カーネルを用いたパターン認識に関する一考察

知識情報処理研究　　3603R017-1　北川暁士
指　　導　松嶋敏泰教授

A Note on Pattern Recognition with Kernel Functions

by Satoshi Kitagawa

1 はじめに

パターン認識とは，認識対象がいくつかの概念に分類できるとき，観測されたパターンをそれらの概念のうちのひとつに対応させる処理のことを言う．この概念をクラスと呼ぶ．その応用範囲は広く，手書き数字認識や文字認識，また，さまざまな因子からある人が病気であるかどうかを判定するという問題も，パターン認識問題である．本研究では，2 クラスの識別問題の学習を考えるが，3 クラス以上の識別問題も複数の 2 クラスの識別問題へと帰着されるため，本研究の内容は多クラスの識別問題への応用も可能である．

近年，Support Vector Machine（以下，SVM）と呼ばれる 2 クラスの識別手法が注目を浴びている．その背景には，カーネル関数と呼ばれるものの存在がある．SVM は当初，2 次計画法を使って学習する線形識別器として提案されたが，現実に線形関数でうまく識別できる問題は限られており，現在のような注目を集めるにはいたらなかった．後に，カーネル関数を利用した非線形写像を使うことで，非線形識別器へと拡張され [5]，現在に至る．非線形に拡張されたことで，適用範囲が大きく広がったこと，また，計算量が他の非線形識別器と比較して少ないこと，局所解の問題がないことなどから，多くの実問題に利用されている [8]．

SVM による識別性能は，カーネル関数に依存する．しかし，一般にカーネル関数は未知であり，実験的にカーネル関数を決めるなどの方法が取られていた．それに対し，従来，SVM の確率構造を仮定することによりカーネル関数のパラメータを推定する手法が考案されている [1],[3]．しかし，そのようにして選ばれたカーネル関数を用いて識別を行ったとしても，識別に対する理論的な保証は得られない．そこで，本研究では，カーネル関数のパラメータが未知である元で，ベイズ基準の下で識別誤り最小にする識別法を示す．

2 準備

2.1 パターン識別問題

ここでは，本研究が扱う 2 クラスのパターン識別問題について述べる．認識する対象が，m 次元ベクトル $\mathbf{X} \in \mathcal{X}, \mathcal{X} \subset \mathbb{R}^m$ で与えられ，対応するクラスラベルを $Y \in \{-1, +1\}$ と表されるとする．このとき，与えられた訓練サンプル集合 $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$ より，新たな入力 $\mathbf{x}_{n+1}$ に対する出力 y_{n+1} を予測する関数 g を求める問題をパターン識別問題と呼ぶ．

$$g(\mathbf{X}) = Y. \tag{1}$$

関数 g の表し方は用いる識別手法によって異なる．

2.2 線形 SVM

一般に線形識別関数は

$$g(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^T\mathbf{x} + b), \tag{2}$$

$$\text{ただし,} \mathrm{sgn}(z) = \begin{cases} 1 & ; z \geq 0, \\ -1 & ; z < 0, \end{cases} \tag{3}$$

で表される．線形 SVM では，未知パラメータ $\mathbf{w} \in R^d$ ，$b \in R$ を識別境界と学習データとの最小距離（マージン）が最大になるように求める．

学習データの各パターンと識別超平面との最小距離は，

$$\min_{i=1,2,\cdots,n} \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{\| \mathbf{w} \|}, \tag{4}$$

で表されるので，$\min |\mathbf{w}^T\mathbf{x}_i + b| = 1$ の制約の下で以下の最小化問題を解くことに帰着する．

$$\min \quad \| \mathbf{w} \|, \tag{5}$$

$$s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, (i = 1, 2, \cdots, n). \tag{6}$$

ここで，ラグランジュの乗数 $\alpha_i (i = 1, 2, \cdots, n)$ を導入し，$\mathbf{w}$ に関する微分によって得られる式を代入すると，

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j, \tag{7}$$

$$s.t. \quad \alpha_i \geq 1, (i = 1, 2, \cdots, n), \tag{8}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{9}$$

の 2 次計画問題となる．重み $\mathbf{w}$ は $\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ と書くことができ，この解を (2) 式に代入すると，

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b \right), \tag{10}$$

が得られる．バイアス項 b は，$b = -\frac{1}{2}(\mathbf{w}^T\mathbf{x}_{+1} + \mathbf{w}^T\mathbf{x}_{-1})$ により求められる．$\mathbf{x}_{+1}$，$\mathbf{x}_{-1}$ はそれぞれ，クラス+1，-1 に属し，$(\mathbf{w}^T\mathbf{x} + b) = 1$ を満たすベクトル（サポートベクトルという）である．

2.3 非線形 SVM とカーネルトリック

ここでは前述のように，線形分離可能でない問題に対して，SVM がどう対処しているかについて述べる．そのためには $\mathbf{x}$ そのままではなく，いったん $\mathbf{x}$ を高次元空間 (M 次元特徴空間とする) $\mathcal{F} \subset \mathbb{R}^M$ に写像した $\Phi(\mathbf{x})$ を (2) 式に入れてやればよい．一般に特徴空間の次元 M を n 以上にすればサンプルは線形分離可能になる．そこで以下では識別関数を，

$$g(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^T\Phi(\mathbf{x}) + b), \tag{11}$$

と定義する．ただし，ここでの $\mathbf{w}$ は (2) 式のものと次元が異なる．すると，(10) 式で与えられる最終的な識別関数は，

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + b \right). \tag{12}$$

で与えられる．

ここで，特徴空間上の 2 つのベクトル間の内積を $k(\mathbf{x}, \mathbf{x}')$ と書くことにする．すなわち，

$$k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}'), \tag{13}$$

となる．この関数 k をカーネル関数と呼び，これを用いて，(12) 式を書き直すと，

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right). \tag{14}$$

と表すことができる．実際に用いられているカーネル関数は数多くあるが，計算が簡単なものとしては，

$$\text{シグモイドカーネル: } \tanh(a < \mathbf{x}, \mathbf{x}' > -b), \tag{15}$$

$$\text{ガウシアンカーネル: } \exp(\frac{-\|\mathbf{x} - \mathbf{x}'\|}{2\sigma^2}), \tag{16}$$

$$\text{多項式カーネル: } (< \mathbf{x}, \mathbf{x}' > +1)^p, \tag{17}$$

などが知られている．

2.4 ソフトマージン

ここでは，写像された空間上でも線形分離できない場合について考える．線形分離できない場合，誤って識別される訓練サンプルが存在する．そこで，そのようなサンプルに対し，損失を置くことにすると，以下の最適化問題となる．スラック変数 $\xi_i, i = 1, 2, \cdots, n$ を導入し，最適化問題を，

$$\min . \quad \frac{1}{2} \| \mathbf{w} \| + C \sum_{i=1}^{n} \xi_i, \tag{18}$$

$$s.t. \quad y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, (i = 1, \cdots, n), \tag{19}$$

とする．ここで，スラック変数 $\xi_i, i = 1, 2, \cdots, n$ が誤識別サンプルに対する損失で，C は損失をどの程度大きくとるかを決めるパラメータである．この最適化問題の変更に伴い，ラグランジュ乗数 α_i に関する問題が，制約式が $0 \leq \alpha_i \leq C$ に変更される．

3 従来研究

3.1 SVM の確率モデル

入力 $\mathbf{x}$ と，出力 y の間に，以下の確率モデルを仮定する [1],[4],[3]．これは，(18) 式で表されるソ

フトマージンSVMの目的関数の第2項をSVMのパラメータの尤度関数であると考えることによる．

$$P(y|\mathbf{x},\mathbf{w}) = \exp(-l(yf)). \tag{20}$$

where
$f = \mathbf{w}^T\Phi(\mathbf{x}) + b$, $l(z) = (1-z)H(1-z)$,
$H(a) = \begin{cases} 1\ ; a \geq 1, \\ 0\ ; \text{otherwise.} \end{cases}$

3.2 カーネル関数のパラメータ推定

パラメータである重み $\mathbf{w}$，バイアス b の事前分布 $P(\mathbf{w})$ を各要素が独立な多変量正規分布とすることにより，パラメータの関数である f についての事前分布 $P(f)$ を正規過程として考えることができる．訓練サンプルが与えられると，カーネルのパラメータ θ の尤度は，

$$P(y^n|\mathbf{x}^n,\theta) = \int d\mathbf{w} P(y^n|\mathbf{x}^n,\mathbf{w},\theta)P(\mathbf{w}|\theta), \tag{21}$$

$$= \int df^n P(y^n|\mathbf{x}^n,f^n,\theta)P(f^n|\theta), \tag{22}$$

で与えられる．ここで，$P(f^n)$ は n 変量正規分布で，分散共分散行列がカーネル関数によって与えられるものになる．この f 上での積分計算への変換によって，高次元の特徴空間に関係なくカーネルパラメータの尤度が計算できるが，この積分は解析的には解けないので，Variational Method を用いて近似的に解くことになる [3]．その他の近似アルゴリズムとして，ラプラス近似 [6] や Markov Chain Monte Corlo[7]，Mean Feild Algorithm[2] などが研究されている．従来研究では，求めた尤度を元に，カーネルのパラメータを推定し，推定したカーネル関数を用いてSVMを行うことで識別している．

3.3 従来研究の問題点

(20) 式で仮定した確率モデルは，SVMから導出されているが，厳密に確率分布の形になっていない．また尤度計算自体も，解析的に扱えないために,近似アルゴリズムを用いる．そのため，カーネルパラメータの尤度が正しく計算されている保証はない．また，計算が正確であるとしても，尤度によるパラメータの推定は，有限個の訓練サンプルのもとでの識別率は理論的に保証されない．

4 本研究の提案

本研究では新たな確率モデルを導入し，ベイズ基準の下で誤り率を最小にする識別法を示す．

4.1 確率モデルの導入

本研究では，カーネル関数として，(16) 式の多項式カーネルを用いる．多項式カーネルは有限な特徴空間への写像を意味し，その写像が明確になる．次数 p の多項式カーネルを用いる場合，その写像は，$M =_{m+p} C_p$ 次元ベクトルとなり，

$$\Phi_p(\mathbf{x}) = \left\{ a_{d_1,d_2,\cdots,d_m} x_1^{d_1} x_2^{d_2} \cdots x_m^{d_m} | \sum_{l=1}^{m} d_l = p \right\}, \tag{23}$$

で与えられる．写像が特定されることにより，特徴空間上でのパラメータ $\mathbf{w}, b$ の学習が可能になるので，ベイズ決定理論を利用してベイズ最適な識別が行える．以下では表記上の簡便さのため，バイアス項 $b = 0$ で議論する．これが成り立たない場合も，$\mathbf{w}$ に含めることで同様の議論ができる．

まずSVMと対応する確率モデルとして，p を写像を示すインデックスとした，以下の確率構造を仮定する．

$$P(Y|\mathbf{X},\mathbf{w},p) = \begin{cases} \frac{\exp(-\mathbf{w}_p^T\Phi(\mathbf{x}))}{1+\exp(-\mathbf{w}_p^T\Phi(\mathbf{x}))}; Y = 1, \\ \frac{1}{1+\exp(-\mathbf{w}_p^T\Phi(\mathbf{x}))}; Y = -1 \end{cases}. \tag{24}$$

4.2 ベイズ基準のもとで最適な識別

上記の確率モデルの下で，ベイズ基準で最適な識別法を導出する．
損失関数 L を以下のように定義する．

$$L(y_{n+1},\hat{y}_{n+1}) = \begin{cases} 1\ ; y_{n+1} \neq \hat{y}_{n+1}, \\ 0\ ; y_{n+1} = \hat{y}_{n+1}. \end{cases} \tag{25}$$

危険関数:

$$R(\hat{y}_{n+1},\mathbf{w},p) = \int dy_{n+1} dy^n L(y_{n+1},\hat{y}_{n+1})$$
$$\times P(y_{n+1}|\mathbf{x}_{n+1},\mathbf{w},b,p)P(y^n|\mathbf{x}^n,\mathbf{w},p). \tag{26}$$

ベイズ危険関数:

$$BR(\hat{y}_{n+1}) = \int R(\hat{y}_{n+1}, \mathbf{w}, p)P(\mathbf{w}|p)P(p)d\mathbf{w}dp$$
$$= \int L(y_{n+1}, \hat{y}_{n+1})P(Y_{n+1}|\mathbf{x}_{n+1}, \mathbf{w}, p)$$
$$\times P(\mathbf{w}, p|y^n, \mathbf{x}^n,)dy_{n+1}dy^n d\mathbf{w}dp. \tag{27}$$

よって，ベイズ危険関数を最小化する（ベイズ基準で最適な）$\hat{y}_{n+1}^{bayes}$ は以下で求められる．

$$\hat{y}_{n+1}^{bayes} = \arg\min_{\hat{y}_{n+1}} \int d\mathbf{w}dpL(y_{n+1}, \hat{y}_{n+1})$$
$$\times P(\mathrm{y}_{n+1}|\mathbf{x}_{n+1}, \mathbf{w}, p)P(\mathbf{w}|p, \mathbf{x}^n, y^n)P(p|\mathbf{x}^n, y^n). \tag{28}$$

よって，w，p の事後分布が求められれば良い．この計算は解析的に求まらないので，次節で近似アルゴリズムを提案する．

4.3 近似アルゴリズム

近似アルゴリズムの着目点は，多項式カーネルのパラメータが離散であること，SVM アルゴリズムを利用することである．真のカーネルパラメータ（認識対象がそのカーネルパラメータから発生している）の候補が分かっている場合を考える．p の候補を $p = 1, 2, \cdots, q$ とする．

【アルゴリズム】

Step1 候補となる各 p の値に対し，SVM で識別関数を学習する（w_p^{svm} を求める）．ただし，$p = 1, 2, \cdots, q$.

Step2 各識別において，
$\Pi_{i|y_i\mathbf{w}^T\Phi(\mathbf{x}_i)<1} \frac{1}{(1+\exp(y_i\mathbf{w}_p^{\mathrm{svm}T}\Phi(\mathbf{x}_i)))}$ を求める．この値を l_p とする．

Step3 近似事後確率 $P_a(p|y^n, x^n)$ を，
$P_a(p|y^n, x^n) = \frac{l_p \times P(p)}{\sum_{p=1}^{q} l_p \times P(p)}$ で求める．

5 考察

従来研究では，カーネルのパラメータを 1 点に決め，それを用いて識別を行っていた．これは，訓練サンプルが有限個であることを考えると，識別という目的に対しては理論的な保証がない．そこで本研究では，仮定した確率モデルで，パラメータの事後確率を用い，ベイズ基準で最適な識別方法を提案した．しかし，パラメータの事後確率は解析的には求まらないので，確率構造と SVM のアルゴリズムの関係を利用し，近似的に事後確率を求める方法を提案した．

6 まとめ

従来研究では SVM のカーネルのパラメータを確率モデルを用いて推定していた．本研究では，多項式カーネルを用いる SVM に対するカーネルのパラメータの重み付け方法を，確率モデルを用いて提案した．これにより，その確率モデルのもとで理論的に誤り最小が保証される．今後は，実験による検証と，SVM と仮定する確率モデルとの整合性を理論的に考察していくことが必要であると考えられる．

参考文献

[1] C. Gold, P. Sollich, Model Selection for Support Vector Machines, Neurocomputing,55,pp221-249,2003.

[2] M. Opper, and O. Winther,“ Gaussian Process Classification: Mean Field Algorithm ”,Neural Computation vol.12, issue 11, pp.2655-2684, 2002.

[3] M. Seeger, Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers, Neural Information Processing Systems 12, pp.603-609,2000.

[4] P. Sollich. Bayesian methods for Support Vector Machines: Evidence and predictive class probabilities, Machine Learning, 46, pp.21-52, 2002.

[5] Vladimir N. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

[6] C. K. I. Williams and D. Barber, “ Bayesian Classification With Gaussian Processes, ”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.20, no.12, 1998.

[7] R. M. Neal,“ Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification ”, Technical Report 9702, Dept. of Statistics, University of Toronto, 1997

[8] 松本 裕治,“ 自然言語処理におけるカーネル法の利用 ”第 5 回 情報論的学習理論ワークショップ予稿集，pp.19-24, 2002.