サポートベクターマシンの概要

小野田　崇

1. はじめに

本稿では機械学習の研究領域で，最近注目をされているパターン認識の一手法である Support Vector Machine(以下，SVM)[1]の概要について述べる．SVM の研究は，80 年代から 90 年代にかけて注目されたニューラルネットワークに比べて優れたパターン認識結果が報告 (参考文献 [2] など) されて以来，数多くの理論的な研究が急速に行なわれてきた [3]．本稿では，Boser, Guyon, Vapnik の文献 [4], Guyon, Boser, Vapnik の文献 [5], Cortes, Vapnik の文献 [6] および Vapnik の文献 [7, 1] に基づき，SVM のアルゴリズムを中心に紹介する．

以下，第 2 章では識別関数の構造と VC 次元との関連について述べる．第 3 章では，線形 SVM について Hard Margin の場合と Soft Margin の場合に分けて紹介する．第 4 章では，第 3 章の線形 SVM の非線形への拡張と 2 クラス分類問題で議論される SVM を多クラス分類問題へ適用する場合について簡単に紹介する．第 5 章「おわりに」では，最近の SVM 関連研究情報の収集が可能なウエッブサイトを紹介する．

2. 超平面集合の構造

本章では，Structual Risk Minimization の概念 [7, 1] から導出されるひとつの学習アルゴリズムとその関数構造について述べる．

まず，SVM の中核をなす超平面識別関数の構造について考える．いま，内積空間 F およびパターンベクトル集合 $\mathbf{z}_1, \ldots, \mathbf{z}_r$ が与えられたとすると，任意の超平面識別関数は次のように表現される．

$$\{\mathbf{z} \in F : (\mathbf{w} \cdot \mathbf{z}) + b = 0\}. \tag{1}$$

この式 (1) は，自由度として係数 $\mathbf{w}$ と非負値である b をパラメータとして有している．式 (1) を図示すると

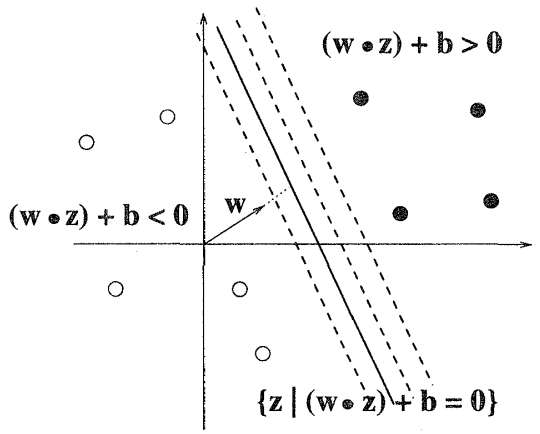

図 1: 線形識別関数

図 1 のようになる．図 1 は 2 次元の観測空間にデータが観測された様子を表している．ここで，白い円と黒い円とを分類したいとする．しかし，式 (1) からだけでは図 1 の実線や破線のように，いくつもの線形識別関数が導かれてしまう．

そこで，以下の式で表現される制約を加えることによって，識別関数となる超平面を $(\mathbf{w}, b) \in F \times \mathbf{R}$ を有する関数に一意に決める．

$$\min_{i=1,\ldots,r} |(\mathbf{w} \cdot \mathbf{z}_i) + b| = 1. \tag{2}$$

つまり，この制約によって $\mathbf{w}$ と b は距離 $1/\|\mathbf{w}\|$ を持つ超平面に最も接近するデータ点を表現することとなる．その様子を示したのが図 2 である．従って，2 クラス分類問題の場合，超平面間を垂直に測ったマージン (開き) [1] は少なくとも $2/\|\mathbf{w}\|$ となる．この超平面集合上の一つの構造の存在は，Vapnik の導出した以下の結果により確認することができる [7].

命題 1 R を点 $\{\mathbf{z}_1, \ldots, \mathbf{z}_r\}$ を含む最も小さい球 $B_R(\mathbf{a}) = \{\mathbf{z} \in F : \|\mathbf{z} - \mathbf{a}\| < R\}(\mathbf{a} \in F)$ の半径とし，

おのだ　たかし　(財) 電力中央研究所　情報研究所
〒 201-8511 狛江市岩戸北 2-11-1

[1] 以下,「マージン上」と言う場合は、マージン (開き) を測定する際の超平面上を意味する．

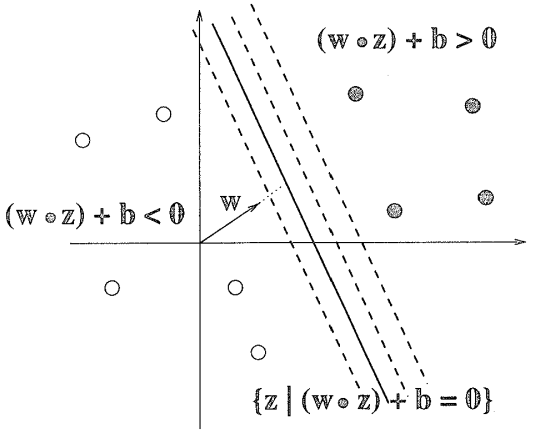

図 2: 制約付線形識別関数

次式がこれらの点を定義する識別関数であるとする.

$$f_{\mathbf{w},b} = \mathrm{sgn}((\mathbf{w}\cdot\mathbf{z}) + b) \tag{3}$$

そのとき，関数集合 $\{f_{\mathbf{w},b} : \|\mathbf{w}\| \le A\}$ は次式を満たす VC-次元 h を有する.

$$h < R^2A^2 + 1. \tag{4}$$

上記定理において，条件 $\|\mathbf{w}\| \le A$ を省くと VC-次元が N_F+1 となる関数の集合を導くことができる．ただし，N_F は空間 F の次元を表す．つまり，条件 $\|\mathbf{w}\| \le A$ により N_F より小さな VC-次元を得ることが可能であり，結果的に高次元空間で識別問題を取り扱うことができる.

命題 1 は次のように解釈できる.

1. マージンと $\|\mathbf{w}\|$ が反比例の関係にあるので，小さいマージンを大きくできれば，小さい VC-次元となることが式 (4) からわかる.

2. 識別を小さなマージンで行なう場合，より大きいクラスの識別問題を扱うことができる.

訓練サンプルから高い汎化能力[2] を有する学習機械を実現するには，訓練サンプルに対する誤識別と VC-次元の両方を小さくする必要がある [7]. つまり，超平面識別関数はマージンを最大化し，同時に可能な限り訓練サンプルを識別できる関数である必要がある．このマージン最大化と訓練サンプルの学習については第 3 章で述べる.

[2]非訓練サンプルの識別に対しても高い識別能力を持つということ.

3. 線形 SVM

3.1 Hard Margin SVM

訓練サンプル $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_\ell, y_\ell), z_i \in F, y_i \in \{\pm 1\}$ が与えられ，次式を満たす識別関数 $f_{\mathbf{w},b} = \mathrm{sgn}((\mathbf{w}\cdot\mathbf{z}) + b)$ を推定する問題を考える.

$$f_{\mathbf{w},b}(\mathbf{z}_i) = y_i, \quad i = 1, \ldots, \ell. \tag{5}$$

この関数が存在すれば，式 (2) の制約は次のように表現できる.

$$y_i \cdot ((\mathbf{z}_i \cdot \mathbf{w}) + b) \ge 1, \quad i = 1, \ldots, \ell. \tag{6}$$

$(\mathbf{w}, b), (-\mathbf{w}, -b)$ のように $\mathbf{w}$ と b の方向の違いにより，同じ超平面識別関数の式が 2 つ存在することとなる．しかし，式 (5) と式 (6) によって識別関数は一意に定めることができる.

命題 1 によって，汎化能力の高い識別関数は式 (6) で表現されるの制約条件の下，次式を最小化することで推定できる.

$$\tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2. \tag{7}$$

この凸最適化問題を解くため，式 (7) の Lagrangian を計算すると

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{\ell} \alpha_i (y_i((\mathbf{z}_i\cdot\mathbf{w}) + b) - 1), \tag{8}$$

ここで，$\alpha_i \ge 0$ は Lagrange 乗数である．この Lagrangian を α_i について最大化し，$\mathbf{w}$ と b について最小化する．パラメータ $\mathbf{w}$ と b についての L の導関数は鞍点において次式のように 0 にならなければならないので，

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \quad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0. \tag{9}$$

式 (9) から次式が成立する.

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \tag{10}$$

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{z}_i. \tag{11}$$

結局，$\mathbf{w}$ は訓練サンプルの展開式となる．$\mathbf{w}$ の解はただ一つに決まるが，係数 α_i はその必要がない.

Karush-Kuhn-Tucker 条件により，鞍点において Lagrange 乗数 α_i は式 (6) を正確に表現し直した次式の制約条件に対して非ゼロでなくてはならない.

$$\alpha_i \cdot [y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) - 1] = 0, \quad i = 1, \ldots, \ell. \tag{12}$$

$\alpha_i > 0$ を有するパターン $\mathbf{z}_i$ を *Support Vectors* と呼ぶ．式(12)より，*Support Vectors* はマージン上に存在することとなる．*Support Vectors* 以外の訓練サンプルは凸最適化問題の解法には関係のないものとなる．つまり，*Support Vectors* 以外の訓練サンプルは式(6)の制約条件を自動的に満たし，式(11)の展開項の部分には現れないのである．

この凸最適化問題を解いて得られる超平面識別関数の汎化能力については，以下の命題が成立する[1]．

命題 2 サンプル数 ℓ の訓練サンプルから得られる *Support Vectors* 数の期待値を $\ell - 1$ で割った値は，非訓練サンプルに対する誤識別率の上限である．

式(8)の Lagrangian に式(10)、式(11)の条件を代入すると，双対問題となる次の凸最適化問題を得ることができる．

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) \\ \text{subject to} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, \ell, \\ & \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{aligned} \tag{13}$$

式(11)の展開式を識別関数の式(5)に代入することによって，式(5)の識別関数を分類されるパターンと *Support Vectors* との内積で評価される次式に書き換えることができる．

$$f(\mathbf{z}) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} \alpha_i y_i (\mathbf{z} \cdot \mathbf{z}_i) + b \right). \tag{14}$$

以上より，式(13)で表現される凸二次計画問題を解くことで，識別関数 $f_{\mathbf{w},b}(\mathbf{z}) = \operatorname{sgn}((\mathbf{w} \cdot \mathbf{z}) + b)$ を得ることができる．これが基本となる線形SVMである．

3.2 Soft Margin SVM

現実問題としては，訓練サンプルを完全に分離できる超平面は存在しない場合が多い．そのような場合，次式で表現される緩和変数を導入して式(6)を満たさない訓練サンプルが存在しても良いようにする[6]．

$$\xi_i \geq 0, \quad i = 1, \ldots, \ell. \tag{15}$$

この緩和変数を使って式(6)の制約条件を次式のように緩和できる．

$$y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, \ell. \tag{16}$$

この緩和変数の導入によって，式(7)と式(6)で表現される凸最適化問題が次式のようになる．

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \tau(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{\ell} \\ \text{subject to} \quad & y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, \ell. \end{aligned} \tag{17}$$

目的関数の右辺第一項は，識別関数クラスのVC-次元の最小化に関連することが式(4)よりわかる．一方，$\sum_{i=1}^{\ell} \xi_i$ は訓練サンプル中で誤識別されるパターンの上限値である．適切な正定数 γ を選択できるとすれば，式(17)で表現される凸最適化問題は任意の関数集合における Structual Risk Minimization の概念の実践的な方法となる[7]．

訓練サンプルが完全に分離できる場合の式(11)と同様に，式(17)の最適解において，$\mathbf{w}$ は次式のように訓練サンプルの展開式となる．

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{z}_i. \tag{18}$$

ここで，係数 α_i が非ゼロとなるのは、訓練サンプル $(\mathbf{z}_i, y_i)$ が制約条件式(16)を満たす場合である．式(17)で表現される最適化問題の双対問題となる以下の凸二次計画問題を解くことで，係数 α_i を求めることができる．

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, \ell, \\ & \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{aligned} \tag{19}$$

Karush- Kuhn-Tucker 条件から，式(19)で表現される凸二次計画問題の最適解は次の条件を満たす．

$$\begin{aligned} \alpha_i = 0 \quad &\Rightarrow \quad y_i f(\mathbf{z}_i) \geq 1 \\ 0 \leq \alpha_i \leq \gamma \quad &\Rightarrow \quad y_i f(\mathbf{z}_i) = 0 \\ \alpha_i = \gamma \quad &\Rightarrow \quad y_i f(\mathbf{z}_i) \leq 1 \end{aligned} \tag{20}$$

この条件より，識別結果 $\operatorname{sgn}(f(\mathbf{z}_i))$ が y_i と一致していて，マージン値 $y_i f(\mathbf{z}_i)$ が1より大きいサンプルに対応する α_i は0になることがわかる．

4. 線形SVMの拡張

4.1 非線形SVMへの拡張

第3章では線形SVMについて述べた．しかし，線形SVMは線形分離可能な場合には高い汎化能力を達

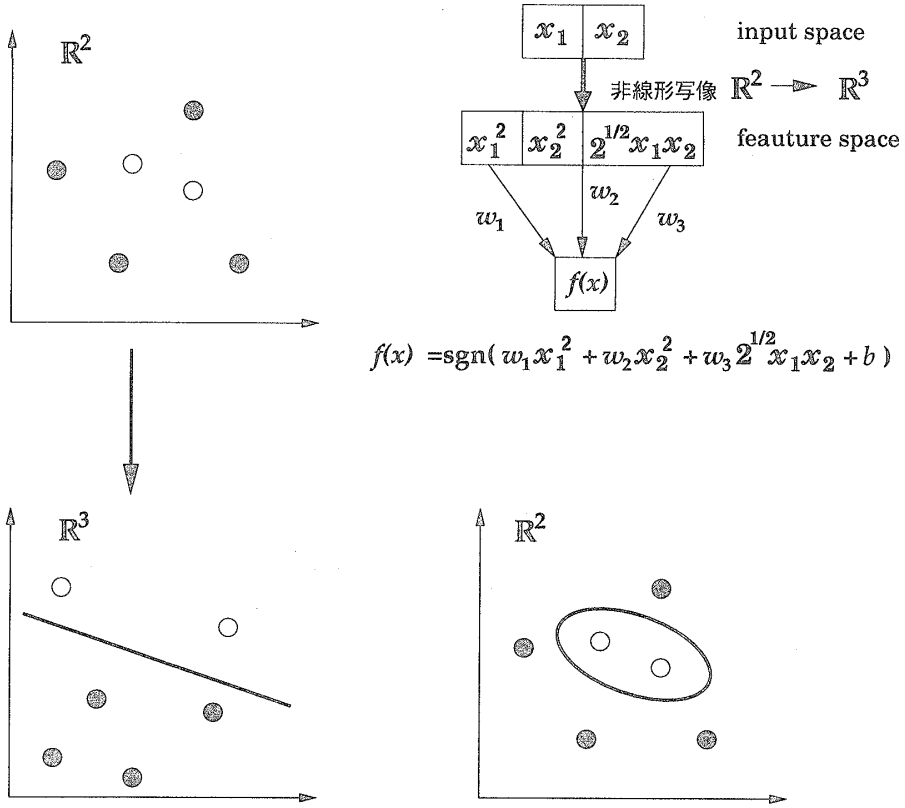

図 3: 非線形 SVM の原理

成できるが，実際の問題では線形分離可能な場合は多くない．そこで，より一般的な識別関数を推定するため，前処理として入力ベクトル $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ を次式のように高次元特徴空間に写像し，その後，その特徴空間で線形 SVM を行うという方法が考えられる．

$$\Phi : \mathbf{x}_i \mapsto \mathbf{z}_i. \tag{21}$$

本章で用いる $\mathbf{z}_i$ と第 3 章で用いた $\mathbf{z}_i$ とは異なり，$\mathbf{z}_i$ は観測された入力パターン $\mathbf{x}_i$ を高次元特徴空間に写像した結果であることに注意されたい．

式 (19) で表現される凸二次計画問題の目的関数を最大化し，式(14)で表現される識別関数を推定するには，高次元空間での以下の内積を計算する必要がる．

$$(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)). \tag{22}$$

式 (22) で表現される内積の計算には膨大な計算が必要となる．Mercer の条件の下，元の観測空間で定義される次式を満たすカーネル関数を用いて，膨大な計算を削減できる．

$$(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)) = k(\mathbf{x}, \mathbf{x}_i). \tag{23}$$

このカーネル関数を用いると，高次元特徴空間での式 (14) に相当する識別関数は次のようになる．

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{\ell} y_i \alpha_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b\right). \tag{24}$$

結局，ユークリッド空間の内積に代わって適切なカーネル関数 k を選択できれば、このカーネル関数 k に基づく非線形SVMには，前章で述べた線形SVMの特性が全て適用できる．図 3 に非線形 SVM の原理を示す．図 3 では，入力空間 (ここでは $\mathbb{R}^2$) 上のデータ (上図左) を非線形の写像を使ってより高次の特徴空間 (ここでは $\mathbb{R}^3$) にマッピングし，特徴空間上で分離可能な超平面を作成することで (下図左)，入力空間では非線形の識別関数になる (下図右)SVM(上図右) が構成できる様子を示している．

また，非線形 SVM は様々なカーネル関数を利用して，多様な学習機械を構成できる．図 4 に非線形SVM の構造を示した．図 4 中のカーネル関数 k には，以下で述べるようなカーネル関数を利用することが可能である．また，SVM を構成するためのパラメータは凸二次計画問題を解くことで推定することが可能で

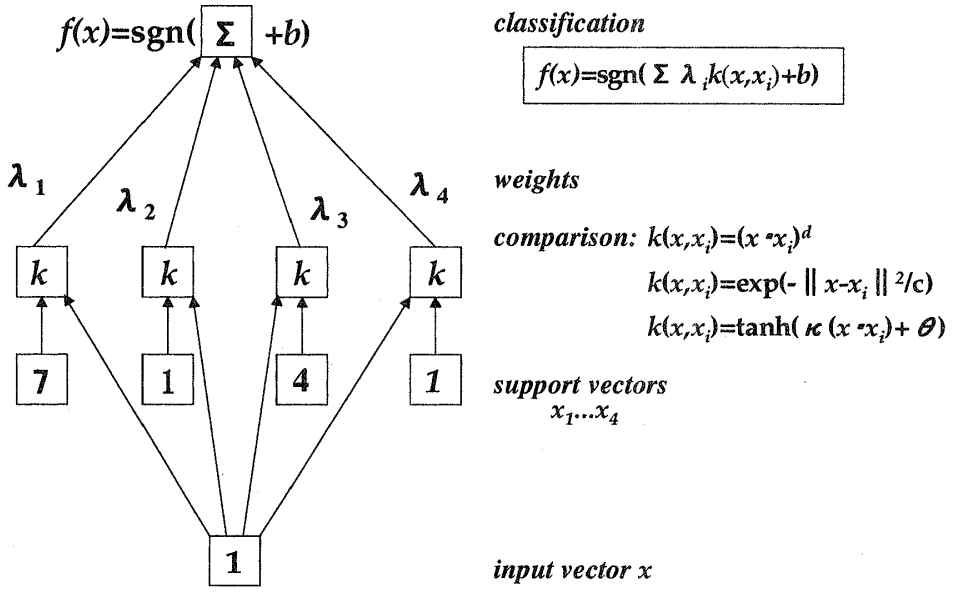

図 4: 非線形 SVM の構造

ある．図 4 中の第一層の $\mathbf{x}_i$ は訓練サンプル集合の補集合 (*Support Vectors*) であり，第二層の λ は Lagrange 乗数 α_i から $\lambda = y_i\alpha_i$ で計算できる．以下に一般的に利用されているカーネル関数を紹介しておく．

Polynomial カーネル関数

$$k(\mathbf{x},\mathbf{x}_i) = (\mathbf{x}\cdot\mathbf{x}_i)^d. \tag{25}$$

Radial basis カーネル関数

$$k(\mathbf{x},\mathbf{x}_i) = \frac{\exp(-\|\mathbf{x}-\mathbf{x}_i\|)^2}{c}. \tag{26}$$

Sigmoid カーネル関数

$$k(\mathbf{x},\mathbf{x}_i) = \tanh(\kappa\cdot(\mathbf{x}\cdot\mathbf{x}_i)+\theta). \tag{27}$$

式 (24) で表現される識別関数を求めるには，以下の最適化問題を解けばよい．

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i\cdot\mathbf{x}_j) \\ \text{subject to} \quad & 0\le\alpha_i\ge\gamma, \quad i=1,\ldots,\ell, \\ & \sum_{i=1}^{\ell}\alpha_i y_i = 0. \end{aligned} \tag{28}$$

カーネル関数 k は Mercer の条件を満たす必要がある．つまり，式 (23) のように高次元特徴空間での内積にカーネル関数 k は一致する必要がある．このことから，$K_{ij} = (y_i y_j k(\mathbf{x}_i,\mathbf{x}_j))_{ij}$ は正値行列となる．つまり，

$$\begin{aligned} &\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \\ &= \left(\sum_{i=1}^{\ell}\alpha_i y_i\Phi(\mathbf{x}_i)\cdot\sum_{j=1}^{\ell}\alpha_j y_j\Phi(\mathbf{x}_j)\right) \ge 0. \end{aligned} \tag{29}$$

式 (16) から $\xi_j = 0$ である *Support Vector* $\mathbf{x}_j$ に対し，次式が成立する．

$$\sum_{i=1}^{\ell} y_i\alpha_i\cdot k(\mathbf{x}_j,\mathbf{x}_i) + b = 0. \tag{30}$$

以上より，変数 b は次式のように，$0\le\alpha_j\le\gamma$ となる α_j を有する *Support Vector* の平均によって得ることができる．

$$b = y_j - \sum_{i=1}^{\ell} y_i\alpha_i\cdot k(\mathbf{x}_j,\mathbf{x}_i). \tag{31}$$

4.2 多クラス問題への拡張

ここまでは，全て 2 クラス分類問題の場合についての議論である．本節では，多クラス問題を SVM で扱う一般的な方法について簡単に紹介する．

k クラスの分類問題を解くため，SVM では 1 クラスとその他の残りのクラスとを識別する 2 クラス分

類SVMの集合 $f^1,\ldots,f^k$ を生成する．そして，符合関数を適用する前に，次式のような最大出力に従って多クラス分類を行ない，2クラス分類SVMの集合 $f^1,\ldots,f^k$ を統合しておく．

$$\underset{j=1,\ldots,k}{\operatorname{argmax}}\, g^j(\mathbf{x}), \tag{32}$$

ここで，$g^j(\mathbf{x})$ は次式で表現される．

$$g^j(\mathbf{x}) = \sum_{i=1}^{\ell} y_i \alpha_i^j \cdot k(\mathbf{x},\mathbf{x}_j) + b^j. \tag{33}$$

最終的に符合関数を適用する際は，以下のようになる．

$$f^j(\mathbf{x}) = \operatorname{sgn}(g^j(\mathbf{x})). \tag{34}$$

興味深いことにこの式(33)で表される関数 $g^j(\mathbf{x})$ の値は，棄却決定にも適用することができる．例えば，識別の信頼性測定として，関数 $g^j(\mathbf{x})$ の値の最大値と2番目に大きい値との差を考えれば，その差に基づき棄却決定を行なうことができるのである．

5. おわりに

本稿で最新の機械学習手法の一つとして紹介したSVMは，ニューラルネットワークに比べて優れたパターン認識結果が報告されて以来，様々な研究者に注目され，盛んに理論的研究がなされてきた．しかし，最適化理論の立場から機械学習を扱う研究は，まだ未開拓であり，様々な研究課題が残されている。

本稿での解説はSVMのさわりに過ぎない，SVMと最適化手法との関連の研究についての詳細を知りたい方は， www.kernel-machines.org を参照されたい．このホームページには SVM に関する研究の論文やSVM用の最適化手法のプログラムなどが掲載されている．また，最適化手法に関わる他の機械学習手法について知りたい方は，boosting.org を参照されたい．このホームページには，本稿では触れなかった最新機械学習法の一つアンサンブル学習と最適化問題との関連についての論文が充実している．

参考文献

[1] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[2] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Sign. Processing*, Vol. 45, pp. 2758 – 2765, 1997.

[3] 小野田崇．電気事業における最新機械学習技術の適用可能性． 調査報告書，(財) 電力中央研究所，1999.

[4] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pp. 144–152, Pittsburgh, PA, 1992. ACM Press.

[5] B.E. Boser, I.M. Guyon, and V.N. Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pp. 147–155, San Mateo, CA, 1993. Morgan Kaufmann.

[6] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, Vol. 20, pp. 273 – 297, 1995.

[7] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.