手書き数字認識におけるカーネル法を用いた識別器の構成

坂本美虹　(指導教員：吉田裕亮)

1　はじめに

パターン認識とは, いくつかの概念に分類できる観測データが存在するとき, 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ことである.

パターン情報処理技術において, さまざまな確率分布モデルに基づいた多種多様な統計的手法が大きな役割を果たしてきた. 現在ではパターン情報処理は統計科学の応用の主要な一分野となっている.

そこで, 本研究では手書き数字におけるパターン認識の一手法として, カーネル法を用いて識別器の構成を行うことを検討する.

2　カーネル法

カーネル法とは, データの高次モーメントを有効に抽出し, かつ必要な計算を効率的に実行可能にするような非線形変換の方法論である.

データが存在する空間を Ω とし, これをある実ベクトル空間 H に写像することによりデータの (非線形) 特徴を抽出することを考える.

$$\Phi : \Omega \to H \tag{1}$$

このとき, Φ を特徴写像, H を特徴空間と呼ぶ. (1) が

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

という内積計算を満たすとき, 2つのデータ点 $\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)$ の内積が

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

というカーネル関数 k の値の評価によって計算される. 多くのデータ解析手法において, カーネル値 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ が得られれば十分であることが多く, さまざまな線形のデータ解析手法を変換後のデータに適用することが可能となる.

3　カーネル PCA

3.1　主成分分析 (PCA)

主成分分析 (PCA) とは, 分布の様子をなるべく保持するようにデータを低次元表現するための方法である.

m 次元データ $X_1, \cdots, X_N$ に対し, データを単位ベクトル u に直交射影した $\{u^T X_i\}_{i=1}^N$ の分散が最大となるような方向 u, すなわち

$$u_1 = \arg\max_{||u||=1} \frac{1}{N} \left\{ \sum_{i=1}^{N} u^T \left(X_i - \frac{1}{N} \sum_{j=1}^{N} X_j \right) \right\}^2$$

$$= \arg\max_{||u||=1} u^T V u$$

を第1主軸という. 射影 $u_1^T Xi$ はデータ X_i の第1主成分と呼ばれる. d 次元表現を求める場合には, すでに得られた $u_1, \cdots, u_{p-1}$ と u_p が直交するという条件下で u_p に射影したデータの分散が最大になるように第 p 主軸 u_p を求める. d 次までの主軸は, 分散共分散行列 V の大きいほうから d 個の固有値に対応する単位固有ベクトル $u_1, \cdots, u_d$ により与えられ, データ X_i の第 p 主成分は $u_p^T X_i$ により与えられる.

3.2　カーネル PCA

カーネル PCA とは, 特徴抽出をした空間で通常の主成分分析を行なって, 低次元の線形部分空間を求める非線形な次元削減手法である.

カーネル PCA ではカーネル関数の選び方によって結果がかなり異なってしまうため, 妥当と思える結果を得るためには, もとの空間やデータの構造をうまく反映した適切なカーネル関数を定義する必要がある.

4　カーネル回帰

4.1　線形回帰

線形回帰とは, 数値データの間の関数関係を推定する際に, 線形モデルの式 (2) をあてはめる手法である.

$$y = \boldsymbol{w}^T \boldsymbol{x} = \sum_{m=1}^{d} w_m x_m \tag{2}$$

これは, データに (原点を通る) 直線をあてはめることに相当する. データの直線からのずれを二乗誤差で取り, すべてのサンプルに対する二乗誤差の総和を最小にする w を求めることで関数推定を行う.

4.2　カーネル回帰

カーネル回帰とは, 線形回帰においてカーネル関数を拡張した式 (3) のモデルを用いることで, 非線形関数をあてはめる手法である.

$$y = \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}^{(j)}, \boldsymbol{x}) \tag{3}$$

カーネル関数を使ったモデルにはサンプル数と同じだけの自由度があるため, 過学習になる可能性がある. そこで, 正則化を行い汎化能力を高める必要がある.

5　提案手法

本研究では, カーネル法を用いた手書き数字データの識別器構成を, 以下のように行うことを提案する.

1. カーネル関数として以下のガウスカーネルを使用する.
$$k(x, y) = \exp\left(-\frac{1}{\sigma^2} ||x - y||^2 \right)$$
2. カーネル PCA により 256 次元の手書き数字データを 2 次元に縮約する.
3. 上の 2. で得られた縮約データのクラス間にカーネル回帰により識別境界を引く.

6　実験結果

ここでは, 公開されているアメリカ合衆国郵便公社 (USPS) が業務で得た実際の手書き数字データを使用する. データは 16×16 の 256 画素で構成されており, 画素値は 256 階調の離散値である. このデータを各数字 100 個ずつ用意し, 50 個を識別器構成のための学習用データ, 残りの 50 個を汎化能力を調べるためのテスト用データとして使用する. ただし, データはランダムに抽出している.

6.1 実験 I 0,2,3,5,6 - 1,4,7,8,9 の 2 段階識別

先行研究で行っていたように, 0〜9 を 0,2,3,5,6 と 1,4,7,8,9 の 2 群に分けたあと, それぞれのグループ内で識別を行うという 2 段階での識別方法をとる.

まず, 0〜9 を 2 群に分ける識別結果は以下の図のようになった. 図 1 が学習データの識別結果, 図 2 がその識別器の汎化能力を調べるために行ったテストデータの識別結果である. このとき, 識別率はそれぞれ 89.8%, 86.2% で, その誤差は −3.6pt であった.

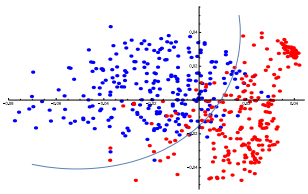

図 1 :学習データ (0,2,3,5,6-1,4,7,8,9)

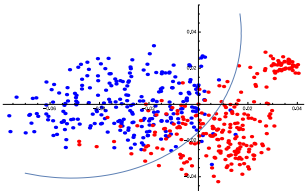

図 2 :テストデータ (0,2,3,5,6-1,4,7,8,9)

次に, 0,2,3,5,6 と 1,4,7,8,9 それぞれのグループ内で識別を行った. 以下の図が識別結果である. 回帰は 2 群での識別となるため, 一度に 5 つに識別することができず, 判別を 4 回行うことによって識別境界を引いた. そのため, 境界が複雑なものになってしまった.

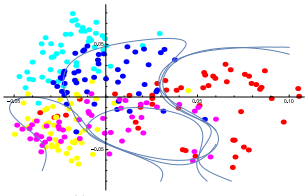

図 3 :学習データ (0,2,3,5,6)

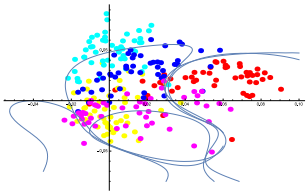

図 4 :テストデータ (0,2,3,5,6)

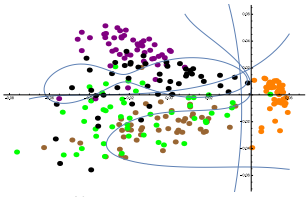

図 5 :学習データ (1,4,7,8,9)

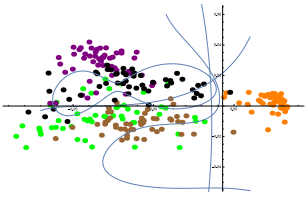

図 6 :テストデータ (1,4,7,8,9)

6.2 実験 II 1,0,6,2,3,8,7,5,4,9 の順に 9 段階で識別

実験　の結果をふまえ, 2 群での判別では識別境界を引くことが可能であったため, 0〜9 を一つずつ取り出し, 9 段階で識別を行う方法をとる. その際, 1,0,6,2,3,8,7,5,4,9 の順に数字を取り出し識別を行ったときが, 学習データの識別率が最も良かったので, この順で識別を行った.

識別率は以下の表 1 に示す. 7 の学習データとテストデータの識別率の誤差が −14% と大きくなってしまった.

表 1 :実験　の識別率

n	学習 (%)	テスト (%)	誤差 (pt)
1	99.8	96.4	− 3.4
0	89.3	82.9	− 6.4
6	93.8	95.8	+ 2.0
2	83.7	91.4	+ 7.7
3	86.6	83.7	− 2.9
8	80.4	80.0	− 0.4
7	83.5	69.5	− 14.0
5	97.3	96.0	− 1.3
4	87.0	86.0	− 1.0
9	87.0	86.0	− 1.0

6.3 実験 III 学習データの識別率 100%

実験　の結果をふまえ, 7 の識別誤差が大きくなる原因として, 識別器の構成に使う学習データの中に, 人の目で見ても読むことができない手書き数字データがあるからではないかと考えた. 例として, 7 を取り出す識別の際に含まれていた 4 の判別が困難なデータを以下の図 7 に示す.

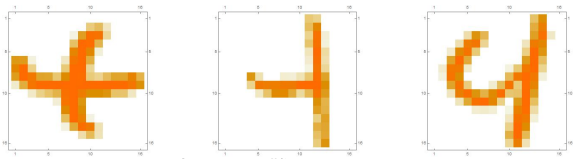

図 7 :判別が困難な 4 データ

そこで, 実験　では学習データの識別率が 100% になるように学習データを抽出し, 実験　と同様の手順で識別を行った. 識別率は以下の表 2 に示す. 結果として 7 の識別誤差は小さくなったが, そのほかの数字での識別誤差が大きくなってしまう結果となった. その原因として, 識別器を読みやすい文字のみを学習して構成してしまったため, テストデータの中に読みにくい文字がある場合にその数字を認識できない可能性があるためと考えられる.

表 2 :実験　の識別率

n	学習 (%)	テスト (%)	誤差 (pt)
1	100.0	99.6	− 0.4
0	100.0	89.1	− 10.9
6	100.0	91.3	− 8.7
2	100.0	86.0	− 14.0
3	100.0	83.3	− 16.7
8	100.0	79.2	− 20.8
7	100.0	94.0	− 6.0
5	100.0	84.7	− 15.3
4	100.0	61.0	− 39.0
9	100.0	61.0	− 39.0

7 まとめと今後の課題

先行研究では, 手書き数字データの識別にカーネル PCA を用いることが有効であると思われるという結果が得られた. 本研究ではその拡張として, カーネル回帰を用いることでデータのクラス間に識別境界を引くことを試みた.

結果として, 2 群の識別では識別境界を引くことができ, 識別器として用いることが有効であると思われる. しかし, 実用化するためには識別誤差が大きすぎるため, より適切なカーネル関数やパラメータを設定することが必要である. また, より良い識別器にするために, どのようなデータを使って学習すれば良いかを考えることも今後の課題である.

参考文献

[1] 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

[2] 赤穂昭太郎, カーネル多変量解析〜非線形データ解析の新しい展開〜, 岩波書店, 2008

[3] 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

[4] 和田萌, カーネル PCA を用いたパターン認識, お茶の水女子大学理学部情報科学科卒業研究, 2013