日本オペレーションズ・リサーチ学会和文論文誌
Transactions of the Operations Research Society of Japan
Vol. 55, 2012, pp. 110–131

不定値カーネルを伴うサポートベクターマシンに対するカーネル最適化

木村 彩英子
東海旅客鉄道株式会社

矢部 博
東京理科大学

(受理 2011 年 8 月 1 日; 再受理 2012 年 5 月 2 日)

和文概要　サポートベクターマシン (SVM) は 2 値分類問題を解くための手法である．SVM において最も重要であるのは，適したカーネル行列の選び方であり，一般に半正定値のものが用いられる．しかし不定値のカーネル行列がデータの特性を表していることも多く，その場合の SVM のための様々な手法が考えられている．本論文では，その手法の一つである Luss and d'Aspremondt のモデルを紹介し，その解法として射影勾配 BB 法を提案する．また，Luss らのモデルは計算量の点で問題があることから，その修正として新たなモデルを定式化し，数値実験により識別能力を評価する．

キーワード: 最適化，数理計画，サポートベクターマシン，カーネル最適化モデル

1. はじめに

サポートベクターマシン（SVM）はパターン認識問題における 2 値分類手法の一つである．SVM はカーネルトリックと呼ばれる方法を用いることで，非線形の識別関数を構成できる．効率のよい識別関数を与えるためにはデータに適したカーネル行列を選ぶことが重要である．カーネル行列はデータ同士の類似度を測るカーネル関数によって構成される行列であり，通常，半正定値行列に選ばれることが多い．この半正定値条件は Mercer 条件 (Vapnik [19]) として知られている．

しかし現実には，データの特徴を反映したカーネル関数は半正定値にならないことがある．たとえば，生物情報科学における DNA とたんぱく質配列との類似度を測るための Smith-Waterman and BLAST scores [1] や，画像分類のための Tangent Distance Kernel [8] や Simpson Score [10] が挙げられる．そのため不定値のカーネル行列を伴う SVM に対する手法が近年考えられている．

不定値カーネルを伴う SVM のための手法として，不定値のカーネル行列を直接変形する方法が考えられていた．元の固有ベクトルを保存したまま固有値を変形することにより半正定値行列に緩和する手法が Wu et al. [20] によって提案されている．また，Lin and Lin [12] は SVM の主問題にカーネルを陽に用いて，目的関数のカーネル行列を単位行列に置き換えている．この手法は良い識別能力を達成するが，単位行列による置き換えに対する解釈は困難である．一方 Haasdonk [7] は，不定値カーネルによる分類問題を擬ユークリッド空間における 2 つの凸包の距離を最小化する問題として定式化し，その幾何学的解釈を与えた．別のアプローチとして，Luss and d'Aspremont [14] は不定値行列の近傍で半正定値行列を探すモデルを提案している．モデルを解く際には射影勾配法を利用しているが，探索におけるステップ幅は与えるパラメータの値に依存し，その与え方に根拠はない．そのため本論文で

は，射影勾配法において Barzilai-Borwein 法 [2] の適用を提案し，収束を加速することを試みる．また， Luss and d'Aspremont のモデルは計算量の点において問題があることを指摘し，その解決策として新たなカーネル最適化モデルを提案する．

本論文の構成は以下の通りである．2 節で SVM およびカーネル最適化問題を示す．3 節では Luss and d'Aspremont のモデルと，新たなモデルの定式化を与える．4 節では計算機による数値実験の結果を示して，提案した数値解法や新しいカーネル最適化モデルの有効性を検証する．

本論文を通じて，次のような記号を用いる: $\boldsymbol{S}^n$ は n 次対称行列全体の集合とし，$\boldsymbol{S}^n_+$ は n 次半正定値対称行列全体の集合とする．行列 X が与えられたとき，$\lambda_i(X)$ は X の i 番目の固有値とする．X_+ は行列 X を $\boldsymbol{S}^n_+$ に射影して得られる行列である．すなわち X のスペクトル分解 $X = \sum_i \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$ に対して $X_+ = \sum_i \max(0, \lambda_i) \boldsymbol{v}_i \boldsymbol{v}_i^T$ と定義する (ただし，$\boldsymbol{v}_i$ は λ_i に属する正規固有ベクトルである)．また，$X \in \boldsymbol{S}^n_+$ を $X \succeq O$ と表す．ベクトル $\boldsymbol{e}$ は成分がすべて 1 のベクトルとし

$$\boldsymbol{e} = (1, \ldots, 1)^T \tag{1}$$

で定義する．trace は行列の対角成分の和，X^T は X の転置を表す．ベクトル $\boldsymbol{w}, \boldsymbol{v}$ の内積を $\langle \boldsymbol{w}, \boldsymbol{v} \rangle = \boldsymbol{w}^T \boldsymbol{v} = \sum_i w_i v_i$ で定義する．行列のノルム $\|\cdot\|$ はフロベニウスノルムを表し，ベクトルに対して $\|\cdot\|_1$ と $\|\cdot\|_2$ はそれぞれ 1 ノルム，2 ノルムを表す．

2. カーネル最適化モデル

2.1 節では SVM の定式化を行い，2.2 節では SVM で用いられるカーネル行列の最適化問題を紹介する．

2.1. クラス判別とサポートベクターマシン

SVM は 2 値分類問題の手法の一つである．2 値分類問題は入力空間 $\mathcal{X} \subset \mathbb{R}^m$ に属する n 個のデータ点 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ とその所属クラス $y_1, \ldots, y_n \in \{+1, -1\}$ を学習し，未知データ $\boldsymbol{x} \in \mathcal{X}$ が与えられたときに学習データに基づき，その所属するクラスを判別することを目的とする．線形 SVM では，アフィン関数 $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ を考える．ここで係数 $\boldsymbol{w} \in \mathbb{R}^m$ は重みベクトル，b は閾値と呼ばれるパラメータである．識別超平面を $\{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) = 0\}$ として，$f(\boldsymbol{x}) \geq 0$ ならばクラスは $+1$，$f(\boldsymbol{x}) < 0$ ならばクラスは -1 と識別する．

以下の 2.1.1 節ではハードマージン SVM，1 ノルムソフトマージン SVM をそれぞれ紹介する．そして 2.1.2 節では非線形の SVM，特に 1 ノルムソフトマージン非線形 SVM について説明する．

2.1.1. ハードマージン SVM と 1 ノルムソフトマージン SVM

学習データが超平面により完全に分離できるとき，ハードマージン SVM は次の凸 2 次計画問題として定式化される．

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b} & \dfrac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle \\ \text{s.t.} & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad i = 1, \ldots, n \end{array} \tag{2}$$

問題 (2) を解いて得られた $\boldsymbol{w}^*, b^*$ を用いて，識別関数 $\hat{f}(\boldsymbol{x}) = \mathrm{sgn}(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle + b^*)$ を得ることができる．

一方，超平面で学習データを完全に分離できない場合には，非負の緩和変数 ξ_i を用いて問題 (2) を以下のように書き換える．

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\langle \boldsymbol{w},\boldsymbol{w}\rangle + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle + b) \geq 1-\xi_i, \quad i=1,\ldots,n \\ & \xi_i \geq 0, \quad i=1,\ldots,n \end{aligned} \tag{3}$$

ここで C はペナルティパラメータであり，$\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^T$ である．そして，問題 (3) のラグランジュ双対問題は次のような凸 2 次計画問題になる．

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle \\ \text{s.t.} \quad & \boldsymbol{0} \leq \boldsymbol{\alpha} \leq C\boldsymbol{e}, \quad \boldsymbol{\alpha}^T\boldsymbol{y}=0 \end{aligned} \tag{4}$$

ただし,

$$\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)^T, \quad \boldsymbol{y}=(y_1,\ldots,y_n)^T$$

であり，$\boldsymbol{0}$ は零ベクトル，$\boldsymbol{e}$ は (1) で定義される．双対問題 (4) の最適解を $\boldsymbol{\alpha}^*$ とすると主問題 (3) の最適解 $\boldsymbol{w}^*, b^*$ はそれぞれ

$$\boldsymbol{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \boldsymbol{x}_i, \qquad b^* = -\frac{1}{2}\Big(\max_{i:y_i=-1}\langle \boldsymbol{w}^*,\boldsymbol{x}_i\rangle + \min_{i:y_i=1}\langle \boldsymbol{w}^*,\boldsymbol{x}_i\rangle\Big) \tag{5}$$

で求められ，したがって識別関数は $\hat{f}(\boldsymbol{x})=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i\langle \boldsymbol{x},\boldsymbol{x}_i\rangle + b^*\right)$ で与えられる．

2.1.2. 非線形 SVM

前節では入力空間 $\mathcal{X}$ でデータ点を超平面で分離することを考えた．しかし一般のデータ点は，入力空間で超平面分離を行うことは不可能である．そのようなデータ点の集合に対しては，写像 Φ によって高次元の特徴空間 $\mathcal{F}\subset\mathbb{R}^l$ に写し，その空間で超平面分離することを考える．これは結果的に入力空間における非線形関数による分離となる．以下では，特徴空間において 1 ノルムソフトマージン SVM を行うことを考える．

いま，写像

$$\Phi : \mathcal{X}\subset\mathbb{R}^m \to \mathcal{F}\subset\mathbb{R}^l$$

が与えられているとする．このとき前節にならえば，学習データ $\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ に対し 1 ノルムソフトマージン SVM は

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\langle \boldsymbol{w},\boldsymbol{w}\rangle + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w},\Phi(\boldsymbol{x}_i)\rangle + b) \geq 1-\xi_i, \quad i=1,\ldots,n \\ & \xi_i \geq 0, \quad i=1,\ldots,n \end{aligned} \tag{6}$$

と定式化できる．さらに，ラグランジュ双対問題を求めると

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle \\ \text{s.t.} \quad & \boldsymbol{0} \leq \boldsymbol{\alpha} \leq C\boldsymbol{e}, \quad \boldsymbol{\alpha}^T\boldsymbol{y}=0 \end{aligned} \tag{7}$$

となる．問題 (7) を解くためには，高次元特徴空間上で内積 $\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ を計算する必要があるが，これは膨大な計算が必要になる．そのため，次の関係式

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle \tag{8}$$

を満たすようなカーネル関数 k が考えられている．カーネル関数 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を (i,j) 成分に持つ n 次のカーネル行列 K を用いれば双対問題 (7) は

$$\begin{array}{ll} \max\limits_{\boldsymbol{\alpha}} & \boldsymbol{\alpha}^T \boldsymbol{e} - \dfrac{1}{2} \boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} \\ \text{s.t.} & \boldsymbol{0} \leq \boldsymbol{\alpha} \leq C\boldsymbol{e}, \quad \boldsymbol{\alpha}^T \boldsymbol{y} = 0 \end{array} \tag{9}$$

となる．ただし，Y は $Y = \mathrm{diag}\,(y_1, \ldots, y_n)$ で定義される対角行列である．この問題の最適解を $\boldsymbol{\alpha}^*$ とすれば，識別関数は

$$\hat{f}(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i^* y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b^*)$$

と導出される．ここで b^* は式 (5) によって求められる．

2.2. カーネル最適化モデル

一般に，識別関数を構成する $\mathcal{F}, \Phi$ を直接求めることは難しく，しかも特徴空間 $\mathcal{F}$ の次元は非常に大きくなる．しかしカーネル関数 k さえ与えられれば，関係式 (8) において写像 Φ を陽に用いることなく識別関数を構成することができる．効率のよい識別関数を構成するためには，効率のよいカーネル関数を与える必要がある．

前述したように，カーネル関数 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ は 2 つの学習データ $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ に対して定義される関数で，対称性 $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$ を満たす．カーネル関数が以下の半正定値性を満たすことは，数学的な理論構成の点で非常に役に立つ．

定義 2.1. $k(\boldsymbol{x}, \boldsymbol{x}')$ が任意の $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ に対して

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \quad \forall c_1, \ldots, c_n \in \mathbb{R}$$

を満たすならば，関数 k は半正定値であるという．

この条件を Mercer 条件といい，これを満たすようなカーネル関数を半正定値カーネルという．このとき，カーネル行列 K は半正定値対称行列になる．半正定値カーネルとして以下のような例がある ([6])．

RBF(Radial Basis Function) カーネル (ガウスカーネル)

$$k_G(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\Big) \qquad (\sigma > 0)$$

ラプラスカーネル

$$k_L(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1) \qquad (\gamma > 0)$$

p 次多項式カーネル

$$k_P(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p \qquad (c \geq 0,\ p \in \mathbb{N})$$

一方，次のようなカーネルも利用されているが，これらはMercer条件を満たすとは限らない．このようなカーネルを不定値カーネルと呼ぶ．

シグモイドカーネル

$$k_{Sg}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\theta_1 \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \theta_2) \qquad (\theta_1, \theta_2 \in \mathbb{R}) \tag{10}$$

シンプソンカーネル [10]

これは2つの白黒画像 $\boldsymbol{x}_i, \boldsymbol{x}_j$ に対して定められるカーネルである．

a： どちらの画像でも白であるピクセル数
b： 画像 $\boldsymbol{x}_i$ では白，画像 $\boldsymbol{x}_j$ では黒であるピクセル数
c： 画像 $\boldsymbol{x}_i$ では黒，画像 $\boldsymbol{x}_j$ では白であるピクセル数

とすれば，シンプソンカーネルは

$$k_{Sm}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{a}{\min\{a+b, a+c\}} \tag{11}$$

で与えられる．

データに基づいて客観的にカーネル行列を選ぶような最適化モデルを構築することは重要であり，以下ではSVMに対するそのようなカーネル最適化モデルを紹介する．まず次の仮定を与える．

仮定 2.2. ラベル $y_i (i = 1, \ldots, n)$ のすべてが1，あるいはすべてが -1 ではないとする．

$K \in \boldsymbol{S}^n_+$ ならば，双対問題 (9) は凸2次計画問題であり，仮定2.2の下では狭義実行可能解が存在する．したがってSlater条件が満たされる．

以下では，$\mathcal{S}_P$, $\mathcal{S}_D$ をそれぞれ主問題 (6)，双対問題 (9) の実行可能領域とし

$$\begin{aligned} \mathcal{S}_P &= \{(\boldsymbol{w}, b, \boldsymbol{\xi}) \in \mathbb{R}^l \times \mathbb{R} \times \mathbb{R}^n \mid y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i) \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \ i = 1, \ldots, n\} \\ \mathcal{S}_D &= \{\boldsymbol{\alpha} \in \mathbb{R}^n \mid \boldsymbol{0} \leq \boldsymbol{\alpha} \leq C\boldsymbol{e},\quad \boldsymbol{\alpha}^T \boldsymbol{y} = 0\} \end{aligned}$$

で定義する．もし主問題が実行可能であれば双対定理より次の関係が成り立つ．

$$\min_{(\boldsymbol{w}, b, \boldsymbol{\xi}) \in \mathcal{S}_P} \ \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{n} \xi_i = \max_{\boldsymbol{\alpha} \in \mathcal{S}_D} \ \boldsymbol{\alpha}^T \boldsymbol{e} - \frac{1}{2} \boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha}$$

双対問題の最適値を行列 K の関数

$$w_C(K) = \max_{\boldsymbol{\alpha} \in \mathcal{S}_D} \ \boldsymbol{\alpha}^T \boldsymbol{e} - \frac{1}{2} \boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} \tag{12}$$

とすれば，$\mathcal{K} \subset \boldsymbol{S}^n_+$ となる集合 $\mathcal{K}$ に対し

$$\min_{K \in \mathcal{K}} w_C(K) \tag{13}$$

を解くことにより，よりよいマージンを与えるカーネル行列 $K^* \in \mathcal{K}$ が得られる．これをカーネル最適化モデルという．

Lanckriet et al. [9] は，カーネルを既知のカーネルの線形結合に制限すれば問題 (13) は半正定値計画問題に変形されること，同様に非負結合に制限すれば2次制約2次計画問題に変形されることを示した．

3. カーネル行列が不定値である場合

カーネル行列 K が半正定値の場合，問題 (12) は凸 2 次計画問題になる．しかしながら実際の問題では，シグモイドカーネルやシンプソンカーネルのように，必ずしも半正定値であるとは限らないカーネルを扱うこともある．したがって，不定値カーネルが与えられた場合の SVM を考えることは応用上，意義のあることである．そこで本節では，こうした場合の対処法について既存研究を紹介する．

ここでは不定値カーネル行列 $K_0 \in \boldsymbol{S}^n$ が与えられているとする．

3.1. 半正定値カーネル行列への変換

以下では不定値カーネル行列 K_0 の固有値を変形することにより，新たに半正定値行列 $\tilde{K}$ をつくることを考える．

K_0 は実対称行列なので，直交行列 V と対角行列 $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$ によって $K_0 = V\Lambda V^T$ とできる．このとき $\tilde{\lambda}_i = \psi(\lambda_i)$ が非負となるような関数 $\psi(\cdot)$ を考え $\tilde{\Lambda} = \mathrm{diag}\,(\tilde{\lambda_1}, \ldots, \tilde{\lambda_n})$ とすれば，$\tilde{K} = V\tilde{\Lambda}V^T$ は半正定値となる．Wu et al. [20] は半正定値行列に変換できるような関数として，例えば以下の 3 つの手法 Denoise，Flip，Shift を提案している．

(i) Denoise: $\psi(\lambda) = \max(0, \lambda)$

これは負の固有値の全てをノイズとして扱い，それらを 0 に置き換える方法である．この方法によって得られる行列 $\tilde{K}$ は，元の不定値カーネル行列 K_0 にフロベニウスノルム $\|\cdot\|$ の意味で最も近い半正定値行列である．すなわち次の定理が成り立つ ([20])．

定理 3.1. 行列 $K_0 \in \boldsymbol{S}^n$ が与えられたとき，最小化問題

$$\begin{array}{ll} \min & \|K - K_0\|^2 \\ \text{s.t.} & K \succeq O \end{array}$$

の解は $K = (K_0)_+$ である．

(ii) Flip: $\psi(\lambda) = |\lambda|$

これは K_0 の固有値を対応する特異値で置き換える方法であるとみなすことができ，このとき $\|\tilde{K}\|_2 = \|K_0\|_2$ となる．

(iii) Shift: $\psi(\lambda) = \lambda + \eta$

これは，ある正の定数 η をすべての固有値に加える方法である．η の値は少なくとも $|\min_i \lambda_i(K_0)|$ を選べば $\tilde{K}$ は半正定値となる．しかし η は 1 ノルムソフトマージン SVM および 2 ノルムソフトマージン SVM におけるパラメータ C と密接に関係しているため，SVM に対して，$\eta = |\min_i \lambda_i(K_0)|$ とすることは必ずしも得策とは限らない [20].

3.2. 単位行列での置き換えによる修正

Lin and Lin [12] は不定値行列 K_0 が 1 ノルムソフトマージン SVM におけるカーネル行列として与えられているとき，主問題 (6) を修正することで，凸計画問題に変形している．その手法を紹介する．

いま $K_0 \succeq O$ ならば，$K_0 = \hat{K}^T\hat{K}$ と分解でき，$\hat{K} = [\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_n)]$ とみることができる．すると

$$y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i \ (i = 1, \ldots, n) \iff Y(\hat{K}^T\boldsymbol{w} + b\boldsymbol{e}) \geq \boldsymbol{e} - \boldsymbol{\xi}$$

である．ここで $\boldsymbol{c}$ を $\boldsymbol{w} = \hat{K}Y\boldsymbol{c}$ であるような変数とする．このとき $\|\boldsymbol{w}\|_2^2 = \boldsymbol{c}^T Y K_0 Y \boldsymbol{c}$, $\hat{K}^T\boldsymbol{w} =$

$K_0Y\boldsymbol{c}$ が成り立つことから，問題 (6) は

$$\begin{array}{ll}\min\limits_{\boldsymbol{c},b,\boldsymbol{\xi}} & \dfrac{1}{2}\boldsymbol{c}^TYK_0Y\boldsymbol{c}+C\boldsymbol{\xi}^T\boldsymbol{e}\\ \text{s.t.} & Y(K_0Y\boldsymbol{c}+b\boldsymbol{e})\geq\boldsymbol{e}-\boldsymbol{\xi},\ \ \boldsymbol{\xi}\geq\boldsymbol{0}\end{array}\tag{14}$$

と表される．問題 (14) において K_0 が不定値行列のとき，Lin and Lin [12] は，目的関数に含まれる行列 K_0 を単位行列に置き換え，かつ，$Y^2=I$(単位行列) であることを利用して，(14) を次の凸計画問題に変形した．

$$\begin{array}{ll}\min\limits_{\boldsymbol{c},b,\boldsymbol{\xi}} & \dfrac{1}{2}\|\boldsymbol{c}\|_2^2+C\boldsymbol{\xi}^T\boldsymbol{e}\\ \text{s.t.} & Y(K_0Y\boldsymbol{c}+b\boldsymbol{e})\geq\boldsymbol{e}-\boldsymbol{\xi},\ \ \boldsymbol{\xi}\geq\boldsymbol{0}\end{array}$$

ここで不定値行列 K_0 は制約条件にまだ残っていることに注意されたい．K_0Y の第 i 行ベクトルの転置が問題 (6) の $\Phi(\boldsymbol{x}_i)$ に対応していることよりこの問題のラグランジュ双対問題は，問題 (9) で $K=(K_0Y)(K_0Y)^T=K_0K_0^T$ とおいた凸計画問題になる．

3.3. Luss and d'Aspremont のモデル

本節では Luss and d'Aspremont [14] の提案するモデルを紹介し，彼らの用いた数値解法の加速を試みる．彼らは与えられたカーネル K_0 の近傍で半正定値カーネル行列を見つけることを目指して，(13) として次の問題を考えた．

$$\min_{\substack{K\succeq O\\ \|K-K_0\|^2\leq\beta^2}} \quad w_C(K)\tag{15}$$

パラメータ $\beta>0$ は K_0 との距離を制御するものである．問題 (15) は β が小さい値の場合には実行不可能になる可能性があるため，不等式制約の代わりにペナルティを課し，以下の問題に置き換えている．

$$\min_{K\succeq O}\ \max_{\boldsymbol{\alpha}\in\mathcal{S}_D}\ \boldsymbol{\alpha}^T\boldsymbol{e}-\frac{1}{2}\boldsymbol{\alpha}^TYKY\boldsymbol{\alpha}+\rho\|K-K_0\|^2\tag{16}$$

ただし ρ はカーネルの最適性と，元のカーネル行列 K_0 との距離のトレードオフを制御するペナルティパラメータである．ここで，角谷 (1941) による以下の定理が役に立つ ([5])．

定理 3.2 (minimax 定理)**.** A と B をノルム線形空間の凸部分集合で，かつ空でないコンパクトな集合であるとする．関数 $f:A\times B\to\mathbb{R}$ は連続で，すべての $\boldsymbol{b}\in B$ に対して $\boldsymbol{a}\to f(\boldsymbol{a},\boldsymbol{b})$ は A で凹であり，すべての $\boldsymbol{a}\in A$ に対して $\boldsymbol{b}\to f(\boldsymbol{a},\boldsymbol{b})$ は B で凸であるならば，

$$\min_{\boldsymbol{b}\in B}\max_{\boldsymbol{a}\in A}\ f(\boldsymbol{a},\boldsymbol{b})=\max_{\boldsymbol{a}\in A}\min_{\boldsymbol{b}\in B}\ f(\boldsymbol{a},\boldsymbol{b})$$

が成り立つ．

この定理を用いて，Luss らは問題 (16) で max と min を入れ替えて，

$$\max_{\boldsymbol{\alpha}\in\mathcal{S}_D}\ \min_{K\succeq O}\ \boldsymbol{\alpha}^T\boldsymbol{e}-\frac{1}{2}\boldsymbol{\alpha}^TYKY\boldsymbol{\alpha}+\rho\|K-K_0\|^2\tag{17}$$

を考えた．このとき定理 3.1 を利用すれば，問題 (17) の内部の最小化問題の最適解 K^* は次のように陽に求めることができる:

$$K^*=\Big(K_0+\frac{1}{4\rho}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T\Big)_+\tag{18}$$

また，ある直交行列 V と対角行列 D を用いて，$K_0 + \frac{1}{4\rho}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T = VDV^T$ と表せる．ただし，

$$\hat{\lambda}_i(\boldsymbol{\alpha}) := \lambda_i\big(K_0 + \frac{1}{4\rho}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T\big), \quad i = 1, \ldots, n$$

とすれば $D = \mathrm{diag}\,(\hat{\lambda}_1(\boldsymbol{\alpha}), \ldots, \hat{\lambda}_n(\boldsymbol{\alpha}))$ である．このとき式 (18) より $K^* = VD_+V^T$ となるので，結局，問題 (17) は次のように表せる．

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & g(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{e} - \rho\sum_i \hat{\lambda}_i(\boldsymbol{\alpha})\max\big(0, \hat{\lambda}_i(\boldsymbol{\alpha})\big) + \rho\ \mathrm{trace}\,(K_0^2) \\ \text{s.t.} \quad & \boldsymbol{\alpha} \in \mathcal{S}_D \end{aligned} \tag{19}$$

問題 (19) の目的関数の導関数を得るため，$\boldsymbol{\alpha}$ について微分可能であるような関数 $h(\boldsymbol{\alpha})$ に対して $\max\big(0, h(\boldsymbol{\alpha})\big)$ の平滑化と固有値の微分を考える．関数 $\max(0, h(\boldsymbol{\alpha}))$ は微分可能ではないため，ある $\varepsilon > 0$ に対して Moreau-Yosida 正則化を用いることで微分可能な以下の関数を定義する．

$$\phi_\varepsilon\big(h(\boldsymbol{\alpha})\big) := \max_{0\le u\le 1}\Big(uh(\boldsymbol{\alpha}) - \frac{\varepsilon}{2}u^2\Big) \tag{20}$$

ここで

$$u^*(\boldsymbol{\alpha}) := \operatorname*{argmax}_{0\le u\le 1}\phi_\varepsilon\big(h(\boldsymbol{\alpha})\big) = \begin{cases} 1, & h(\boldsymbol{\alpha}) \ge \varepsilon \\ h(\boldsymbol{\alpha})/\varepsilon, & 0 \le h(\boldsymbol{\alpha}) \le \varepsilon \\ 0, & h(\boldsymbol{\alpha}) \le 0 \end{cases}$$

を定義すれば，関数 (20) の導関数は $\nabla\phi_\varepsilon\big(h(\boldsymbol{\alpha})\big) = u^*(\boldsymbol{\alpha})\nabla h(\boldsymbol{\alpha})$ で表される．次に変数 $\boldsymbol{\alpha} \in \mathbb{R}^n$ に対して関数 $X : \mathbb{R}^n \mapsto \boldsymbol{S}^n$ を定義する．λ を $X(\boldsymbol{\alpha})$ の固有値，$\boldsymbol{v}$ を λ に属する固有ベクトルとすると，

$$\nabla_{\boldsymbol{\alpha}}\lambda = \left(\boldsymbol{v}^T\frac{\partial X(\boldsymbol{\alpha})}{\partial\alpha_1}\boldsymbol{v}, \ldots, \boldsymbol{v}^T\frac{\partial X(\boldsymbol{\alpha})}{\partial\alpha_n}\boldsymbol{v}\right)^T$$

が成り立つ．したがって，問題 (19) の目的関数 $g(\boldsymbol{\alpha})$ の代わりに

$$\tilde{g}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{e} - \rho\sum_i \hat{\lambda}_i(\boldsymbol{\alpha})\phi_\varepsilon\big(\hat{\lambda}_i(\boldsymbol{\alpha})\big)\big) + \rho\ \mathrm{trace}\,(K_0^2)$$

を考えれば，勾配ベクトル $\nabla\tilde{g}(\boldsymbol{\alpha})$ が得られるので，様々な勾配法を利用することができる．

Luss and d'Aspremont [14] は問題 (19) を解くための手法として，射影勾配法と解析的中心切除平面法を提案している．しかし後者は各反復で解析的中心を求めるための手間がかかることから，射影勾配法の方がより効果的であることが示されている．以下に制約付き最大化問題に対する射影勾配法のアルゴリズムを紹介する．

射影勾配法

step 1: 初期点 $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$ を選び，$\nu = 0$ とおく．

step 2: ステップ幅 t_ν を求めて，$\hat{\boldsymbol{\alpha}}_{\nu+1} = \boldsymbol{\alpha}_\nu + t_\nu\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)$ とおく．

step 3: $\boldsymbol{\alpha}_{\nu+1} = p_A(\hat{\boldsymbol{\alpha}}_{\nu+1})$ とおく．ただし，$p_A(\cdot)$ は実行可能領域 $\mathcal{S}_D$ への正射影である．

step 4: もし停止条件が満たされていれば停止し，さもなければ $\nu = \nu + 1$ として step 2 へいく．

ここで，t_ν は ν 回目の反復におけるステップ幅を意味している．一般の無制約最適化問題では，山登り法 (最小化問題に対する最急降下法に対応している) は Armijo 条件を満たす直線探索を実行すれば，大域的に収束することが知られている．今回の場合，直線探索は次のアルゴリズムで与えられる．

直線探索 (Armijo 条件)

step 1: 現在の近似解 $\boldsymbol{\alpha}_\nu$，パラメータ $0<\xi<1,\ 0<\tau<1$ を与える．

step 2: $\sigma_{\nu,0}=1,\ i=0$ とおく．

step 3: $g\big(\boldsymbol{\alpha}_\nu+\sigma_{\nu,i}\nabla\tilde{g}(\boldsymbol{\alpha})\big)\geq g(\boldsymbol{\alpha}_\nu)+\xi\,\sigma_{\nu,i}\,\|\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)\|^2$ ならば停止し，ステップ幅を $\sigma_{\nu,i}$ とする．さもなければ step 4 へいく．

step 4: $\sigma_{\nu,i+1}=\tau\sigma_{\nu,i},\ i=i+1$ として，step3 へいく．

しかしながら，Luss and d'Aspremont のモデル (19) は目的関数に固有値が陽に現れているので，t_ν を選ぶ際に直線探索を用いると，目的関数を評価する度に固有値の計算が必要となり，結果的に収束は非常に遅くなる．そのため，Luss らは直線探索を用いないステップ幅の選び方として，ある正の定数 δ に対して

$$t_\nu=\frac{\delta}{\nu+1} \tag{21}$$

を利用している．δ の値は収束の速さにとって重要であるが，その選び方には経験則は存在しないため，決して効率的であるとはいえない．

そこで本研究では，射影勾配法の step 2 において Barzilai-Borwein 法 (BB 法)[2] を適用した射影勾配 BB 法を利用することを提案する．BB 法では，$\boldsymbol{s}_\nu=\boldsymbol{\alpha}_\nu-\boldsymbol{\alpha}_{\nu-1},\ \boldsymbol{y}_\nu=\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)-\nabla\tilde{g}(\boldsymbol{\alpha}_{\nu-1})$ を定義したとき，ステップ幅は

$$t_\nu=\frac{\boldsymbol{s}_\nu^T\boldsymbol{s}_\nu}{\boldsymbol{s}_\nu^T\boldsymbol{y}_\nu} \tag{22}$$

で与えられる．以上より，BB 法を利用した射影勾配法のアルゴリズムは次の通りである．

射影勾配 BB 法

step 1: 初期点 $\boldsymbol{\alpha}_0\in\mathbb{R}^n$ を選び，$\nu=0$ とおく．

step 2: $\hat{\boldsymbol{\alpha}}_{\nu+1}=\boldsymbol{\alpha}_\nu+\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)$ とおいて step 5 へいく．

step 3: $\boldsymbol{s}_\nu=\boldsymbol{\alpha}_\nu-\boldsymbol{\alpha}_{\nu-1},\ \boldsymbol{y}_\nu=\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)-\nabla\tilde{g}(\boldsymbol{\alpha}_{\nu-1})$ とおく．

step 4: $\hat{\boldsymbol{\alpha}}_{\nu+1}=\boldsymbol{\alpha}_\nu+\dfrac{\boldsymbol{s}_\nu^T\boldsymbol{s}_\nu}{\boldsymbol{s}_\nu^T\boldsymbol{y}_\nu}\nabla\tilde{g}(\boldsymbol{\alpha}_\nu)$ とおく．

step 5: $\boldsymbol{\alpha}_{\nu+1}=p_A(\hat{\boldsymbol{\alpha}}_{\nu+1})$ とおく．ただし，$p_A(\cdot)$ は実行可能領域 $\mathcal{S}_D$ への正射影である．

step 6: もし停止条件が満たされていれば停止し，さもなければ $\nu=\nu+1$ として step 3 へいく．

BB 法は無制約狭義凸 2 次関数最小化問題に対して大域的収束することが知られており ([15])，また一般の無制約最小化問題に対して適当な非単調直線探索法と組み合わせれば，大域的収束することが知られている ([18])．Luss らの提案するステップ幅 (21) を用いた射影勾配法と，(22) を用いた射影勾配 BB 法の数値実験比較は 5 節で与える．

4. 新しいモデルの提案

不定値カーネルが与えられた場合の最適化モデルとして，本節では新しいモデルを提案する．

3.3 節で与えた Luss and d'Aspremont の手法は各反復で固有値計算が必要となるため，学習データが多くなると問題のサイズが増大し固有値計算がボトルネックとなる．そのため本研究では各反復で固有値計算を必要としないモデルを提案する．

与えられた不定値カーネル K_0 を半正定値錐 $\boldsymbol{S}^n_+$ に射影し，その近傍で半正定値行列 K を見つけることを目標とする．すなわちカーネル最適化問題 (13) において

$$\mathcal{K} = \{K \in \boldsymbol{S}^n \mid K \succeq O,\ \|K - (K_0)_+\|^2 \leq \beta^2\} \tag{23}$$

と定義し，以下の問題を考える:

$$\min_{K\in\mathcal{K}} \max_{\boldsymbol{\alpha}\in\mathcal{S}_D} \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} \tag{24}$$

パラメータ $\beta > 0$ は $(K_0)_+$ との距離を制御するものである．問題 (15) と異なり，β の値によらず実行可能である．定理 3.2 より，問題 (24) は max と min を入れ替えることができて，次のようになる．

$$\max_{\boldsymbol{\alpha}\in\mathcal{S}_D} \min_{K\in\mathcal{K}} \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} \tag{25}$$

このとき問題 (25) に関して，以下の定理が得られる．

定理 4.1. 不定値行列 $K_0 \in \boldsymbol{S}^n$，ラベル行列 $Y = \mathrm{diag}\,(y_1, \ldots, y_n)$ が与えられているとする．このとき問題 (25) は変数 $\boldsymbol{\alpha}$ だけの最適化問題

$$\max_{\boldsymbol{\alpha}\in\mathcal{S}_D} \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y\big((K_0)_+ + \beta I\big)Y\boldsymbol{\alpha} \tag{26}$$

に帰着される．

証明 $\boldsymbol{\alpha}$ が与えられたとき，関数 $w(K;\boldsymbol{\alpha})$ を

$$w(K;\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha}$$

と定義し，最小化問題

$$\min_{K\in\mathcal{K}} w(K;\boldsymbol{\alpha}) \tag{27}$$

を解く．

(i) $\boldsymbol{\alpha} = \boldsymbol{0}$ のとき:

$$\min_{K\in\mathcal{K}} w(K;\boldsymbol{0}) = \min_{K\in\mathcal{K}} 0 = 0$$

となるので，任意の $K \in \mathcal{K}$ が最適解になる．

(ii) $\boldsymbol{\alpha} \neq \boldsymbol{0}$ のとき:

問題 (27) の実行可能領域 (23) において $K \succeq O$ を緩和した問題

$$\min_{\|K-(K_0)_+\|^2\leq\beta^2,\ K\in\boldsymbol{S}^n} w(K;\boldsymbol{\alpha})$$

を考える．この問題のラグランジュ関数 $\mathcal{L}(K,\gamma)$ は

$$\mathcal{L}(K,\gamma) := \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} + \gamma(\|K - (K_0)_+\|^2 - \beta^2)$$

で定義されるので，KKT 条件は次式で与えられる．

$$\nabla_K \mathcal{L}(K,\gamma) = -\frac{1}{2}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T + 2\gamma(K-(K_0)_+) = O, \tag{28a}$$

$$\nabla_\gamma \mathcal{L}(K,\gamma) = \|K-(K_0)_+\|^2 - \beta^2 \leq 0, \tag{28b}$$

$$\gamma(\|K-(K_0)_+\|^2 - \beta^2) = 0, \tag{28c}$$

$$\gamma \geq 0. \tag{28d}$$

もし KKT 条件を満たす点で $\gamma = 0$ が成り立つならば，(28a) より $-\frac{1}{2}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T = O$ となり，$\boldsymbol{\alpha} \neq \mathbf{0}$ に矛盾する．よって (28d) より $\gamma > 0$ が成り立つので，(28a) を変形し

$$K = (K_0)_+ + \frac{1}{4\gamma}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T \tag{29}$$

が得られる．ここで相補性条件 (28c) より $\|K-(K_0)_+\|^2 - \beta^2 = 0$ が成り立つので，この式に (29) を代入すれば

$$\gamma = \frac{1}{4\beta}\|Y\boldsymbol{\alpha}\|_2^2 \tag{30}$$

となる．(30) を (29) に代入することで

$$K^* = (K_0)_+ + \frac{\beta}{\|Y\boldsymbol{\alpha}\|^2}(Y\boldsymbol{\alpha})(Y\boldsymbol{\alpha})^T \tag{31}$$

が得られ，これは $K^* \succeq O$ を満たすことから，問題 (27) の最適解である．よって，問題 (27) に (31) を組み込めば，

$$\min_{K\in\mathcal{K}} w(K;\boldsymbol{\alpha}) = w(K^*;\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T Y((K_0)_+ + \beta I)Y\boldsymbol{\alpha}$$

を得る．$\min_{K\in\mathcal{K}} w(K;\boldsymbol{\alpha})$ を $\boldsymbol{\alpha}$ の関数として見れば，これは $\boldsymbol{\alpha} = \mathbf{0}$ で連続である．

以上より，問題 (25) は $\boldsymbol{\alpha} = \mathbf{0}, \boldsymbol{\alpha} \neq \mathbf{0}$ のいずれの場合でも，問題 (26) で表すことができる． □

変数 $\boldsymbol{\alpha}$ の凸 2 次計画問題 (26) は Sequential Minimal Optimization(SMO) [16] で解くことができる．SMO は 2 つの変数を選び他の変数は定数とみなして，2 次計画問題を繰り返し解く手法である．行列演算が不要になり，高速に SVM を解くことができるので，Luss and d'Aspremont のモデルよりも高速に最適解を求めることができる．なお凸 2 次計画問題 (26) は，$\beta = 0$ の場合には 3.1 節で紹介した Denoise になることに注意されたい．

5. 数値実験

本節では 2 種類の数値実験を行う．まず，5.1 節では，3.3 節で紹介した Luss and d'Aspremont のモデルを彼らの提案した射影勾配法と本研究で提案した射影勾配 BB 法で解いて，収束の速さを数値実験により比較する．次に 5.2 節では，4 節で定式化した新しいモデルの識別能力を 3 節で紹介した手法と比較する．

実験環境はいずれも，CPU: Intel® Core™2 Duo P8400(2.26GHz)，メモリ: 2.99GB，OS: Windows XP Professional SP3 であり，言語は MATLAB である．SVM の計算には LIBSVM [3] を使用した．

使用したデータは，USPS handwritten digit data [17]，MNIST database of handwritten digits [11]，及び UCI Machine Learning Repository [4] の Ionosphere，German である．USPS handwritten digit data と MNIST database of handwritten digits は手書きの数字の画像データであり，画素数はそれぞれ 16×16pixels，28×28pixels である．0 から 9 までの数字の画像データから 2 種類の数字を選び，シンプソンカーネル (11) を適用して実験に用いた．以下の表記として，例えば USPS 3-5 は USPS handwritten digit data のうち数字 3 のデータと数字 5 のデータを扱うことを意味する．Ionosphere，German にはシグモイドカーネル (10) を適用し，実験に用いた．各データのトレーニングデータ数 (#Train) 及びテストデータ数 (#Test)，カーネル行列の最大固有値 $\lambda_{\max}$ 及び最小固有値 $\lambda_{\min}$ は表 1 の通りである．

表 1: トレーニングデータ数，テストデータ数，最大固有値と最小固有値

Dataset	#Train	#Test	$\lambda_{\min}$	$\lambda_{\max}$
USPS 4-6	737	767	-31.64	3668.51
USPS 3-5	857	829	-33.15	3531.27
MNIST 4-6	1895	1902	-33.58	1479.82
MNIST 3-5	1994	1940	-33.28	1329.16
Ionosphere	234	117	-0.35	205.90
German	666	334	-0.39	130.25

5.1. Luss らの射影勾配法と射影勾配 BB 法の比較

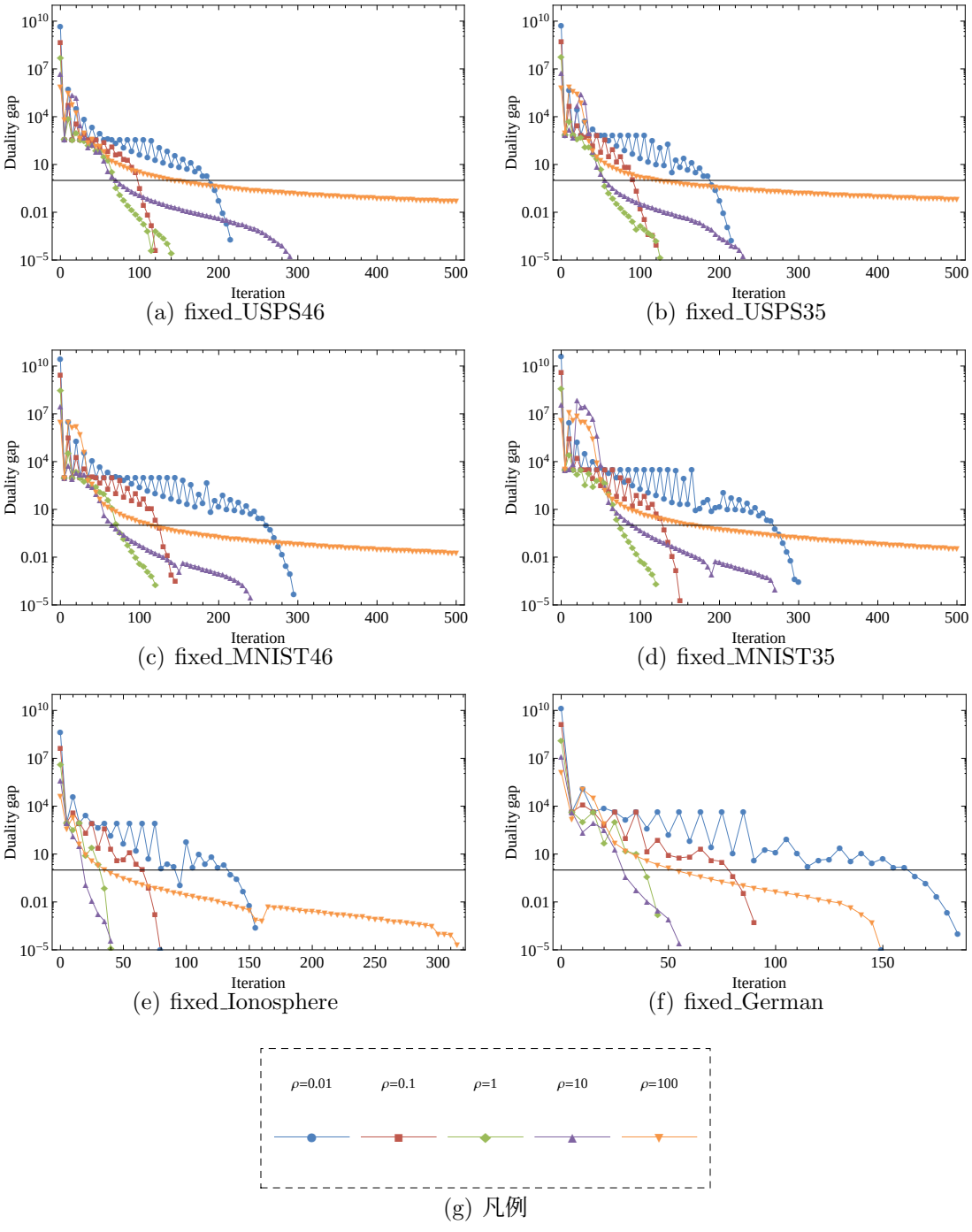
Luss and d'Aspremont の定式化したモデル (19) を，彼らの提案するステップ幅 (21) を用いた射影勾配法と，ステップ幅 (22) を用いた射影勾配 BB 法とでそれぞれ解き，収束するまでの反復回数と計算時間を比較する．実験のための Matlab コードは Luss らの用いたものを基にしている ([13])．また，収束判定条件は双対ギャップ ([14]) が許容誤差以下になった場合とし，最大反復回数は 500 回とした．実験におけるパラメータの値は，$C = 10$，許容誤差を 10^{-5} とし，$\rho = 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$ のそれぞれの値で実験を行った．また，正則化のための関数 (20) とステップ幅 (21) における ε と δ の選び方は [13] に倣って $\varepsilon = 10^{-4}, \delta = 5$ とした．

表 2 は，以上の実験の結果をまとめたものである．この表によれば，射影勾配 BB 法を利用して解くことにより反復回数及び計算時間の大幅な減少がみられる．図 1，図 2 は双対ギャップの減少の様子を片対数でプロットしたグラフである．

射影勾配 BB 法は非単調なアルゴリズムであるが，Ionosphere，German では反復の初期段階において双対ギャップ振動が特に大きく見られた．そのため双対ギャップの単調減少を保証するために，Armijo の直線探索法を利用した射影勾配 BB 法で実験を行った．結果は表 3 の通りであり，双対ギャップの減少の様子は図 3 で与えた．双対ギャップは単調に減少し，データまたはパラメータ ρ の値によっては反復回数も少なくなる．しかしほとんどの場合，全体の計算時間は多くなっている．これは，直線探索をする度に目的関数の計算，すなわち固有値を計算する必要があり，1 反復ごとの計算時間が長くなってしまうためである．したがってサイズの大きいデータに対しては，特に直線探索は有効であるとはいえない．

表 2: Luss and d'Aspremont の方法，射影勾配 BB 法で問題 (19) を解いたときの反復回数と計算時間

		Luss & d'Aspremont の方法		射影勾配 BB 法	
Data set	ρ	反復回数	実行時間 (秒)	反復回数	実行時間 (秒)
USPS 4-6	0.01	219	241.52	18	26.79
	0.1	121	131.44	20	30.84
	1	143	257.72	24	37.62
	10	292	361.92	28	35.45
	100	500	550.77	44	80.73
USPS 3-5	0.01	219	206.65	20	20.88
	0.1	122	170.00	23	42.29
	1	126	326.42	27	78.08
	10	233	285.92	30	46.72
	100	500	595.09	43	105.96
MNIST 4-6	0.01	297	3659.25	21	308.92
	0.1	148	1922.37	21	339.15
	1	123	1807.30	25	376.43
	10	242	4503.40	32	541.05
	100	500	6560.26	44	691.34
MNIST 3-5	0.01	303	3263.67	29	365.81
	0.1	150	1672.51	22	289.76
	1	122	1722.16	22	332.09
	10	273	5951.90	34	643.28
	100	500	6481.47	52	1111.32
Ionosphere	0.01	159	40.24	82	20.78
	0.1	80	21.19	37	9.74
	1	40	10.74	23	6.38
	10	42	19.35	16	4.96
	100	319	300.06	17	5.63
German	0.01	189	513.68	107	294.62
	0.1	95	259.42	46	130.22
	1	48	135.62	24	75.14
	10	56	185.19	18	57.89
	100	150	454.30	24	84.84



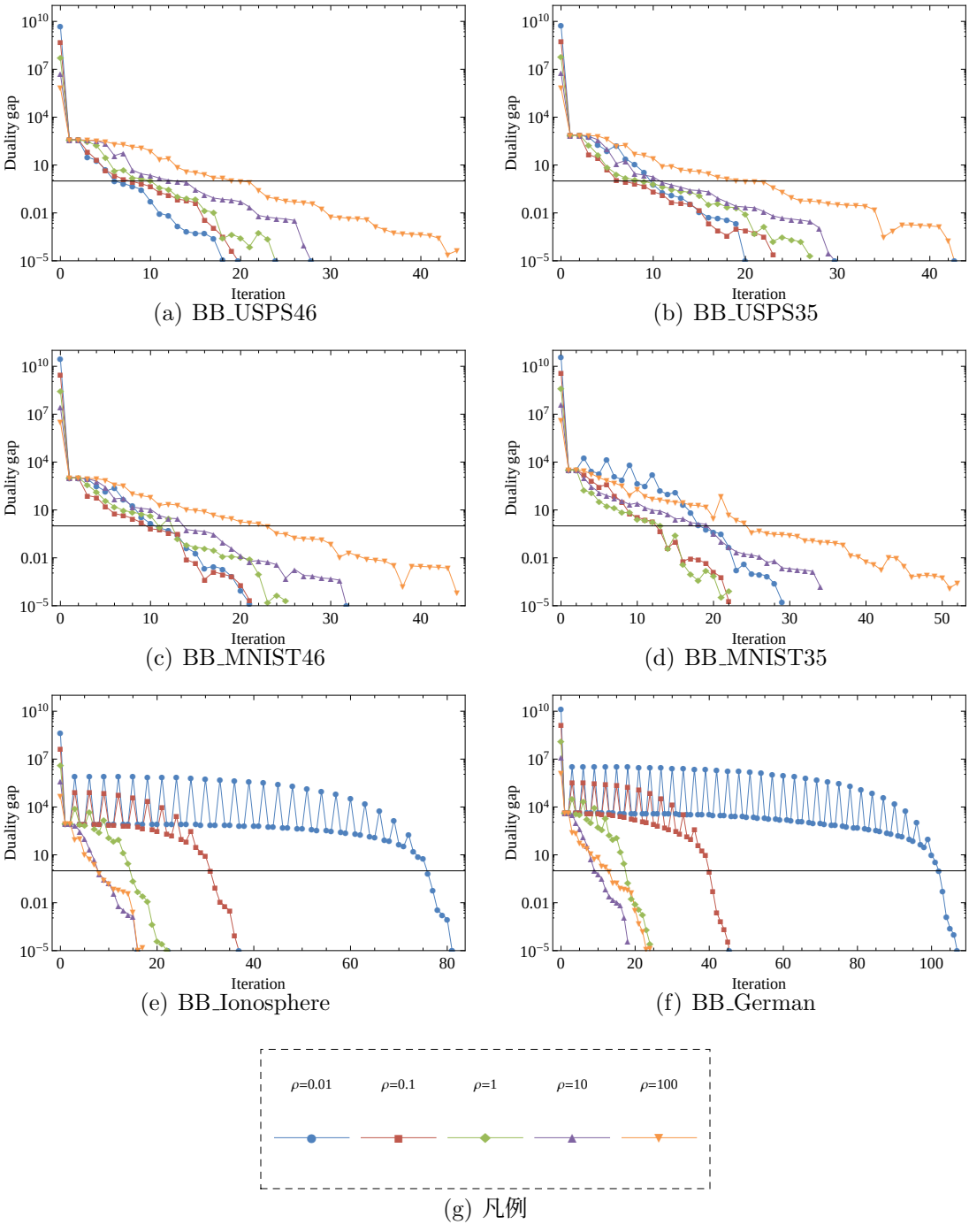
(a) fixed_USPS46 (b) fixed_USPS35

(c) fixed_MNIST46 (d) fixed_MNIST35

(e) fixed_Ionosphere (f) fixed_German

(g) 凡例

図 1: Luss and d'Aspremont の方法で解いたときの反復回数と双対ギャップ



(a) BB_USPS46 (b) BB_USPS35

(c) BB_MNIST46 (d) BB_MNIST35

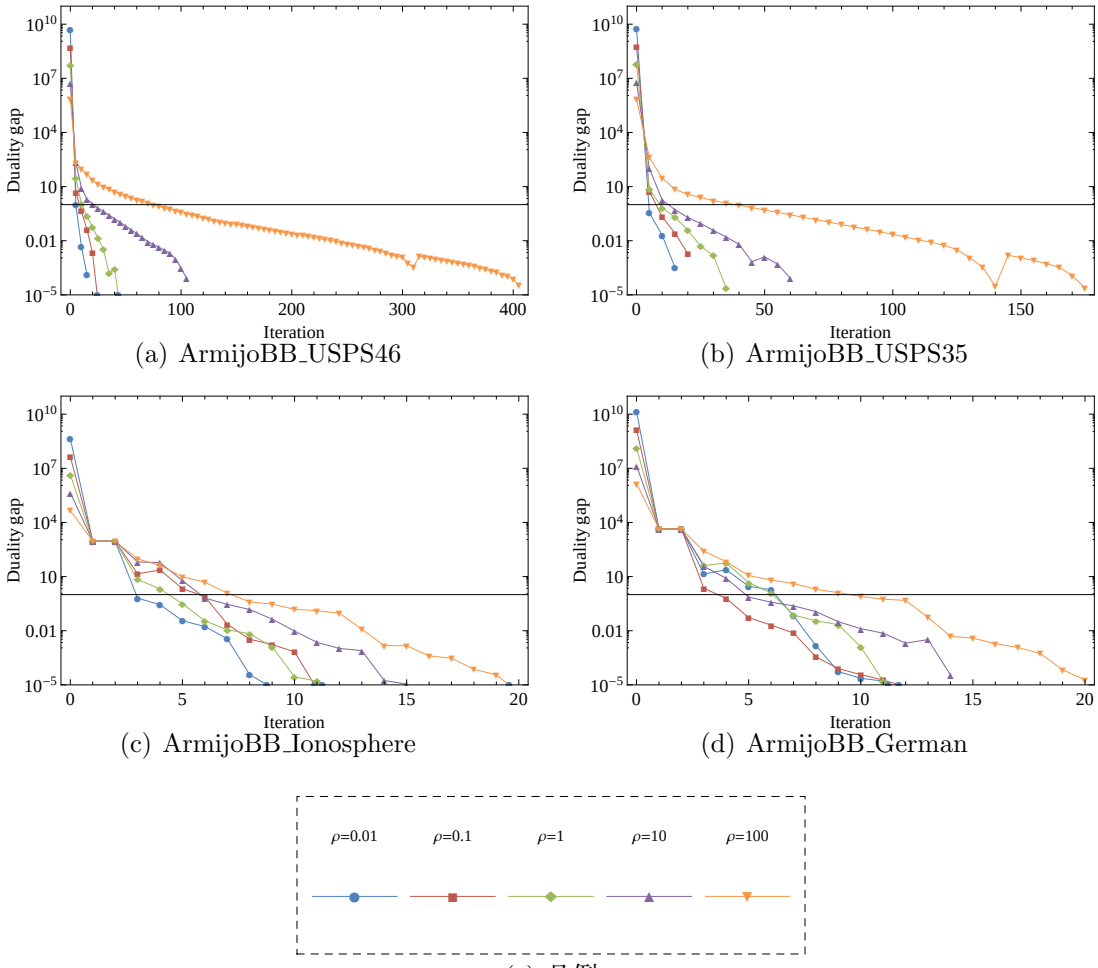
(e) BB_Ionosphere (f) BB_German

(g) 凡例

図 2: 射影勾配 BB 法で解いたときの反復回数と双対ギャップ

表 3: 射影勾配 BB 法，直線探索付き射影勾配 BB 法で問題 (19) を解いたときの反復回数と計算時間

		射影勾配 BB 法		射影勾配 BB 法 + 直線探索	
Data set	ρ	反復回数	実行時間 (秒)	反復回数	実行時間 (秒)
USPS 4-6	0.01	18	26.79	17	48.26
	0.1	20	30.84	26	55.94
	1	24	37.62	45	142.49
	10	28	35.45	109	365.87
	100	44	80.73	409	2017.54
USPS 3-5	0.01	20	20.88	19	40.34
	0.1	23	42.29	23	51.13
	1	27	78.08	36	99.04
	10	30	46.72	63	219.71
	100	43	105.96	178	845.26
Ionosphere	0.01	82	20.78	9	8.58
	0.1	37	9.74	11	8.54
	1	23	6.38	12	8.53
	10	16	4.96	15	9.88
	100	17	5.63	20	17.68
German	0.01	107	294.62	12	101.46
	0.1	46	130.22	12	101.93
	1	24	75.14	11	88.36
	10	18	57.89	14	111.45
	100	24	84.84	20	164.52



(a) ArmijoBB_USPS46 (b) ArmijoBB_USPS35

(c) ArmijoBB_Ionosphere (d) ArmijoBB_German

(e) 凡例

図 3: 直線探索付き射影勾配 BB 法で解いたときの反復回数と双対ギャップ

5.2. モデルの識別能力の比較

本研究で提案したモデル (26) の識別能力を，3 節で紹介した各種の既存手法と比較する．モデルに含まれるパラメータ C, β, ρ は，$\{2^{-3}, 1, 2^3, 2^6, 2^9, 2^{12}\}$ の候補の中からトレーニングデータに対する 5-fold 交差検定により決定した．選ばれたパラメータの値は表 4 の通りである．表の中で「-」はパラメータが該当しない箇所を表す.

表 4: 既存手法と提案手法のパラメータの値

Dataset	パラメータ	IndSVM	Denoise	Flip	Shift	ModSVM	LussSVM	提案モデル
USPS 3-5	C	0.125	8	8	0.125	0.125	8	8
	ρ	-	-	-	-	-	4096	-
	β	-	-	-	-	-	-	0.125
USPS 4-6	C	0.125	8	8	0.125	1	1	8
	ρ	-	-	-	-	-	512	-
	β	-	-	-	-	-	-	0.125
MNIST 3-5	C	1	8	8	0.125	64	1	64
	ρ	-	-	-	-	-	4096	-
	β	-	-	-	-	-	-	0.125
MNIST 4-6	C	0.125	64	8	0.125	64	1	8
	ρ	-	-	-	-	-	4096	-
	β	-	-	-	-	-	-	0.125
Ionosphere	C	8	64	64	8	64	8	64
	ρ	-	-	-	-	-	4096	-
	β	-	-	-	-	-	-	0.125
German	C	0.125	0.125	0.125	8	8	1	1
	ρ	-	-	-	-	-	64	-
	β	-	-	-	-	-	-	1

識別能力の比較の指標として，Accuracy(正確度) と Recall(再現率)，及びその平均値 (Average = (Accuracy + Recall)/2) を用いる．Accuracy は，すべてのテストデータに対して，正しく分類されたデータの割合である．Recall(再現率) は，+1 に分類されるデータに対して正しく分類されたデータの割合であり，完全性の指標である．それぞれ次のように定義され，値が大きいほど良い.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

ただし，TP，FP，FN，TN は表 5 で定義されるようなデータ数である．例えば FP は，正しいクラスは「-1」であるがテストで得られたクラスが「$+1$」となったデータの数である.

表 5: TP，FP，FN，TN の定義

		正しいクラス +1	正しいクラス −1
テストで得られたクラス	+1	TP	FP
	−1	FN	TN

表 6 が実験結果である．表において IndSVM は与えられた不定値カーネルを直接 SVM に利用した手法である．Denoise，Flip，Shift は 3.1 節，ModSVM は 3.2 節，LussSVM は 3.3 節でそれぞれ紹介した手法である．太字となっている値は，すべての手法の結果の中で最も良い値を表している．

表 6: 既存手法と提案手法の識別能力の比較

Dataset	Measure	IndSVM	Denoise	Flip	Shift	ModSVM	LussSVM	提案モデル
USPS 3-5	Accuracy	67.14	95.47	95.08	91.07	93.66	95.08	**95.99**
	Recall	63.88	94.98	94.98	93.30	93.06	**97.61**	95.93
	Average	65.51	95.22	95.03	92.19	93.36	**96.35**	95.96
USPS 4-6	Accuracy	77.25	98.37	98.48	90.43	97.90	94.63	**98.60**
	Recall	73.81	99.10	99.10	88.04	98.42	**100.00**	99.32
	Average	75.53	98.73	98.79	89.23	98.16	97.32	**98.96**
MNIST 3-5	Accuracy	63.67	94.74	**95.53**	75.45	94.53	87.80	94.79
	Recall	68.02	94.95	**96.14**	87.62	95.05	77.62	95.15
	Average	65.84	94.85	**95.83**	81.54	94.79	82.71	94.97
MNIST 4-6	Accuracy	67.94	98.61	**98.81**	87.63	98.66	93.81	98.56
	Recall	68.64	98.88	98.78	86.25	98.78	**100.00**	98.68
	Average	68.29	98.74	**98.80**	86.94	98.72	96.91	98.62
Ionosphere	Accuracy	94.02	76.07	83.76	93.16	68.38	**94.87**	**94.87**
	Recall	**97.20**	74.77	83.18	95.33	66.36	**97.20**	96.26
	Average	95.61	75.42	83.47	94.24	67.37	**96.03**	95.57
German	Accuracy	74.00	74.00	74.00	73.40	71.00	**78.44**	72.20
	Recall	92.56	92.56	91.96	91.37	69.94	92.86	**94.35**
	Average	83.28	83.28	82.98	82.38	70.47	**85.65**	83.27

IndSVM で得られた解は定常点でしかなく，最適解である保証は無い．数値実験においても良い結果はほとんど得られていない．Denoise，Flip，Shift は，例えば MNIST3-5 に対する Flip のように他の手法に比べ良い結果を得られることもあるが，そのばらつきは比較的大きい．ModSVM も同様の傾向がみられる．これらに比べて，Luss らの提案するモデル LussSVM と今回提案したモデルは，全体として良い結果が得られていることが確認できる．特にシンプソンカーネルを用いた 4 つの画像データ集合すべてに対して，LussSVM より提案モデルの方が Accuracy が高く，Average も概ね高い．シグモイドカーネルを用いた 2 つのデータ

集合 Ionosphere，German に対しては，Luss らのモデルにやや劣っているが，その一方で提案手法は固有値計算の必要が無いという利点をもつ．以上より，計算量と識別能力を共に考慮すれば，今回提案した手法は Luss らの手法に対する有効な修正であるとみなせ，更に他の手法と比較しても遜色がない．

6. おわりに

本論文では，不定値カーネル行列を伴うサポートベクターマシンに着目し，いくつかの既存手法を紹介した．そして，Luss and d'Aspremondt のモデルを解くための射影勾配法に対する Barzilai-Borwein 法の適用を提案し，数値実験により収束の加速を確認した．更に，Luss らのモデルは計算量の点で問題があることから，その修正として新たなモデルを定式化するとともに，数値実験でその識別能力を評価した．今回提案したモデルによる識別能力の顕著な改善はみられなかったものの，Luss らのモデルに比べ反復回数及び計算時間は大幅に減少した．また他の手法に比べて誤判別率の平均は比較的低く，また分散が小さくなる傾向もみられた．今後は，他の不定値カーネル及びデータ集合を用いて，提案手法の有効性をより詳しく検証する必要がある．また行列の類似性尺度の一つであるアラインメントを利用して，与えられた不定値行列の近傍で有効な半正定値行列を見つける方法に新たな情報を組み込むことも考えられる．

謝辞

多くの有意義なコメントを頂いたレフェリーの方々に感謝します．なお，本研究の一部は独立行政法人日本学術振興会の科学研究費補助金基盤研究（C）(課題番号 21510164) の支援のもとに行われた．

参考文献

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman: Basic local alignment search tool. *Journal of Molecular Biology*, **215** (1990), 403–410.

[2] J. Barzilai and J.M. Borwein: Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, **8** (1988), 141–148.

[3] C.C. Chang and C.J. Lin: LIBSVM: A Library for Support Vector Machines. (`http://www.csie.ntu.edu.tw/~cjlin/libsvm`).

[4] A. Frank and A. Asuncion: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, (2010), (`http://archive.ics.uci.edu/ml`).

[5] J.B.G. Frenk, G. Kassay, and J. Kolumbán: On equivalent results in minimax theory. *European Journal of Operational Research*, **157** (2004), 46–58.

[6] 福水健次: カーネル法入門 —正定値カーネルによるデータ解析— (朝倉書店, 2010).

[7] B. Haasdonk: Feature space interpretation of SVMs with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27** (2005), 482–492.

[8] B. Haasdonk and D. Keysers: Tangent distance kernels for support vector machines. *Proceedings of the 16th International Conference on Pattern Recognition*, **2** (2002), 864–868.

[9] G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan: Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, **5** (2004), 27–72.

[10] J. Laub and K.R. Müller: Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, **5** (2004), 801–818.

[11] Y. LeCun and C. Cortes: The MNIST Database of Handwritten Digits. (`http://yann.lecun.com/exdb/mnist/`) .

[12] H.T. Lin and C.J. Lin: A study on sigmoid kernels for SVM and the training of non-psd kernels by SMO-type methods. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[13] R. Luss: homepage, (`http://www.tau.ac.il/~rluss/`) または (`http://www.eecs.berkeley.edu/~rluss/`).

[14] R. Luss and A. d'Aspremont: Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, **1** (2009), 97–118.

[15] Y. Narushima, T. Wakamatsu, and H. Yabe: Extended Barzilai-Borwein method for unconstrained minimization problems. *Pacific Journal of Optimization*, **6** (2010), 591–613.

[16] J.C. Platt: Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola (eds.): *Advances in Kernel Methods-Support Vector Learning* (MIT Press, 1998), 185–208.

[17] C.E. Rasmussen and C.K.I. Williams: Gaussian Processes for Machine Learning. (`http://www.gaussianprocess.org/gpml/`) .

[18] M. Raydan: The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization*, **7** (1997), 26–33.

[19] V. Vapnik: *Statistical Learning Theory* (John Wiley & Sons, New York, 1998).

[20] G. Wu, E.Y. Chang, and Z. Zhang: *An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines*. Technical Report, Department of Electrical and Computer Engineering, University of California, Santa Barbara, June 2005.

矢部博
東京理科大学理学部数理情報科学科
〒162-8601 東京都新宿区神楽坂 1-3
E-mail: `yabe@rs.kagu.tus.ac.jp`

ABSTRACT

A METHOD FOR SUPPORT VECTOR MACHINE CLASSIFICATION WITH INDEFINITE KERNELS

Saeko Kimura
Central Japan Railway Company

Hiroshi Yabe
Tokyo University of Science

Support vector machines (SVMs) have been paid attention to for solving binary classification problems. SVMs usually use a positive definite kernels in many applications. On the other hand, SVMs with indefinite kernels are studied in this decade, because such SVMs take advantage of application-specific structure in data. Recently Luss and d'Aspremont (2009) formulated a convex optimization problem to deal with them. Their formula came from a max-min problem with a penalized term which controled the distance between the original indefinite kernel matrix and the proxy positive semidefinite kernel matrix. They gave a projected gradient method to solve the problem. However their method needs to calculate eigenvalues and vectors of a matrix corresponding to a given indefinite kernel matrix. In this paper, we first introduce the Barzilai and Borwein method instead of the gradient method of Luss and d'Aspremont to accelerate the method in practical computation. Secondly, we propose a new formulation of SVMs with indefinite kernels to overcome the defect that the model of Luss and d'Aspremont needs eigenvalues and vectors of a matrix. Since our formula is represented by a quadratic optimization problem, it can be easily solved by a suitable numerical method like the SMO method. Finally we give some numerical experiments to investigate numerical performance of our method and the generalization performance of our formulation.