグランジュ乗数とし，ラグランジュ関数

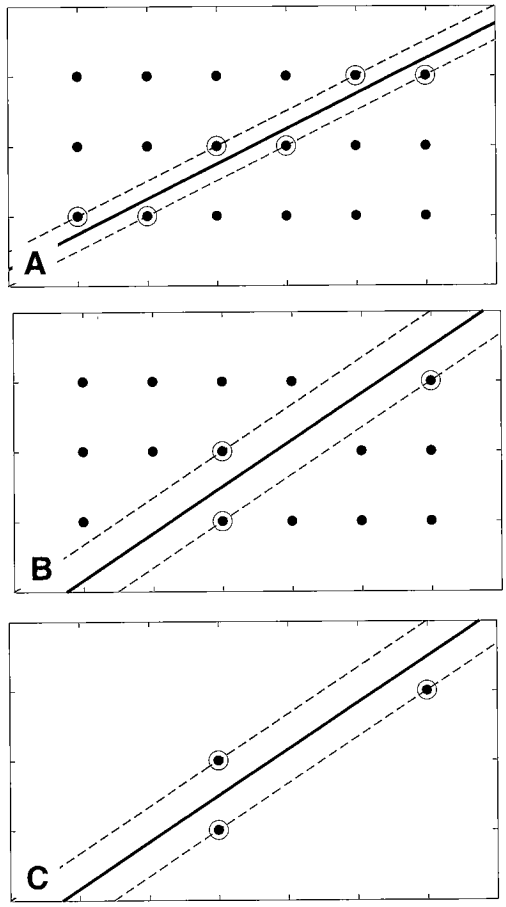

図-4　線形SVMによる識別境界（黒線）とサポートベクトル（二重丸）

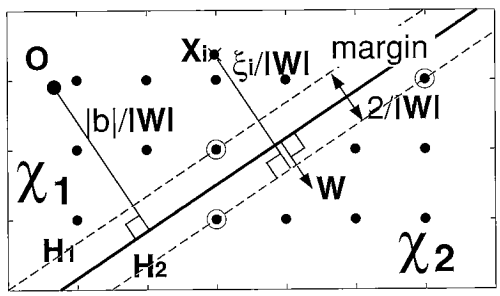

図-5　線形分離不可能な場合の線形SVM

$$L_P(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\gamma})$$

$$\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\lambda_i[y_i\cdot(\mathbf{w}^t\mathbf{x}_i+b) - 1 + \xi_i]$$

$$-\sum_{i=1}^{n}\gamma_i\xi_i + c\sum_{i=1}^{n}\xi_i \tag{19}$$

を$\mathbf{w}, b, \boldsymbol{\xi}$で偏微分すると，

$$\mathbf{w} - \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i = 0 \tag{20}$$

$$\boldsymbol{\lambda}^t\mathbf{y} = 0 \tag{21}$$

$$c - \lambda_i - \gamma_i = 0 \tag{22}$$

となり，最小化問題(18)は，

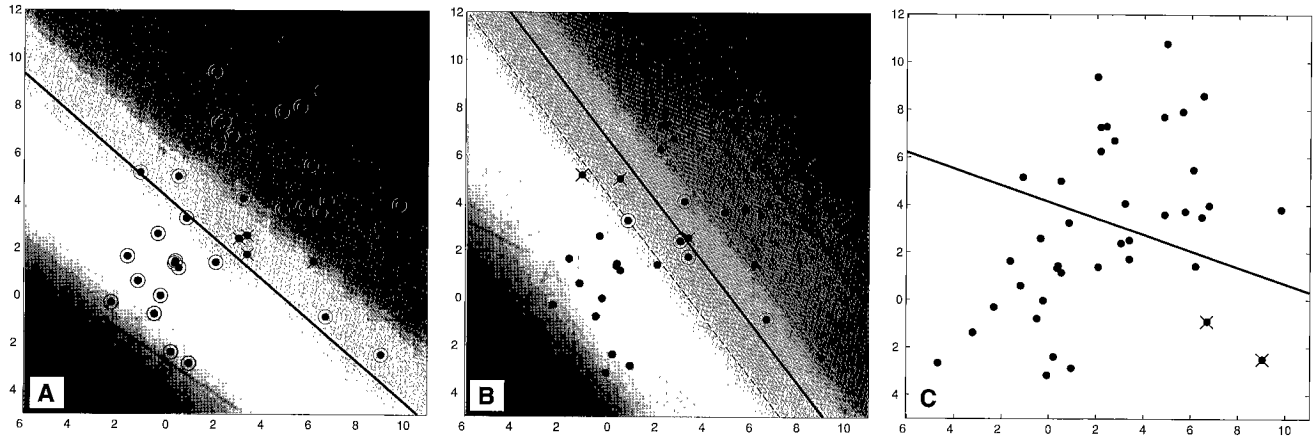

図-6　線形SVMによる識別境界（c=0.001（A），1（B））とFisherの線形判別法による識別境界（C）

$$\underset{\boldsymbol{\lambda}}{\text{Minimize}} \quad F(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\mathbf{1} - \frac{1}{2}\boldsymbol{\lambda}^t\mathbf{D}\boldsymbol{\lambda} \tag{23}$$

$$\text{s.t.} \quad \boldsymbol{\lambda}^t\mathbf{y} = 0$$

$$0 \leqq \boldsymbol{\lambda} \leqq c\mathbf{1}$$

という2次計画法の問題に帰着する．一方，相補性条件より，

$$\forall i, \quad \lambda_i^*[y_i \cdot (\mathbf{w}^{*t}\mathbf{x}_i + b^*) - 1 + \xi_i] = 0 \tag{24}$$

$$\forall i, \quad \gamma_i\xi_i = 0 \tag{25}$$

が成立する．以上の結果を用いると，式（18）を満たすb^*は$0<\lambda_i^*<c$となる任意のλ_i^*に対応する$\mathbf{x}_i$を用いて式（24）より求まり，識別関数$f(\mathbf{x})$は前章と同じ形となり式（15），式（16）で表すことができる．

これまでの結果から式（23）で求まるλ_i^*とλ_i^*に対応するパターン$\mathbf{x}_i$との関係をまとめると次のようになる．以下，$g(\mathbf{x}_i)$とy_iは式（1），式（4）の定義による．まず，$\lambda_i^* = 0$の時，$\mathbf{x}_i$は非SVであり，式（26）が成り立つ．$\mathbf{x}_i$は，図-5においてマージン領域の外側，すなわちH_1，H_2の外側に存在し，正しく識別されるパターンである．図において右下の5個の赤丸と左上の7個の青丸がこれに該当する．次に，$0<\lambda_i^*<c$の時，$\mathbf{x}_i$はSVであり，式（27）が成り立つ．$\mathbf{x}_i$は，ちょうど平面H_1，H_2上に存在し，正しく識別されるパターンである．図において3個の二重丸のパターンがこれに該当する．最後に，$\lambda_i^* = c$の時，$\mathbf{x}_i$はSVであり，式（28）が成り立つ．図のクラス2の場合，H_2より識別境界側にあるパターンに相当する．H_2と識別境界の間にあれば正しく識別され，識別境界よりH_1側にあれば誤識別される（×印のついた赤丸がこの例）．

$$g(\mathbf{x}_i)/y_i > 1, \quad \gamma_i = c, \quad \xi_i = 0 \tag{26}$$

$$g(\mathbf{x}_i) = y_i, \quad 0 < \gamma_i < c, \quad \xi_i = 0 \tag{27}$$

$$g(\mathbf{x}_i)/y_i < 1, \quad \gamma_i = 0, \quad \xi_i \neq 0 \tag{28}$$

図-6は線形分離不可能な分布に対して，線形SVMを適用した結果（A，B）とよく知られた線形識別法である**Fisherの線形判別法**☆4を適用した結果（C）である．パターンは，以下の平均と共分散行列で定義される2次元正規分布$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$に従う乱数から各クラス20個を用いた．

$$\chi_1 : \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix} \tag{29}$$

$$\chi_2 : \boldsymbol{\mu}_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 8 & 0 \\ 0 & 16 \end{pmatrix} \tag{30}$$

式（18）におけるcの値は，それぞれc=0.001（A），c=1（B）とした．破線は$g(\mathbf{x})=\pm 1$となるところを示し，背景色は$g(\mathbf{x})$の大小を示す．ただし色のスケールは図により異なる．AとBを比較して分かるように，cの値がより小さいと平面$g(\mathbf{x})=\pm 1$間の距離が大きくなり，マージン領域内に存在するパターンの数が増える．したがって，SV，すなわち識別境界の決定に関与するパターンが増える．線形SVMとFisherの方法との性能比較はこの結果からだけではできないが，得られる最適識別平面が大きく異なる場合があるということに注意したい．

非線形SVM：カーネルの魔術

2クラスの識別境界が超平面で近似できる場合は，線形識別関数によって実用上十分な性能を実現することができる．ところが，図-1のように複雑な識別境界を持つ場合には，非線形関数によって識別関数を記述しないかぎり，良い識別性能を得ることができない．スカラーを出力する任意のd'個の非線形関数$\phi_i(\mathbf{x})(i=1,\ldots,d')$を用いて関数$\boldsymbol{\phi}$を

$$\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_{d'}(\mathbf{x}))^t \tag{31}$$

と定義する．ここで，$\boldsymbol{\phi}(\mathbf{x})$を新たなパターンと見なし，前章までの$\mathbf{x}$を$\boldsymbol{\phi}(\mathbf{x})$で置き換えることが可能である．これは，パターン$\mathbf{x}$を非線形変換$\boldsymbol{\phi}(\mathbf{x})$によって変換し，変換後の空間において線形SVMを適用したことに相当する．変換後の$\boldsymbol{\phi}(\mathbf{x})$空間における線形識別境界は，$\mathbf{x}$の原

☆4 厳密には，Fisherの方法では識別平面の向きだけが決まる．

特徴空間では非線形な識別境界をなす．

式(12)，式(23)における$\mathbf{D}$の代わりに，

$$\mathbf{D}_{ij} = y_i y_j \boldsymbol{\phi}(\mathbf{x}_i)^t \boldsymbol{\phi}(\mathbf{x}_j) \tag{32}$$

とおくと，解くべき最大化問題は，式(23)とまったく同じ形に書け，識別関数は，

$$f(\mathbf{x}) = \mathrm{sign}\,(\mathbf{w}^{*t}\boldsymbol{\phi}(\mathbf{x}) + b^*) \tag{33}$$

$$= \mathrm{sign}\left(\sum_{i=1}^{n} y_i \lambda_i^* \boldsymbol{\phi}(\mathbf{x}_i)^t \boldsymbol{\phi}(\mathbf{x}) + b^*\right) \tag{34}$$

と求まる．

以上のような方法で非線形SVMが形式的に求まるが，非線形SVMの大きな特徴は実はここから先にある．今，2つの特徴ベクトルを引数とするある関数$K(\mathbf{x}, \mathbf{y})$があって，

$$K(\mathbf{x}, \mathbf{y}) \equiv \boldsymbol{\phi}(\mathbf{x})^t \boldsymbol{\phi}(\mathbf{y}) = \sum_{i=1}^{d'} \phi_i(\mathbf{x})\,\phi_i(\mathbf{y}) \tag{35}$$

が成立するものとする．このKを**カーネル関数**と呼ぶ．この時，式(32)，式(34)は，

$$\mathbf{D}_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{36}$$

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i \lambda_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*\right) \tag{37}$$

となる．式(36)，式(37)は式(35)を満たすKの関数として書くことができ，$\boldsymbol{\phi}$を陽には含まない．したがって，$\boldsymbol{\phi}(\mathbf{x})$空間上での線形SVM，すなわち$\boldsymbol{\phi}(\mathbf{x})$によって定義される非線形SVM(式(37))を求める時，$\boldsymbol{\phi}$の内積形(式(35))さえ定義されていれば$\boldsymbol{\phi}(\mathbf{x})$を計算する必要もなければ$\boldsymbol{\phi}(\mathbf{x})$の具体的な形も知る必要がない．式(35)を満たす$\boldsymbol{\phi}$が存在するためのカーネル関数$K$の条件はすでに知られており[3]，そのような$K$の例として，

$$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^t\mathbf{y})^p \tag{38}$$

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\delta^2}\right) \tag{39}$$

で定義される，多項式型カーネルとガウシアン型カーネルがある．

ここで，Kと$\boldsymbol{\phi}$の具体例をみてみよう．特徴空間の次元を2，パターンを$\mathbf{x}=(x_1, x_2)^t$とする．今，3次の多項式型カーネルを採用すると

$$K(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x})^t \boldsymbol{\phi}(\mathbf{y}) = (1 + \mathbf{x}^t\mathbf{y})^3 \tag{40}$$

となる．この時，式(35)を満たす$\boldsymbol{\phi}(\mathbf{x})$が，

$$\boldsymbol{\phi}(\mathbf{x}) = (1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1x_2, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, x_1^3, x_2^3)^t$$

となることが確かめられる．容易に推測がつくように，特徴空間の次元dと多項式の次数pに依存して$\boldsymbol{\phi}$の次元d'は非常に大きくなり，多項式型カーネルの場合$d' = {}_{d+p}C_p - 1$となる．ところが，最適な非線形SVMを求めたり，得られた非線形識別関数を使って未知パターンのクラスを判定したりする際に，高次元のベクトル演算は必要なく式(38)，式(39)のような低次元演算で足りるのである．これは，**カーネルトリック**と呼ばれており(非)線形SVMの大きな特徴である．次章で触れるように，カーネルトリックにはいろいろな応用が考えられ，これを使って非線形超高次元の旅を手軽に楽しむことが可能になる．

非線形SVMによって得られる識別境界の例を図-7，図-8に示す．パターンは式(29)，式(30)で定義される2次元正規分布に従う乱数から各クラス50パターンを用いた．図-7は多項式型カーネルのSVMを適用した例である．多項式の次元は2(A)，4(B)，6(C)である．パターンの母集合は2次元正規分布に従うから，母集合を最もよく識別する識別境界は本来2次曲面で記述できる．p=6の時(C)，すべての学習パターンに対して誤りなく識別できる識別境界が生成されている．しかし，同じ分布から得られる未知パターンに対して，CがAやBに比べて良いとはいえないことは図からも自明である．図-8は図-7と同じパターンに対しガウシアン型カーネルのSVMを適用した例である．ガウシアンの広がりδによりSVおよび識別境界が変化する様子が分かる．

古くて新しいサポートベクトルマシン

SVMの根幹を成す技術的なアイディアは，実は古くから知られていた．複数のそうしたアイディアが結び付いてSVMという1つの手法が確立し，多くの研究者の注目を集めることとなった．その意味でSVMは古くて新しい手法だといえる．本章では，古くから知られているいくつかのパターン認識手法と比較しながら，SVMの特徴について述べてみたい☆5．

まず初めに，従来の線形識別関数の学習との比較をしよう．線形識別関数の学習法には，(古典的)**パーセプトロン**に代表されるような誤識別パターン数を最少にする方法と，図-6で用いたFisherの線形判別法に代表的される$g(\mathbf{x}_i)$と教師信号y_iの自乗誤差を最小化する方法とに大別される．前者はいわば識別境界付近に着目して誤りをなるべく減らそうという発想なのに対し，後者は分布全体を考慮に入れて2つの分布をなるべくよく分離する識別境界を見つけようという方法である．SVMは基本的に前者のタイプに属するが，従来技術にない特色はマージン最大化という基準を導入した点にある．線形分離可能な場合に，識別境界候補の中からマージンがより大きなものを選ぶというのは自然な考え方である．パターン認識が扱う課題の多くにおいて線形分離不可能であるのが普通であり，通常上記のような状況は発生しなかった．ところが，$\boldsymbol{\phi}(\mathbf{x})$の非常な高次元で線形識別を行う非線形

☆5 本章で言及する従来技術の詳細については文献1)，4)などを参照．

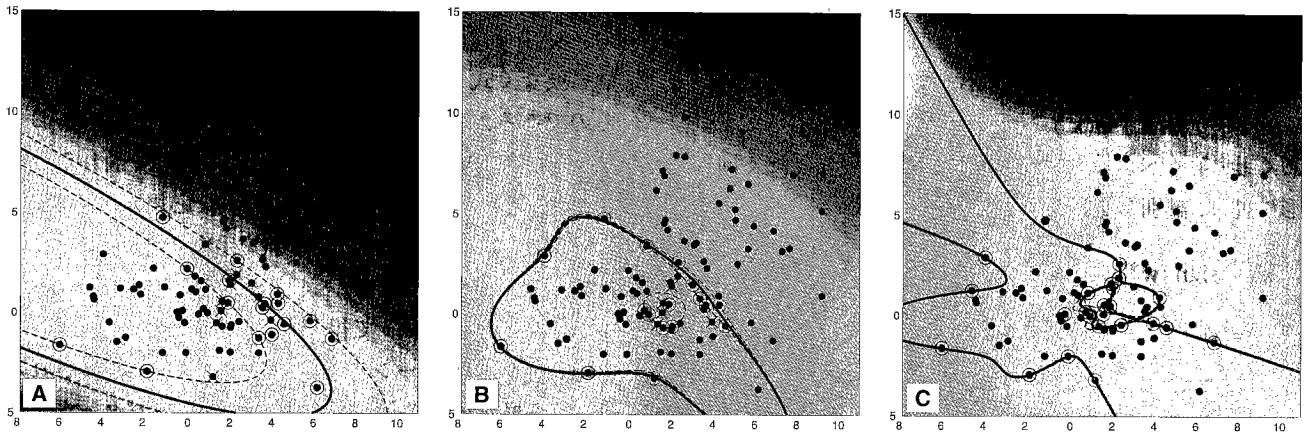

図-7　多項式型カーネルを利用した非線形**SVM**による識別境界の例（多項式の次元p=2 (**A**)，4 (**B**)，6 (**C**)，c=10）

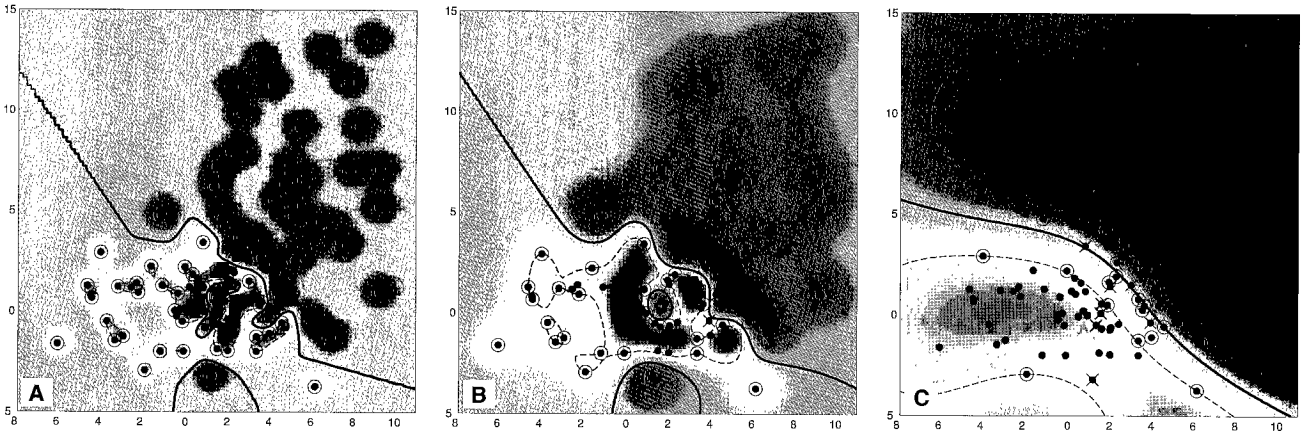

図-8　ガウシアン型カーネルを利用した非線形**SVM**による識別境界の例（ガウシアンの広がりδ=0.5 (**A**)，1 (**B**)，5 (**C**)，c=10）

SVMや，数万次元の学習パターンが各クラス数千しかないような条件で識別関数の最適化を行うテキスト分類のような課題に対しては，マージン最大化基準が有効に働くと期待できる☆6.

次に，**一般識別関数**と非線形SVMとの比較をしよう．前章で，式 (31) の$\boldsymbol{\phi}(\mathbf{x})$を用いて非線形識別関数

$$f(\mathbf{x}) \equiv \mathrm{sign}\,(g(\mathbf{x})) = \mathbf{w}^t\,\boldsymbol{\phi}(\mathbf{x}) + b \qquad (41)$$

を定義できることを述べた．この識別関数は，一般識別関数または**Φ関数**と呼ばれ，線形識別関数の学習を利用して最適化が可能である．1960年代から知られているこの方法により任意の非線形識別関数の最適化が可能であり，非線形SVMはいわばその応用例の1つに過ぎない．ところが，課題に応じた一般識別関数を実現するための$\phi_i(\mathbf{x})$ の条件は一般には分からない．使えるカーネルの選択肢が限られるという点を除くと，SVMも同じ状況にある．ただし，SVMの場合には$\boldsymbol{\phi}(\mathbf{x})$の計算が不要であるという大きな長所がある．

さらに，カーネルトリックについて補足をする．カーネル関数というアイディアは，関数の線形和によって確率密度関数を記述するノンパラメトリック確率密度推定の一手法である**Parzen window法**，これを発展させたパターン認識手法である**ポテンシャル関数法**まで遡ることができる．ポテンシャル関数法では，識別関数$f(\mathbf{x})$ をカーネル関数Kと学習パターン$\mathbf{x}_i$を用いて

$$f(\mathbf{x}) = \sum_{i=1}^{n} q_i K(\mathbf{x}, \mathbf{x}_i) \qquad (42)$$

と定義し，学習によってパラメータq_iの最適値を決定する．1960年代から70年代にかけて，特に，式 (35) の形に展開可能なカーネル関数のクラスについて，学習アルゴリズムの収束性能のなどさまざまな研究がなされた．ただし，ポテンシャル関数法はq_iの推定精度，識別性能を高めるために非常に多くの学習パターンを必要とするという欠点があり，特に特徴空間の次元が高い時には深刻な問題であった．SVMはこのカーネル関数の特性を利用した，非線形識別関数の新しい構成法である．

カーネルトリックは一種の計算技術であるから，これ

☆6 n個のd次元パターンに対してクラス1，2を割り付けるとするとその割り付け方は2^n通りある．今その中の1通りを任意に選んだ時，クラス1とクラス2が線形分離となる確率 $p(n, d)$ はn=2(d+1)の時に1/2となる．したがって本来次元数dの数倍のパターンを使って学習を行うのが望ましいとされる[1), 4)].

を利用してさまざまな手法を容易に非線形へ拡張することが可能である．ただし，その演算が内積計算で求まるものでなければならない．その例として，**非線形PCA**（主成分分析）や，従来から文字認識で使われているパターン認識手法である部分空間法を拡張した**カーネル非線形部分空間法（KNS法）**などがあり，KNS法は学習が高速でかつ高い識別性能を持つことが実証されている[5),6)]．

最後に，他の非線形識別手法との比較をしてみよう．非線形識別関数を実現する方法としては，**k最近傍法，ニューラルネットワーク**がよく知られている．k最近傍法は原理が非常に簡単であること，最適化（学習）の操作が必要がないなどの利点を有しており，識別実験の際に参照用手法としてよく利用される．ただし，特徴空間の次元が高くなると識別性能を高めるために多くの学習パターンを必要とするという欠点がある．ニューラルネットワークは，任意の識別関数を実現できる識別性能の高い非線形パターン認識法である．ところが，最適化すべきパラメータが多いこと，しかも多くの場合局所最適解しか得られない．その点，SVMは最適解が一意に定まるという点で使いやすい．

SVMにも弱点はある

以上みてきたように，SVMはマージン最大化基準を採用した識別手法であり2次計画法を解くことにより最適な識別関数が得られること，カーネルトリックを利用して容易に非線形へ拡張できることの2つを特色とする．その振る舞いと識別性能に関する研究は，現在，理論，実験の両面から世界中の研究者が取り組んでいる．これまでに，以下のような弱点が指摘されており，今後の研究課題とされている．

第1に，SVMは原理的に2クラスを識別する手法であって，文字認識など多クラスの識別にそのままの形では適用できない．複数の識別関数を組み合わせて多クラスの識別を実現することは可能だが，k最近傍法やニューラルネットワークのように多クラスを考慮に入れた識別関数の最適化をすることができない．

第2に，2次計画法を解くための計算量の問題がある．この2次計画法の計算量は式(23)，式(36)の$\mathbf{D}$の大きさに依存する．$\mathbf{D}$は学習パターン数(n)次の行列であるため，パターン数が多くなると計算量が深刻な問題となる．学習パターン数とともに得られる識別関数の性能は良くなるのでなおさらである．

第3に，カーネルの選択の問題がある．問題に適したカーネルの明確な選択方法は知られていない．カーネルの最適型，カーネルの持つパラメータの最適値，式(18)のcの最適値などは，実験的に求めなければならない．ただし，式(18)のc，式(38)のp，式(39)のδの準最適な値を見つけることはそれほど大変ではない．

より詳しく学びたい人のために

以上本稿では，ここ数年国際的に関心が高まっている「新しい」パターン認識手法，サポートベクトルマシン（SVM）のポイントを概観した．SVMやパターン認識についてより詳しく知りたい方のために参考文献をいくつかあげておこう．

SVMに関するよく書けたレビューに文献7)，8)，日本語で読める解説記事に文献9）がある．簡単にポイントをつかみたい人には文献1）の該当部分，文献10）の序章が良い．要領よくまとまった教科書としては文献11）を薦める．背景となる理論，歴史を知るには考案者Vapnik自身の手になる大著[3)]がある．最新の研究成果を知るには，SVM関連技術の情報をまとめたホームページ（http://www.kernel-machines.org/）や学習理論を中心テーマとする国際会議NIPS（Neural Information Processing Systems）の会議資料（MIT Pressより刊行，http://nips.djvuzone.org/で閲覧可能）が参考になる．また，パターン認識全般を学びたい人には第2版が出たばかりの世界的名著[1)]を，日本語のものでは文献4）を推す．

SVMの骨格を成す複数のアイディアが発表されたのは30年以上前に遡る．1990年代にそれらが有機的に結び付いてSVMという成果として結実し，90年代後半になって脚光を浴びるに至った．この過程の最後では，Vapnikをはじめとする数人の研究者が中心的役割を果たした．日本で盛んになるのはさらにその数年後のことである．その流れを垣間見ていると，研究の潮流を作る上での理論研究と実証研究とのバランス，そして国際的な研究人脈の重要性を強く感ずる．本稿をまとめながら自戒の念とともに感じた率直な感想である．最後に，本稿の執筆にあたり貴重な助言をいただいた同僚諸氏に感謝いたします．

参考文献
1) Duda, R.O., Hart, P.E. and Stork, D.G.: Pattern Classification, John Wiley & Sons, 2nd edition (2000).
2) 今野　浩, 山下　浩: 非線形計画法, 日科技連 (1978).
3) Vapnik, V.: The Nature of Statistical Learning Theory, Springer (1995).
4) 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 (1998).
5) 津田宏治: ヒルベルト空間における部分空間法, 信学論 (D-II), J82-DII, 4, pp.592-599 (1999).
6) 前田英作, 村瀬　洋: カーネル非線形部分空間法によるパターン認識, 信学論 (D-II), J82-DII, 4, pp.600-612 (1999).
7) Burges, C.: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, pp.1-47 (1998).
8) Osuna, E., Freund, R. and Girosi, F.: Support Vector Machines: Training and Applications, A.I. Memo 1602, MIT A. I. Lab. (1997).
9) 津田宏治: サポートベクトルマシンとは何か, 信学会誌, 83, pp.460-466 (2000).
10) Schölkopf, B., Burges, C. and Smola, A. eds.: Advances in Kernel Methods: Support Vector Learning, MIT Press (1998).
11) Cristianini, N. and Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge Univ. Press (2000).

（平成13年5月31日受付）