サポートベクターマシンとパターン認識手法による統計的判別

M2007MM033 山田俊哉
指導教員：田中豊

1 はじめに

2 クラス分類問題を解く学習機械として Vladmir N Vapnik により提唱されたサポートベクターマシン (Support Vecter Machines. SVM) は, カーネル法を導入したことにより, 非線形への拡張がなされ脚光を浴びることとなった [1]. この学習機械は既存の判別方法とは異なり, 複雑かつ大規模問題に対してもスムーズに対応できる側面を持っており幅広い問題に応用ができる. 本研究では線形判別関数, ロジスティック回帰モデルと SVM との比較および, カーネル法の適応の有無における SVM の性能比較, ソフトマージン SVM におけるノルムの取り方による正判別率の変化など観察するための実験を行う.

2 サポートベクターマシン

SVM は 2 クラスの分類問題を解くために作られた学習機械 (学習アルゴリズム) である. SVM が 2 クラス識別器として優れているモデルである理由に, クラス分類を行う超平面の決定の基準に「マージン最大化」と言う明確な基準が設けられている点と, カーネル学習法により非線形の判別問題へ拡張することができる, と言う 2 点が挙げられる. ここで「マージン」とは識別面と学習データベクトルとのユークリッド距離である. カーネル学習法は後述する. SVM の最も単純なモデルでありながらも, マージン最大化など SVM の主要な要素を備えた「線形分離可能な学習データに対する SVM」を初めに説明する.

2.1 線形分離可能な学習データに関する SVM

学習データの集合が $(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \cdots, (\mathbf{x}_n, t_n)$ $\forall i, \mathbf{x}_i \in R^d, t_i \in \{-1, 1\}$ と与えられたとする. ここで $\mathbf{x}_i = (x_1, x_2, \cdots, x_d)^T$ は個体の特徴ベクトル, t_i はクラスラベルである. 入力に対し SVM は識別関数

$$f(x) = \mathrm{sign}\big(g(\mathbf{x})\big) \tag{1}$$

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} - b \tag{2}$$

により,2 値の出力値を計算する. ここで,$\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$ は線形識別器の重みベクトルと呼ばれるパラメータである. また b はバイアス項と呼ばれるパラメータであり, この $\mathbf{w}$ と b により識別面 $g(x)$ を決定している. sign(y) 関数は $y > 0$ のとき 1 をとり $y \leq 0$ のとき -1 をとる符号関数である. 学習データが線形分離可能であるため, $t_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 \ \ i = 1, \cdots, n$ を満たすような $\mathbf{w}$ と b が存在する. つまり $H1 : \mathbf{w}^T\mathbf{x} - b = -1$ と $H2 : \mathbf{w}^T\mathbf{x} - b = 1$ の 2 枚の超平面で学習データが完全に分離されており, この間にプロットされる学習データは存在しないことを意味する. この 2 枚の超平面 $H1$ と $H2$ 上の学習データをそれぞれ,$\mathbf{x}^-$ と $\mathbf{x}^+$ とすると

$$\mathbf{w}^T\mathbf{x}^{\pm} - b = \pm 1 \tag{3}$$

となり, このマージン γ は以下のようになる

$$\gamma = \frac{1}{2}\left(\frac{\mathbf{w}^T\mathbf{x}^+}{\|\mathbf{w}\|} - \frac{\mathbf{w}^T\mathbf{x}^-}{\|\mathbf{w}\|}\right) = \frac{1}{\|\mathbf{w}\|} \tag{4}$$

線形分離可能な場合, マージンは必ず $\frac{1}{\|\mathbf{w}\|}$ になることが解る.

2.2 モデルの定式化

線形分離可能な場合 SVM は以下のような最適化問題になる.

$$\min_{w} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{5}$$

$$\text{制約条件:} \qquad t_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 \ \ i = 1, \cdots, n \tag{6}$$

これは数理計画法の分野では凸 2 次計画問題として知られている問題であり, 様々な計算方法が提唱されている, 今回は双対問題に帰着し単純な勾配法を用いて解く方法を採用する. まず,Lagrange 乗数 $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_n)$ を導入すると, 式 (5) の目的関数は以下のようになる

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\lambda_i\left\{t_i(\mathbf{w}^T\mathbf{x}_i - b) - 1\right\} \tag{7}$$

最適解においては,L の勾配が 0 になるので,$\frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \mathbf{w}} = 0$ となり, そこから得られた条件を式 (7) に代入すると次の双対問題が得られる.

$$\max_{\lambda} \quad L_D(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j \mathbf{x}_i^T\mathbf{x}_j \tag{8}$$

$$\text{制約条件:} \qquad \sum_{i=1}^{n}\lambda_i t_i = 0, \ \lambda_i \geq 0, \ i = 1, \cdots, n \tag{9}$$

となり, 双対定理より目的関数から,$\mathbf{w}$,b が消え,$\boldsymbol{\lambda}$ のみに関する最大化問題になる.

2.3 サポートベクター

前節で求めた最適な λ_i を λ_i* と表す.$\lambda_i* = 0$ に対応する学習データ $\mathbf{x}_i$ はパラメータ $\mathbf{w}$ の決定に関与していない. つまり, 全ての学習データの中で $\lambda_i* > 0$ となる一部の $\mathbf{x}_i$ を「Support Vector」と呼び SVM の名前の由来もなっている. また各クラスに属するサポートベクターを $\mathbf{x}_s^+$,$\mathbf{x}_s^-$ とおくと最適な b は

$$b = -\frac{\min_{t_i=1}(\mathbf{w}^T\mathbf{x}_s^+) + \max_{t_i=-1}(\mathbf{w}^T\mathbf{x}_s^-)}{2} \tag{10}$$

より求められる.

2.4 線形分離不可能なデータへの拡張

線形分離可能なデータは現実の問題では実用に向かず, 実際は非線形で複雑な識別面を持つ場合が多い. これに対応する方法は識別面より他群への進入を許す「ソフトマージン法」とカーネル関数を用いて高次空間へ非線形写像する「カーネル法」を導入する方法があり, その二つを同時に導入した「カーネルソフトマージン法」を用いる方法もある.

2.4.1 ソフトマージン法

ソフトマージン法では, マージン $1/\|\mathbf{w}\|$ を最大化しながら, 識別面の反対側に入る事を許す. 反対側にどれくらい入り込んだかの距離を,$\xi_i(\geq 0)$ を用いて,$\xi_i/\|\mathbf{w}\|$ とあらわす. 反対側に入り込んだ距離の和は小さいことが望ましいため最適な識別面は下のような最適化問題になる.

$$\min_{w,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i^2 \tag{11}$$

$$\text{制約条件:} \quad t_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 - \xi_i \; i = 1, \cdots, n \tag{12}$$

ここでパラメータ C はマージンの大きさとはみ出しの距離に対するペナルティ項とのバランスを調整する重みのパラメータである. 反対側に入り込んだ距離 ξ_i を上記のように L2 ノルム ($\|\xi\|_2 = \sum_i \xi_i^2$) とする場合と L1 ノルム ($\|\xi\|_1$) を用いる場合がある,L2 ノルムの場合以下のような最適化問題が得られる.

$$\max_{\boldsymbol{\lambda}} \quad L_D(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j \left(\mathbf{x}_i^T\mathbf{x}_j + \frac{1}{C}\delta_{ij}\right), \tag{13}$$

$$\text{制約条件:} \sum_{i=1}^{n}\lambda_i t_i = 0, \quad \lambda_i \geq 0, i = 1, \cdots, n, \tag{14}$$

また ξ_i に L1 ノルム ($\|\xi\|_1$) を用いた場合は以下のようになる

$$\max_{\lambda} \quad L_D(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j \mathbf{x}_i^T\mathbf{x}_j \tag{15}$$

$$\text{制約条件:} \sum_{i=1}^{n}\lambda_i t_i = 0, 0 \leq \lambda_i \leq C, \; i = 1, \cdots, n \tag{16}$$

これらソフトマージンを用いた SVM を本研究では 1 ノルムソフトマージン SVM,2 ノルムソフトマージン SVM と呼び, これに対して前述の識別面からの進入を許さない SVM をハードマージン SVM と呼ぶ. なお以降 1 ノルムソフトマージン SVM を"L1-SVM",2 ノルムソフトマージン SVM を"L2-SVM", ハードマージン SVM を"H-SVM" と呼称する.

2.4.2 カーネル法

非線形で複雑な識別面に対応する方法として, 特徴ベクトルを高次元空間に非線形写像して識別する方法である. 元の特徴ベクトル $\mathbf{x_i}$ を非線形写像 $\phi(\mathbf{x_i})$ によって変換すると, 元々, 式 (8) は入力データの内積に依存しているため, 非線形に写像した $\phi(\mathbf{x_i})$ と $\phi(\mathbf{x_j})$ の内積が $<\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)>= K(\mathbf{x}_i, \mathbf{x}_j)$ のように元の特徴ベクトルからカーネルと呼ばれる $K(\mathbf{x}_i, \mathbf{x}_j)$ が計算できれば高次元空間で $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ を計算しなくても良い. このカーネルトリックを用いると目的関数 $L_D(\boldsymbol{\lambda})$ と識別関数 $f(x)$ は

$$\max_{\lambda} \; L_D(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{17}$$

$$f(x) = \text{sign}\left((\sum_{i=1}^{n}\lambda_i t_i K(\mathbf{x}_i, \mathbf{x}_j) - b)\right) \tag{18}$$

となり. カーネルトリックにより識別面のパラメータ b を求める. このとき $\mathbf{w}$ を直接もとめずに識別関数が計算が可能となる. またカーネル法を用いても前述の 1 ノルムソフトマージン,2 ノルムソフトマージンの考え方を適応することが可能となる.

3 カーネルロジスティック回帰分析

ロジスティック関数は判別手法として一般的に良く知られた方法である. 本研究では,SVM 以外の判別手法にもカーネル関数を適応する例として, このロジスティク回帰モデルにカーネルを適応することを考える. 第 2 節に述べたように学習データが $\{(\mathbf{x}_1, t_1),(\mathbf{x}_2, t_2),\cdots,(\mathbf{x}_n, t_n)\}, \mathbf{x}_i \in R^d$ として与えられた場合, ロジスティック回帰モデルは以下のように定義される.

$$\mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}^T\mathbf{w}, g(\mu_i) = log\frac{\pi_i}{1-\pi_i}, \pi_i = \frac{\mu_i}{n} \tag{19}$$

$$\mathbf{g}(\boldsymbol{\mu}) = (g(\mu_1), g(\mu_2), \ldots, g(\mu_n))^T \tag{20}$$

$$\mu_i = E(y_i) \tag{21}$$

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n), \mathbf{1} = (1, 1, \ldots, 1)^T, \tag{22}$$

ここで $\mathbf{w}$ は重みベクトルである. 入力空間上の $\mathbf{x}_i$ が $\phi(\mathbf{x_i})$ に非線形写像されるならば, ヒルベルト空間 H 上の Φ は $\Phi = (\mathbf{1}, \phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots,)$ となり, この空間 H において, カーネルロジスティック回帰モデルは以下のようにあらわされる.

$$\mathbf{g}(\boldsymbol{\mu}) = \Phi^T\mathbf{w} \tag{23}$$

ここで $\Phi \in H$ かつ $\mathbf{w}^* \in H$ である. $\mathbf{w} \in H$ は $\mathbf{w}^* =< \Phi, \boldsymbol{\alpha} > +\mathbf{u}, \mathbf{u} \in \text{span}(\Phi)^{\perp}$ となる. よって $< \Phi^T\mathbf{u} >= 0$ より,$\mathbf{u}$ は $\mathbf{g}(\boldsymbol{\mu})$ の説明に寄与しないため $\mathbf{w}^* \in \text{span}(\Phi)$ と考えることができる. そのためこのモデルは以下のように展開される.

$$\mathbf{y} = \Phi^T\Phi\boldsymbol{\alpha} + \epsilon \Rightarrow \mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \epsilon \tag{24}$$

ここで $\mathbf{K}$ は $\mathbf{K}_{ij} = K(\mathbf{x_i}, \mathbf{x_j}) =< \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) >$ となるカーネル行列である.

4 統計ソフト R への実装

後述の実験において, これまで述べた SVM の最適化問題を解く為の手法は最適化問題が凸二次計画である事を利用し iris での実験は最急降下法の一つである慣性法を用いている. また, 人工データの数値実験においては逐次最小最適化アルゴリズム (SMO algorithm: Sequential Minimal Optimisation algorithm) を採用した. これはそれぞれの SVM において双対変数である λ を更新するとき, 各ステップで更新すべき 2 点 (λ_a, λ_b) が選択されれば $\lambda_a^{new} + \lambda_b^{new} = \lambda_a^{old} + \lambda_b^{old} = C_{\text{定数}}$ が成り立つため 2 点において解析的に最適化問題を解くことができるという, 分解法のアルゴリズムをより極端にした SVM のためのアルゴリズムである. プログラミングに関しては R には"e1071"や"klab などのパッケージにより SVM が利用可能であるが, 同一の条件における L1,L2-SVM が実装されていないなどの理由により, 本研究ではこれらパッケージの SVM 関数は採用せず, 作成したプログラムによって比較検討を行った.

5　実データを用いた数値的検討

5.1　Fisher の iris データ

実験には Fisher のアイリスデータを使用した. データ中の線形分離不可能な 2 群 (versicolor,virginica) を用いた. データは 4 変数, 各群 50 例の合計 100 例で構成されている. 今回はそのデータの中から各群25 例ずつ 50 例をランダムに抽出し学習データとし, 残り 50 例のデータをテストデータとし性能の比較を行った. 実験では 20 組の学習データとテストデータの対を用いていた.

5.2　カーネル法を用いない場合

Fisher の線形判別関数 (LDF),L1-SVM,L2-SVM を用いた. パラメータ C は 1 ノルム,2 ノルムの場合でそれぞれ $C=1, C=2$ とした. それぞれ 20 組のデータセットで実験した時の正判別率の平均を表 (1) に示す.

表 1　カーネルを用いない場合の正判別率平均

	L1-SVM	L2-SVM	LDF
正判別率	0.936	0.905	0.908

ここで 20 組の正判別率について符号検定を行うと以下の結果が得られた.

表 2　正判別率の符号検定

	正判別率	p-value
L1-SVM : L2-SVM	12 : 3	0.03516
L1-SVM : LDF	12 : 4	0.07681
LDF : L2-SVM	11 : 7	0.4807

ここで表中の”12 : 3”とは 20 組のデータセットによる実験中 12 組で正判別率が L1-SVM が勝り 3 組で L2-SVM が,2 組は同じ正判別率であったことを示している. この実験では L1-SVM の正判別率が最も良く,L2-SVM での正判別率は線形判別関数とさほど変わらない結果を示した.

5.3　カーネル法を用いた場合

カーネルとしては, 多項式カーネル,Gaussian カーネルの 2 種類を用いた. 本要旨には Gaussian カーネルの結果を記す.

$$\text{GaussianKernel: } K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

ここで σ は Gaussian カーネルパラメータでありこれを決定する必要がある. 実験より正判別率が最も高い時のパラメータを採用した. パラメータを以下のように決定した.

Gaussian kernel	σ	C
H-SVM	2	-
L1-SVM	3	2
L2-SVM	5	4
ロジスティック回帰	2	-

このパラメータを用いて,5.2 節と同じ 20 組のデータセットで検討した. 各判別手法での 20 組の正判別率の平均を表 1(表中数値左列) に示す. また, 各カーネルを用いた H-SVM でカーネル法の適応により線形分離可能になったデータセットの正判別率の平均も記す (表中数値右列). 線形分離可能となったデータセットは Gaussian カーネルでは 15 組であった.

表 3　カーネルを用いた場合の正判別率平均

Gaussian カーネル	20 組 (全て)	15 組 (分離可能)
H-SVM	-	0.923
L1-SVM	0.955	0.954
L2-SVM	0.940	0.939
ロジスティック回帰	0.928	0.929

このデータにおいては正判別率は Gaussian カーネルを用いた場合の L1-SVM が最も優れた判別性能を示した. これはカーネルを使用しない SVM に対してカーネルを使用した場合, 学習データの取り方による正判別率の変化が少なくなり判別性能が同じソフトマージン SVM であっても安定することが確認されている. さらに多項式カーネルでは判別性能が下がってしまう結果が得られ,SVM, ロジスティック回帰の両方で判別性能が下がっていることからデータに合わせたカーネル関数の選び方が重要な問題になっていることを示している.

6　人工データによる数値的検討

人工データには識別面を視認できる 2 次元のデータを人工的に作成し実験を行う. 作成にあたり, 多変量正規分布, 多変量 t 分布, 混合正規分布のデータを作成したが, データ作成には R パッケージである mvtnorm パッケージを利用した. 人工データは 2 変数 2 群データであり各群 5000 例の 10000 例からなる. また変数とは別に 2 群はクラス 1= -1, クラス 2= 1 のクラスラベルを持つ. ここから各群 50 例ずつの 100 例を復元抽出し学習データに, 残った 9900 例をテストデータにしている. この様な学習データとテストデータの組を 100 例作成し, それぞれにおいて正判別率, 及び識別境界面を算出した.

6.1　多変量正規分布に従うデータによるシミュレーション

多変量正規分布に従う 2 次元データの作成においてはクラス 1 の 5000 例を分布の中心を $(x, y) = (0, 0)$ とし分散共分散行列を $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ の下で作成, クラス 2 の分布の中心を $(x, y) = (2, 2)$ とし同様の分散共分散行列を用いて 5000 例を作成しそこから前述の方法により 100 組のデータセットを得た. 100 例の平均正判別率を符号検定で見ていくと, 表 4 のようになった.

表 4　多変量正規分布における判別率の符号検定

	正判別率	p-value
L1-SVM : L2-SVM	46 : 53	0.5467
LDF:L1-SVM	85 : 15	4.825×10^{-13}
LDF:L2-SVM	83 : 17	1.310×10^{-11}

正判別率の観点では LDF>L1-SVM>L2-SVM となっ

た. つまり正規分布が完全に仮定できるならば群の端点の情報のみを用いるSVMに比べ,LDFの方が優れていることが分かる.

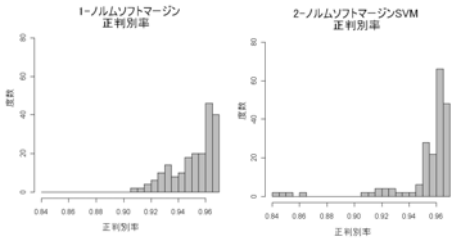
6.2 多変量t分布に従うデータによるシミュレーション

正規分布よりも外れ値が出やすいt分布を仮定した場合のシミュレーションを行った. 多変量t分布では自由度(df)の変化, 群の中心距離の変化などいくつかのケースで比較を行ったが本要旨ではdf=3のケースのみ説明するデータは正規分布の場合と同様の分散共分散行列を用いる. t分布の自由度は$df = 3$とし, 各群の中心はクラス1が$(x, y) = (0, 0)$, クラス2が$(x, y) = (4, 4)$となるように各群5000例ずつ,10000例を作成しシミュレーションを行った.100例の平均正判別率を符号検定で見ていくと,

表5 多変量t分布における判別率の符号検定

	正判別率	p-value
L1-SVM : L2-SVM	42 : 57	0.1591
LDF : L1-SVM	88 : 12	1.911×10^{-15}
LDF : L2-SVM	81 : 19	2.703×10^{-10}

となり, わずかながらL2-SVMの判別率がL1-SVMより高い.



上のようなヒストグラムのようL2-SVMははずれ値に影響されると大きく外れてしまうが, 大半はL1-SVMよりも良い判別性能を示している. L2-SVMは外れ値に影響を受けやすい反面, 外れ値の影響がない場合L1-SVMより優れた境界面を引くことができると考えられる.

6.3 多変量混合正規分布におけるシミュレーション

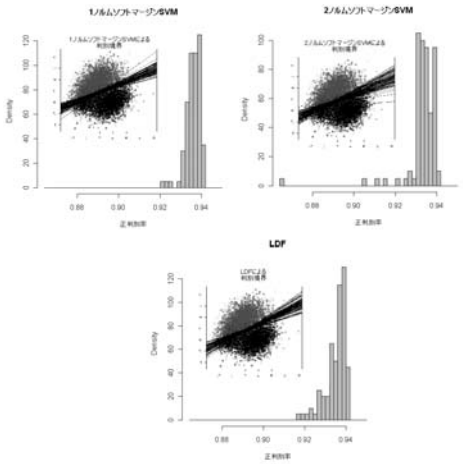
前節までは正規分布,t分布といった1つの中心の下に広がる分布であったが次に混合正規分布を用いて, シミュレーションを行う. 混合正規分布に使用した正規分布の平均(μ)と分散共分散行列Σ, 正規分布の混合比率(π)は以下のようになるクラス1では$\mu_{1a} = (0, 0), \Sigma_{1a} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$と$\mu_{1b} = (3.5, 0), \Sigma_{1b} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$の正規分布を混合比$\pi_1 = (0.2, 0.8)$で混合した分布に従うデータを5000例クラス2では$\mu_{2a} = (0, 2), \Sigma_{2a} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$と$\mu_{2b} = (2, 4), \Sigma_{2b} = \begin{pmatrix} 2 & 0.6 \\ 0.6 & 2 \end{pmatrix}$正規分布を混合比$\pi_2 = (0.3, 0.7)$で混合したデータを5000例作成し, 上述と同様の方法で100組のデータセットを作成している.

実験により得られた正判別率の符号検定は以下のようになった.

表6 混合正規分布における判別率の符号検定

	正判別率	p-value
L1-SVM : L2-SVM	64 : 36	0.0066
L1-SVM:LDF	61 : 39	0.0352
LDF:L2-SVM	59 : 39	0.05439

また100組の実験の正判別率のヒストグラム及び境界線は以下のようになった.



これはirisデータでの実験と似た状況を示しており, 分布がはっきりしない場合などにおいてはSVMが有効であることが確認された.

7 結果と考察

実験を通じ,SVMとその他の手法において比較を行ってきた. 初めのirisデータにおいてはSVMの有用性が改めて確認された. 特にカーネル法を導入したSVMはLDFに比べて大きく判別性能が優れていた. また人工データにより線形SVMと線形判別関数との比較を行った結果,LDFの前提条件である正規分布が仮定できる場合においてSVMよりもLDFの方が優れるという場面が確認され, 自由度が小さいt分布のような場合,L2-SVMは学習データのによっては大きく判別性能が下がるが多くは優れた判別性能を示す.L1-SVMは外れ値の影響が受けにくく判別性能が極端に低くなる事は少ないが, 全体的にはL2-SVMより若干劣る性能を示した. これに対し, 混合正規分布のような分布を仮定する場合,irisデータによる実験と同様にL1-SVM>L2-SVM≃LDFという優劣関係が確認され, 盲目的にSVMに優位性を信じるべきではなく, データの形状や外れ値といったものによってSVMよりもLDFといった判別手法が優れるということ様な状況考えられることが確認された.

参考文献

[1] Bernhard Scholkopf, Alexander J. Smola: Learning With Kernels〜Support Vector Machines, Regularization, Optimization and Beyond The MIT Press(2001)