サポートベクターマシンによる統計的判別
—線形判別関数と比較した統計的性質—

山田 俊哉* 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 [2]. 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, カーネル利用の有無における 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数値実験を行った.

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 [10]. 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 [3]. 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. カーネル学習法については後述する. 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

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

となり, このマージン γ は以下のようになる

*南山大学 数理情報研究科
†南山大学 数理情報学部

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

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

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. 今回は双対問題に変換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.

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

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

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. パラメータ $\mathbf{w}$ の決定には $\lambda_i^* = 0$ に対応する学習データ $\mathbf{x}_i$ は関与していない. つまり, 全ての学習データの中で $\lambda_i^* > 0$ となる一部の $\mathbf{x}_i$ のみが判別境界決定に関与し, それらは「Support Vector」と呼ばれ SVM の名前の由来もなっている [5]. 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

線形分離可能性は現実の問題では必ずしも満たされず実際は非線形で複雑な識別面が有効な場合が多い. 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元空間 (ヒルベルト空間) 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. i 番目のデータが反対側にどれくらい入り込んだかの距離を $\xi_i(\geq 0)$ と表す時, $\{\xi_i\}$ が全体としてできる限り小さいことが望ましい. 反対側に入り込んだ距離の 2 乗和を小さくすることを考えると, 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 ($\|\xi\|_1 = \sum_i \|\xi\|$) を用いる場合がある.L2 ノルムの場合, 双対空間における最適化問題は以下のようになる.

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

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

また 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がある. 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の計算が可能となる. 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用することが可能となる [4].

3　カーネルロジスティック回帰分析

ロジスティック回帰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用する例として, このロジスティク回帰モデルにカーネルを適用することを考える [7][17]. 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 $\mathbf{w}^* =< \Phi, \boldsymbol{\alpha} > +\mathbf{u}, \mathbf{u} \in \mathrm{span}(\Phi)^{\perp}$ と表されるが $< \Phi^T\mathbf{u} >= 0$ より, $\mathbf{u}$ は $\mathbf{g}(\boldsymbol{\mu})$ の説明に寄与しないため $\mathbf{w}^* \in \mathrm{span}(\Phi)$ 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 $\mathbf{K} = (\mathbf{K}_{ij}), \mathbf{K}_{ij} = K(\mathbf{x_i}, \mathbf{x_j}) =< \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) >$ となるカーネル行列である.

4　統計ソフト R への実装

後述の数値実験において, 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 [8]. また, 人工データの数値実験においては, 計算を高速化するため逐次最小最適化アルゴリズム (SMO algorithm: Sequential Minimal Optimisation algorithm) を採用した.SMO は反復の各ステップ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 [9]. 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 [1]. しかし同一の条件における L1-SVM, 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, 独自に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. そのデータの中から 50 例 (各群 25 例) をランダムに抽出して学習データとし, 残り 50 例のデータをテストデータとして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を用いた [14][15].

5.2　カーネル法を用いない場合

Fisher の線形判別関数 (LDF) および L1-SVM, L2-SVM を用いた. パラメータ C は実験で用いた 20 組のデータセットと別に 3 組のデータセットを作成し実験することにより, 正判別率が最も高い時のパラメータを

採用した. 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: iris データにおけるカーネルを用いない場合の平均正判別率

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

20 組の正判別率について符号検定を行うと以下のようになった.

表 2: iris データにおけるカーネルを用いない場合の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. ここでは 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. 実験で用いた 20 組のデータセットと別に 3 組のデータセットを作成し事前に実験することに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. 各判別手法を用いた時の, 20 組の正判別率の平均を表 3(表中数値左列) に示す. また, カーネルを用いた H-SVM でカーネル法の適用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: iris データにおける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. カーネルを使用しない SVM と比較して, 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と思われる結果が得られた. 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

6.1　データ作成

人工データには識別面を視覚的に見やすくするため 2 次元のデータを作成し実験を行った. 作成にあたり, 多変量正規分布, 多変量 t 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, 残ったもの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.

6.2　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

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{25}$$

とした. またクラス 2 の分布の中心を $(x, y) = (2, 2)$ とし同じ分散共分散行列を用いて 5000 例を作成し, そこから 6.1 節で述べた方法により 100 組のデータセットを得た. 以下には L1-SVM, L2-SVM, LDF によるそれぞれの識別境界直線を示す.

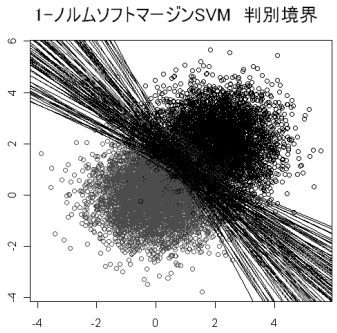


図 1: L1-SVM による多変量正規分布に従う 2 群の判別直線

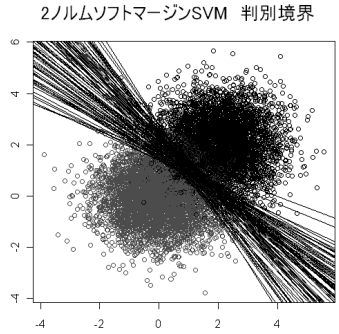


図 2: L2-SVM による多変量正規分布に従う 2 群の判別直線

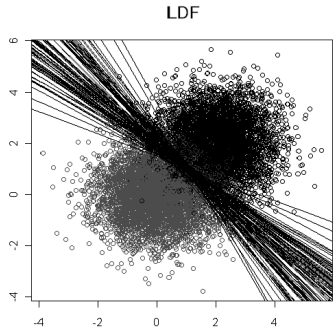


図 3: LDF による多変量正規分布に従う 2 群の判別直線

このように判別直線を見ていくと,100 本の判別直線は LDF が最もまとまりが良く, L1-SVM が最もまとまりが悪い結果になった. 100 組のデータセットでの正判別率を符号検定で見ていくと以下のようになった.

表 4: 多変量正規分布の場合の正判別率の符号検定と平均正判別率

	number of successes	p-value
L2-SVM : L1-SVM	59 : 38	0.04173
LDF : L1-SVM	77 : 23	5.514×10^{-8}
LDF : L2-SVM	72 : 28	1.258×10^{-5}
	平均正判別率	
L1-SVM	0.912729	
L2-SVM	0.914374	
LDF	0.916193	

正判別率の観点では LDF>L2-SVM>L1-SVM という優劣順が確認された. 100 組のデータセットによる正判別率のヒストグラムは以下のようになった.

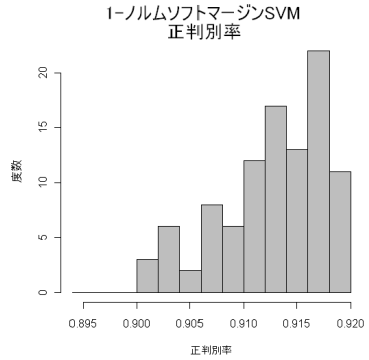


図 4: 多変量正規分布の場合の L1-SVM の正判別率ヒストグラム

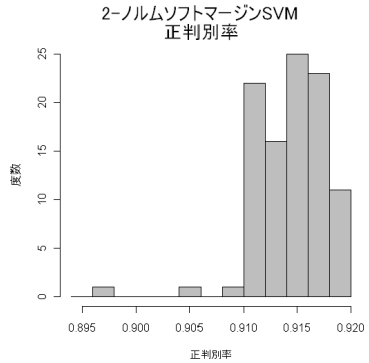


図 5: 多変量正規分布の場合の L2-SVM の正判別率ヒストグラム

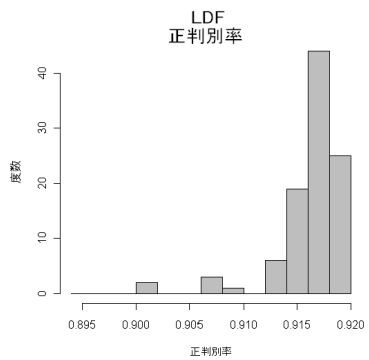


図 6: 多変量正規分布の場合の LDF の正判別率ヒストグラム

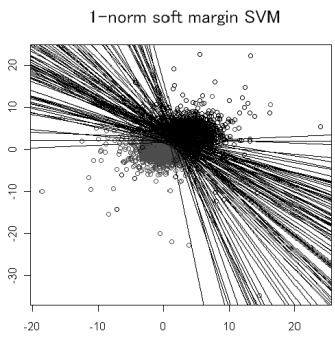
これは正規分布を仮定しているため Fisher の線形判別関数の前提条件を満たす場面である. つまり正規分布が仮定できるならば群の端点の情報のみを用いる SVM に比べ, 分布情報を用いる LDF の方が優れていることが分かる.

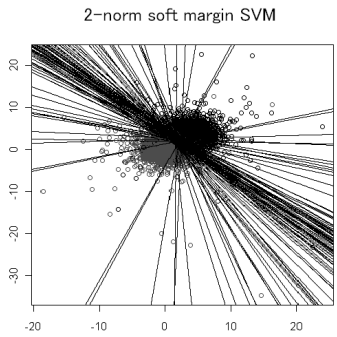
6.3 多変量 t 分布に従うデータによるシミュレーション

前節では 2 クラスとも正規分布に従う場合においてシミュレートした. 本節では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) の変化により外れ値の出現を変化させたいくつかのケースで比較した. すべてのケースにおいて 2 変量 10000 例 (各群 5000 例) とクラスラベル (-1,1) を持つデータを用いた.

6.3.1 ケース 1:df=3, 中心 (0,0),(4,4) の場合

データは正規分布の場合と同様, 分散共分散行列は式 (25) を用いた. 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 例を作成した. L1-SVM, L2-SVM, LDF における判別直線, 正判別率の変化はそれぞれ以下のようになった.





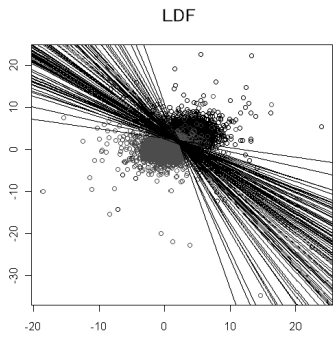


図 7: L1-SVM による多変量 t 分布 (df=3) に従う 2 群の判別直線

図 8: L2-SVM による多変量 t 分布 (df=3) に従う 2 群の判別直線

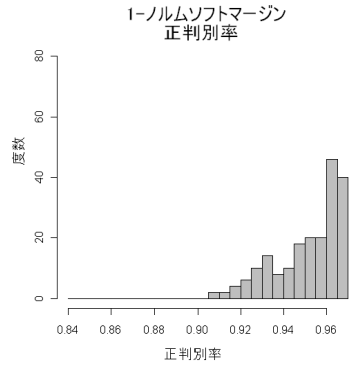
図 9: LDF による多変量 t 分布 (df=3) に従う 2 群の判別直線

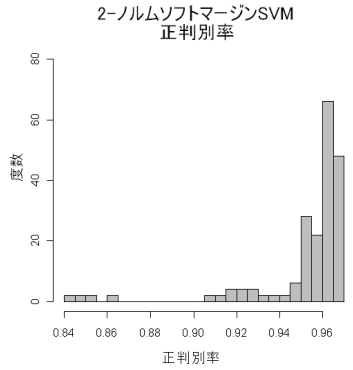
ここでは特に L2-SVM において直線の傾きのばらつきが極端に大きいことが解る. 同様に 100 例の平均正判別率を符号検定で見ていくと表 7 のようになった.

表 5: 多変量 t 分布 (df=3) の場合の判別率の符号検定

	number of successes	p-value
L1-SVM : L2-SVM	42 : 57	0.1591
LDF : L1-SVM	82 : 16	7.380×10^{-12}
LDF : L2-SVM	76 : 22	3.896×10^{-8}
	平均正判別率	
L1-SVM	0.951542	
L2-SVM	0.952621	
LDF	0.965022	

ここで着目すべきは L2-SVM であるが, 図 8 により境界線のばらつきは大きいものの中央付近にも境界線が集中しており, 高い正判別率を持つ境界線と低い境界線を持つ場合が混在していることが解る. L1-SVM と L2-SVM の 100 例の判別率のヒストグラムで見てみると以下のようになった.





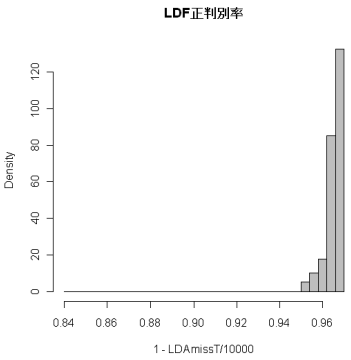


図 10: 多変量 t 分布 (df=3) の場合の L1-SVM の正判別率ヒストグラム

図 11: 多変量 t 分布 (df=3) の場合の L2-SVM の正判別率ヒストグラム

図 12: 多変量 t 分布 (df=3) の場合の LDF の正判別率ヒストグラム

このように,L2-SVM は判別率の低い判別直線が引かれているものの大半が判別率が 1 ノルムソフトマージンよりも優れている. このことは以下の判別率の基本統計量からもわかる

表 6: 多変量 t 分布 (df=3) の場合の正判別率

	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値	標準偏差
1 ノルムソフトージン SVM	0.9069	0.9420	0.9554	0.9515	0.9639	0.9677	0.0152
2 ノルムソフトージン SVM	0.8406	0.9535	0.9608	0.9526	0.9648	0.9677	0.0245
LDF	0.9503	0.9647	0.9661	0.9650	0.9671	0.9680	0.0034

6.3.2 ケース 2:df=6, 中心 (0,0),(4,4) の場合

次に 6.3.1 節の多変量 t 分布に比べ, 自由度を上げ外れ値を抑えた場合を考えた. 多変量 t 分布に用いる分散共分散行列は先ほどと同様であるが自由度を df= 6 とした場合を考えた. 学習データ, テストデータのデータセットはこれまでと同じように, 各組 5000 例の 10000 例を用いて実験するそれぞれの手法における 100 組のデータセットの判別直線は以下のような結果が得られた.

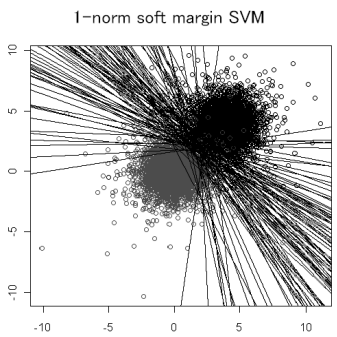


図 13: L1-SVM による多変量 t 分布 (df=6) に従う 2 群の判別直線

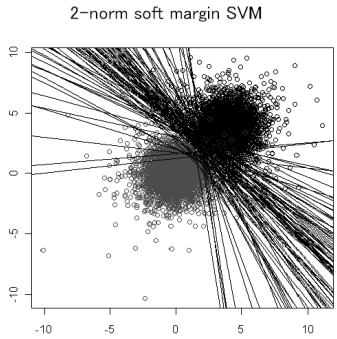


図 14: L2-SVM による多変量 t 分布 (df=6) に従う 2 群の判別直線

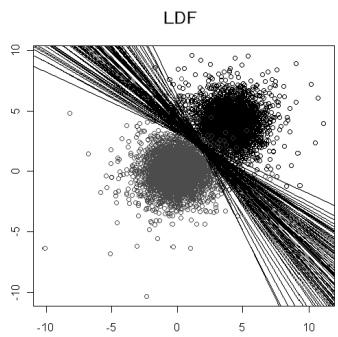


図 15: LDF による多変量 t 分布 (df=6) に従う 2 群の判別直線

100 例の平均正判別率を符号検定で見ていくと表 7 のようになった.

表 7: 多変量 t 分布 (df=6) の場合の正判別率の符号検定

	number of successes	p-value
L1-SVM : L2-SVM	46 : 54	0.4841
LDF : L1-SVM	89 : 10	2.2×10^{-16}
LDF : L2-SVM	86 : 14	8.284×10^{-14}
	平均正判別率	
L1-SVM	0.972694	
L2-SVM	0.975721	
LDF	0.986379	

表 8: 多変量 t 分布 (df=6) の場合の判別率

	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値	標準偏差
1 ノルムソフトージン SVM	0.8915	0.9695	0.9790	0.9727	0.9844	0.9882	0.0178
2 ノルムソフトージン SVM	0.9217	0.9694	0.9808	0.9757	0.9845	0.9882	0.0132
LDF	0.9808	0.9858	0.9866	0.9864	0.9872	0.9881	0.0013

このように自由度を 6 とし外れ値が少なくなったことで L2-SVM の性能が良くなったことがわかる. しかし, 外れ値の影響がなくなり群内分散が低くなったことで LDF の判別性能がさらに良くなっていることもうかがえる結果となった.

6.3.3 ケース 3:df=2, 中心 (0,0),(4,4) の場合

6.3.2 節では自由度をあげて外れ値の出方を抑えたが, 本節では自由度を df= 2 に下げ, 外れ値を多く出した場合を考えた.100 組の学習データ, テストデータのデータセットにおける判別境界はそれぞれ以下のようになった. ここで, 大きく外れたデータがあるため, 作図範囲を両軸ともに $[-10, 10]$ にした図と全体図を記す.

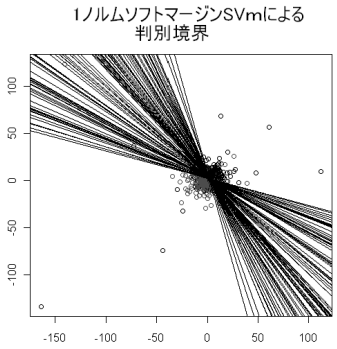


図 16: L1-SVM による t 分布 (df=2) に従う 2 群の判別直線

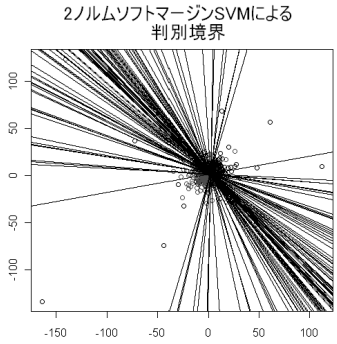


図 17: L2-SVM による t 分布 (df=2) に従う 2 群の判別直線

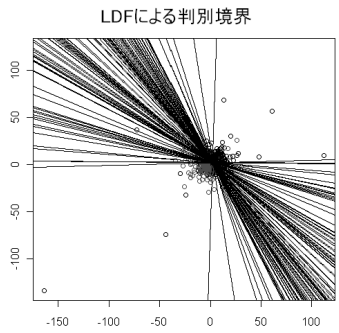


図 18: LDF による t 分布 (df2) に従う 2 群の判別直線

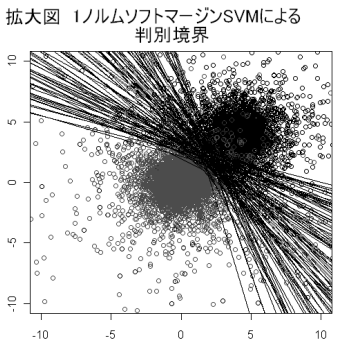


図 19: L1-SVM による t 分布 (df=2) に従う 2 群の判別直線:拡大図

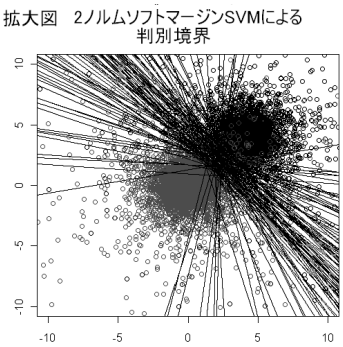


図 20: L2-SVM による t 分布 (df=2) に従う 2 群の判別直線:拡大図

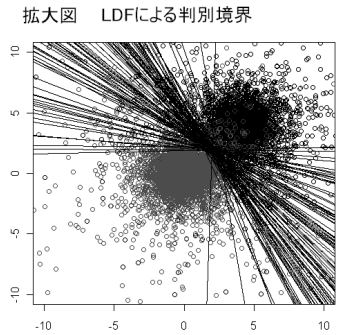
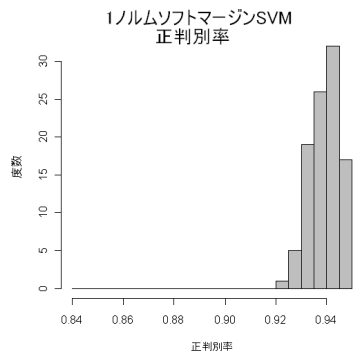
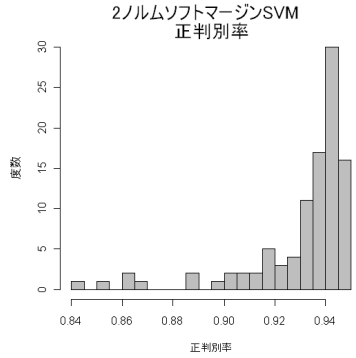


図 21: LDF による t 分布 (df=2) に従う 2 群の判別直線:拡大図

このように外れ値が多く存在する場合,L1-SVM の判別境界はまとまって出るのに対し, L2-SVM と LDF は外れ値の影響が色濃く出てしまう結果となった.

次に, 正判別率のヒストグラムを見てみると以下のようになった.





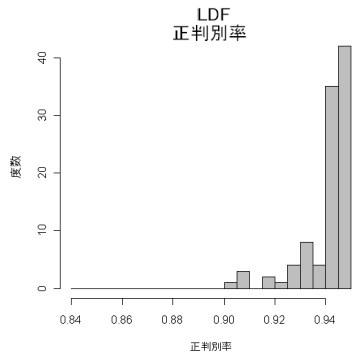


図 22: 多変量 t 分布 (df=2) の場合の L1-SVM の正判別率ヒストグラム

図 23: 多変量 t 分布 (df=2) の場合の L2-SVM の正判別率ヒストグラム

図 24: 多変量 t 分布 (df=2) の場合の LDF の正判別率ヒストグラム

判別境界と正判別率の分布をみると L1-SVM の判別性能が比較的まとまっていることがわかる. 判別性能の優劣を見るため符号検定を行うと

表 9: 多変量 t 分布 (df=2) の場合の正判別率の符号検定

	number of successes	p-value
L1-SVM : L2-SVM	57 : 39	0.0822
LDF : L1-SVM	64 : 35	0.004641
LDF : L2-SVM	67 : 32	0.0005622
	平均正判別率	
L1-SVM	0.939252	
L2-SVM	0.931146	
LDF	0.940577	

となり, 判別境界は L1-SVM がまとまっていたものの, LDF が最もすぐれた判別性能を示す結果となった.

6.4 多変量混合正規分布に従うデータによるシミュレーション

これまで, 正規分布,t 分布に従う場合の比較を行ってきた. そこで次に分布を単一で同一な分布ではなく混合正規分布に従うデータセットを考えた. 混合正規分布に従うデータセットの作成ではクラス i において 2 つの異なる正規分布 ($N_{ia}(\boldsymbol{\mu}_{ia}, \Sigma_{ia})$,$N_{ib}(\boldsymbol{\mu}_{ib}, \Sigma_{ib}), i \in \{1, 2\}$) を混合比 π_i で混合した. それぞれのクラスにおける混

合正規分布を以下に記す.

クラス $1(t=-1)$

$$N_{1a}(\boldsymbol{\mu}_{1a},\Sigma_{1a}): \quad \boldsymbol{\mu}_{1a}=(0,0), \Sigma_{1a}=\begin{pmatrix}1 & 0.4\\ 0.4 & 1\end{pmatrix} \tag{26}$$

$$N_{1b}(\boldsymbol{\mu}_{1b},\Sigma_{1b}): \quad \boldsymbol{\mu}_{1b}=(3.5,0), \Sigma_{1b}=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix} \tag{27}$$

$$\text{混合比}: \quad \pi_1=(N_{1a},N_{1b})=(0.2,0.8) \tag{28}$$

クラス $2(t=1)$

$$N_{2a}(\boldsymbol{\mu}_{2a},\Sigma_{2a}): \quad \boldsymbol{\mu}_{2a}=(0,2), \Sigma_{2a}=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix} \tag{29}$$

$$N_{2b}(\boldsymbol{\mu}_{2b},\Sigma_{2b}): \quad \boldsymbol{\mu}_{2b}=(2,4), \Sigma_{2b}=\begin{pmatrix}2 & 0.6\\ 0.6 & 2\end{pmatrix} \tag{30}$$

$$\text{混合比}: \quad \pi_2=(N_{2a},N_{2b})=(0.3,0.7) \tag{31}$$

このデータセットに対して L1-SVM, L2-SVM, LDF による判別境界は以下のようになった.

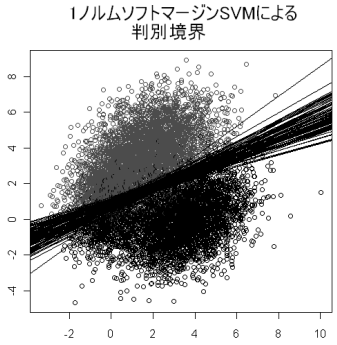


図 25: L1-SVM による混合正規分布に従う 2 群の判別直線

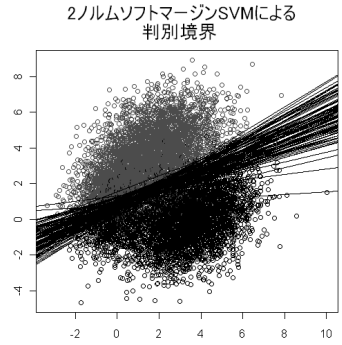


図 26: L2-SVM による混合正規分布に従う 2 群の判別直線

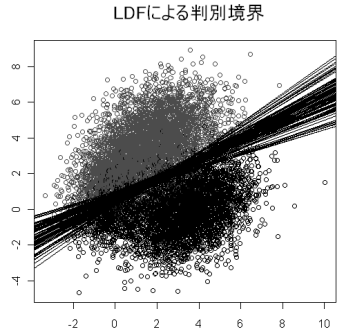


図 27: LDF による混合正規分布に従う 2 群の判別直線

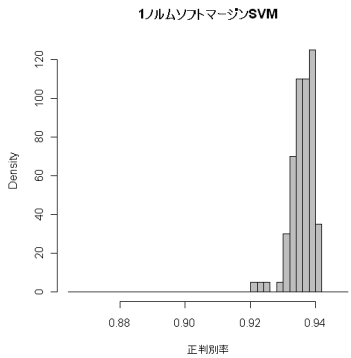


図 28: 混合正規分布の場合の L1-SVM の正判別率ヒストグラム

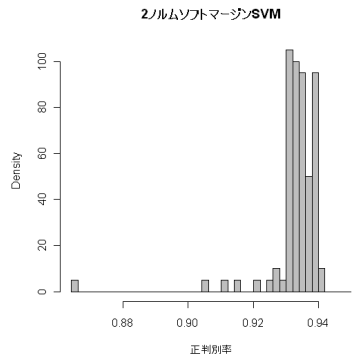


図 29: 混合正規分布の場合の L2-SVM の正判別率ヒストグラム

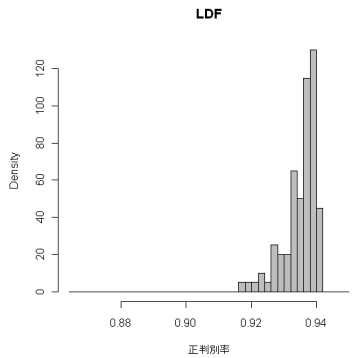


図 30: 混合正規分布の場合の LDF の正判別率ヒストグラム

ここで特筆すべきはこれまで様相の異なっていた判別境界がほとんどまとまっている点, L1-SVM, L2-SVM

ともに LDF よりも性能が優れているという 2 点である. 正判別率で見ていくと符号検定の結果及び正判別率の平均は以下のような結果が得られた.

表 10: 混合正規分布の場合の正判別率の符号検定

	number of successes	p-value
L1-SVM : L2-SVM	64 : 36	0.006637
L1-SVM :LDF	61 : 39	0.0352
LDF : L2-SVM	59 : 39	0.05439
	平均正判別率	
L1-SVM	0.936041	
L2-SVM	0.932911	
LDF	0.935132	

このような結果となり, 前章の Iris データでの実験の L1-SVM>L2-SVM≃LDF という結果が再現できたと考えられる. このように混合正規分布のような複雑な分布においては SVM の方が判別性能が優れることが確認された.

7 結果と考察

この研究ではサポートベクターマシンの統計フリーソフト R でのプログラムの実装, 様々なデータに対して既存の判別手法と SVM との判別性能と特徴の比較を行ってきた. プログラムの作成ではアルゴリズムとして慣性法に加え逐次最小最適化 (SMO) アルゴリズムを採用し, さらにソフトマージンについて判別境界からのはみ出しの距離のノルムの取り方を 2 種類 (1 ノルム, 2 ノルム) にとる方法を採用し SVM プログラムを作成した. 実験ではこれまで行ってきた”iris”データについての解析に加えて, 人工的に多変量正規分布および多変量 t 分布により作成したデータについて実験を行った.”iris”のようなデータに対しては SVM は既存の手法に比べ高い判別性能を示した. 以下に結論をまとめると, iris データの実験より

1.カーネルを使用しない場合

$$\text{L1-SVM} > \text{L2-SVM} \simeq \text{LDF} \tag{32}$$

2.カーネルを使用した場合

$$\text{L1-SVM} > \text{その他の手法} \tag{33}$$

3.カーネル関数の適用の有無による比較

$$\text{カーネル L1-SVM} > \text{非カーネル L1-SVM} \tag{34}$$

ここでこの実験の前に本研究においていくつかの作業仮説を持っていた. その作業仮説は以下の通りである.

1. 線形分離可能なデータの場合, ハードマージン SVM の方がソフトマージン SVM よりも優れている
2. 線形分離可能なデータの場合, カーネル法を用いない SVM の方がカーネルを用いた SVM よりも優れている
3. 線形分離不可能なデータの場合, 非線形に識別面が構成できるカーネル法を適用した SV mの方がカーネル法を用いない場合の SVM よりも優れている.
4. L2-SVM はデータが正規分布に従うような状況では効果的だが, 外れ値が存在する場合は L1-SVM の方が優れている.

5. 完全に正規分布が仮定できるとき, LDF が SVM よりも優れている

Fisher の iris データは比較的正規分布に近いデータと考えていたが, このデータの実験において, 最も高い判別性能を示したのは Gaussian カーネル L1-SVM であった. カーネルを使わない場合でも L1-SVM>LDF であり上記の作業仮説 4 に反し, 疑問であった. そのため, 次に人工データにより多変量正規分布や多変量 t 分布に従う場合のデータに対しカーネルを用いない線形 SVM と LDF との比較実験を行った. 数値実験の結果得られたそれぞれの分布において, 符号検定 (有意水準 10%) と平均正判別率から得られた L1-SVM, L2-SVM と LDF の判別性能の順位は下表のようになる.

表 11: カーネルを用いない場合の判別性能の順位

分布	LDF	L2-SVM	L1-SVM
正規分布	1	2	3
t 分布 ($df = 6$)	1	2	2
t 分布 ($df = 3$)	1	3	2
t 分布 ($df = 2$)	1	3	2
混合正規分布	2	3	1

今回正規分布, t 分布のいずれの場合も 2 群の分散共分散行列が等しい場合について検討した. これらの数値実験の範囲では以下のこと言えるだろう.

1. LDF の前提条件である分散共分散行列が等しい正規分布の場合, LDF が SVM より明らかに優れる.
2. 自由度がかなり小さい t 分布においても LDF が SVM より優れる.
3. L1-SVM と L2-SVM を比較すると, 自由度が小さい t 分布 (df=2,3) のときは L1-SVM が, 自由度が大きい t 分布 (df=6) や正規分布のときは L2-SVM が優れる
4. L1-SVM はロバストな性質をもつ. すなわち, 例えば, df=2 の t 分布の場合, LDF が平均的には優れる場合であっても, 最悪のケースを比較すると, LDF の最悪なケースは L1-SVM の最悪なケースより正判別率が小さくなる. このことは L2-SVM と L1-SVM との比較でも同様のことが言える.

また、混合正規分布のように分布が複雑な場合は LDF よりも L1-SVM が優れた判別性能を示す結果も実験より得られた. 線形判別分析に限って言えば, 正規分布や t 分布のように単一の分布が仮定できる場合 LDF が安定した判別性能を示すことが多く, SVM は分布の情報ではなく群の端点の情報のみを使用しているという SVM の特徴が短所となってしまうような場面も見られた. しかし混合正規分布のような分布の場合, 端点のみの情報を用いる SVM が LDF より優れることも確認ができた.

今回の比較では SVM の利点といわれる高次元データへの適用の優位性やカーネル法の導入の容易さは考慮していないが, SVM 以外にもカーネル法を導入した判別手法が昨今開発されており, カーネル法が適用しやすいというアドバンテージが SVM のものだけではなくなってきている中では線形 SVM と LDF の比較によって得られたこのような統計的な判別手法としての SVM の性質, つまり SVM が必ずしも万能な判別手法ではなく得手不得手とされるデータ構造が存在すること確認される結果となった.

参考文献

[1] Alexandros Kartzoglou , Meyer David, Hornik Kurt : *Support Vector Machines in R*, Wirtschafts universitat (2005)

[2] 赤穂昭太郎：カーネル多変量解析-非線形データの新しい展開, 岩波書店 (2008)

[3] 麻生英樹, 津田宏治, 村田昇：パターン認識と学習の統計学-新しい概念と手法, 岩波書店 (2003)

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

[5] Cristianini Nello ,John Shawe-Taylor, 大北 剛　訳：サポートベクターマシン入門, 共立出版 (2005)

[6] 江口真透: 統計的パターン認識 線型判別からアダブーストまで, 日本化学会情報化学部会誌 Vol.25 No.3 68-75 (2007)

[7] Ikeda.S,Tsuchiya.J and Sato.Y: *Kernel Regression and Variable Selection Problem*, Proceedings of The Ninth Japan-China Symposium on Statistics,75-80 (2007)

[8] S.L.S.Jacoby ,J.S.Kowalik ,J.T.Pizzo, 関根智明　訳：非線形最適化問題の反復解法, 培風館 (1976)

[9] John C. Platt: *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Technical Report MSR-TR-98-14 (1998)

[10] 鳥脇純一郎：認識工学-パターン認識とその応用, コロナ社 (1993)

[11] 豊田秀樹：非線形多変量解析〜ニューラルネットによるアプローチ, 朝倉書店 (1996)

[12] 津田宏治：パターン認識と学習の統計学-第 2 部:カーネル法の理論と実際, 岩波書店 (2003)

[13] Vladimir Naumovich Vapnik: *Statistical Learning Theory*, Wiley (1998)

[14] Toshiya Yamada,Yutaka Tanaka: *Statistical Classification using Support Vector Machines* Proceedings of The Ninth Japan-China Symposium on Statistics,361-366 (2007)

[15] Toshiya Yamada,Yutaka Tanaka: *Support Vector Machines and Statistical Classification with and without Kernel Functions* Proceedings of IASC2008,1717-1726 (2008):

[16] 山下浩, 田中茂：サポートベクターマシンとその応用, 第 13 回日本 OR 学会 RAMP シンポジウム論文集 (2001)

[17] Ji Zhu,Trevor Hastie: *Kernal Logistic Regression and the Import Vector Machine*, Journal of Computational and Graphical Statistics Vol.14, No.1, 185-205 (2005)