カーネル主成分分析を用いた判別分析の研究

M2012MM001 荒切彰太
指導教員：松田眞一

1　はじめに

個人の筆跡は，その個人の特徴や癖がある．これまでその特徴や癖を量的データに変換し，統計的手法で個人の筆跡を判別してきた．それらの研究の中で矩形診断法に着目する．矩形診断法による研究の中では，文字種の研究 (古橋ら [6])，漢字情報を用いた研究 (中村 [5])，そして文字の交点を使用した研究がある．このようにいくつかの判別方法があるが，本研究ではサポートベクターマシンで使用されるカーネル法を使用する．カーネル法を用いた主成分分析であるカーネル主成分分析を使用して判別分析を行うことで，線形のモデルでは判別が困難であったデータでもこの方法で判別率が向上するのではないかと考えた．

2　カーネル法とは

カーネル法とカーネル主成分分析の理論は、赤穂 [1] を参照する．カーネル法は，何か特定の一つの分析手法を指すものではない．複雑なデータをカーネル法を用いることで解析しやすくする方法である．カーネル法の例としてはサポートベクターマシンが代表的であるが，他にもカーネル主成分分析やカーネル判別分析，そしてカーネル正準相関分析などがある．カーネル法は線形手法の非線形化を行うことで，線形手法では分からなかったデータの特徴を知ることができる．具体的には，データを特徴写像で高次元の特徴空間上に変換し、もとの空間データではなく，変換した特徴空間上で解析を行う．
このとき，特徴空間上での内積をカーネル関数を用いて計算することによって，計算量を減らすことができる．

2.1　カーネル関数

本研究では，以下の 3 つのカーネル関数を使用した．

- 線形カーネル

$$k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y} \tag{1}$$

- 多項式カーネル

$$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y} + c)^d \tag{2}$$

ただし，d は自然数，$c \geq 0$ である

- ガウスカーネル

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{\sigma}\|\mathbf{y}-\mathbf{x}\|^2\right) \tag{3}$$

ただし，$\sigma > 0$ である．

2.2　カーネル主成分分析

カーネル法を主成分分析に拡張したのがカーネル主成分分析である．高次元の特徴ベクトルに変換してから，通常の主成分分析を行って，低次元の線形部分空間を求める．
なお，本研究では主成分得点を算出し，それを元に判別分析を行った．

3　カーネル主成分分析の R 関数

パッケージ kernlab には，カーネル主成分分析の関数 kpca がある．関数 kpca の書式を次に示す．

$$kpca(x, kernel = "...dot", features = ..., kpar = list(...))$$

引数 x はマトリックス形式のデータである．引数 kernel には用いるカーネル関数を指定する．引数 features では求める主成分の数を指定する．引数 kpar はカーネル関数に用いるパラメータ指定する．結果としては，固有値 eig，主成分ベクトル kpc，主成分得点 rotated などが返される．(金 [3] 参照)

4　判別分析

判別分析とは, サンプルの持っている特性から, そのサンプルが, どの群に属するかを判別する手法である．(菅 [4] 参照) 今回，線形判別関数による判別を行った．

5　交差確認法

交差確認法とは, データセット全体を n 部分に均等に分割し, その中のひとつをテストデータとし, それ以外の $(n-1)$ を学習データとして用いる. データセットを n 部分に分割したときを n 分割交差確認法という．n 分割交差確認法では, 一つのデータセットに対し,n 回のモデルの構築とテスト (確認、検証) を行い, その n 回のテスト結果を全体の評価に用いて, その平均の値を判別率とする．
今回の研究では，奇数行と偶数行に分割する二分割交差確認法を使用した.(赤穂 [1] 参照)

6　筆跡データによる判別

6.1　変数説明

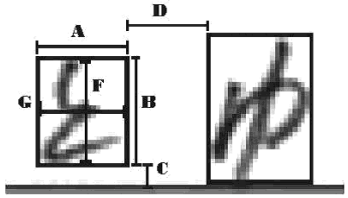

図 1　変数説明

矩形診断をしたとき，漢字情報を用いた筆跡の矩形診断に関する研究 (中村 [3]) を参考に「横の長さ (A)」,「縦の長さ (B)」,「アンダーラインから底辺までの高さ (C)」,「隣り合う文字の間隔 (D)」,「『縦』/『横』の比 (E)」,「交点の縦の比率 (F)」,「交点の横の比率 (G)」を変数とする．

6.2　5 文字の判別率

筆跡データは荒切 [2] のデータを使用する．判別分析，主成分分析を用いた判別分析，カーネル主成分分析を用いた判別分析の結果を次に示す．

A，B，C，D はそれぞれ「あまめみさ」,「みまあめさ」,「ほもとゆね」,「とゆねもほ」を表わす．

全体的には，カーネル主成分分析を用いた判別分析よりも主成分分析を用いた判別分析の方が良い結果が得られた．比較的よい結果となったカーネル関数は，10 次元の多項式カーネル (3 次) であった．どのデータもカーネル関数を変えても，10 次元または 18 次元に次元を減らした方がよい結果となることが分かった．

表 1　5 文字の判別率

分析方法	A	B	C	D
判別分析	91.11	91.11	95.55	93.33
主成分	83.33	96.66	100	97.77
ガウス (18 次元,0.01)	80.00	90.00	71.11	86.66
線形　(18 次元)	83.33	97.77	94.44	88.88
多項式 (18 次元,2 次)	83.33	97.77	94.44	88.88
多項式 (10 次元,3 次)	85.55	93.33	95.55	83.33

6.3　1 文字の判別率

全体的に一文字で判別した場合，判別分析を行った方が良い判別率が得られた．また，線形カーネルと多項式カーネルが主成分分析を用いた判別分析の結果よりも良い結果が得られるものもあった．しかし，ガウスカーネルについては，他の手法よりも劣る結果となった．

7　iris データによる判別の比較

Fisher の研究で使用された iris データを用いて判別を行う．これは判別問題でよく扱われるデータで，統計ソフト R に組み込まれている．iris にはアヤメのがくの長さ (Sepal.Length)，がくの幅 (Sepal.Width)，花弁の長さ (Petal.Length)，花弁の幅 (Petal.Width) の 4 次の特徴ベクトルと，アヤメの種類 (Species) が 3 種類 (setosa,versicolor,virginica) 収録されている．3 種のアヤメはそれぞれ 50 例あり，計 150 例である．判別を行う際，変数を「アヤメのがくの長さ (Sepal.Length)」,「がくの幅 (Sepal.Width)」,「花弁の長さ (Petal.Length)」,「花弁の幅 (Petal.Width)」とし，3 群に分ける．筆跡データと同様に,「判別分析」,「主成分分析を用いた判別分析」,「カーネル主成分分析を用いた判別分析」の判別率を比較する．なお，カーネル主成分分析では，ガウスカーネル，線形カーネル，多項式カーネルを使用する．次元数を 4 次元から 3 次元，2 次元へと主成分分析とカーネル主成分分析のそれぞれで減らし，判別分析を行った結果を次の表に示す．

その結果，4 次元のガウスカーネル (0.01)，3 次元のガウスカーネル (0.001)，3 次元の線形カーネルの場合，97.99 %と判別率が上昇し，判別分析と主成分分析を用いた判別分析の結果よりも高い判別率となった．ガウスカーネル (0.001) と線形カーネルの判別率について，3 次元で行った場合と 4 次元で行った場合を比較すると，97.99 %から 97.33 %へと次元が高くなると減少している．これは，4 次元のときの情報量に判別ができない何らかのノイズが含まれているからだと考えられる．

また 3 次元でも 4 次元と等しい判別率が得られ，これは次元の呪いの回避に繋がる．

表 2　iris データによる判別率

	2 次元	3 次元	4 次元
判別分析			97.33
主成分	91.99	97.33	97.33
ガウス (0.1)	90.66	94.66	97.33
ガウス (0.01)	96.66	97.33	(97.99)
ガウス (0.001)	96.66	(97.99)	97.33
線形	96.66	(97.99)	97.33
多項式 (2 次)	95.99	96.66	97.33
多項式 (3 次)	94.66	96.66	97.33

8　人工データによるカーネル関数の検証

8.1　2 群における人工データの判別

人工データを作成し，3 つのカーネル関数がどのようなデータについて有効であるのかを検証する．人工データは 2 次元で円形にプロットされるデータと，同じく 2 次元で正三角形にプロットされるデータを用意する．なお，これらは一様に乱数を発生させ，内側の図形と外側の図形の 2 群を判別する．これにノイズを加え，カーネル主成分分析を用いた判別分析を行い，それぞれのカーネル関数の有効性を検証する．

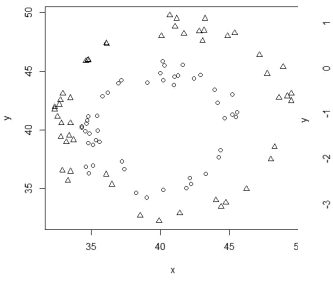

図 2　2 群の円

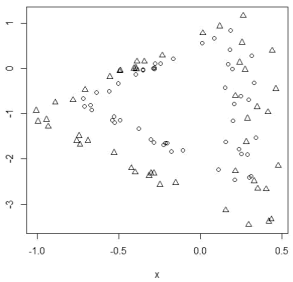

図 3　2 群の三角形

円の人工データと正三角形の人工データは，図 2，3 のように 2 群あると仮定し，データ数はそれぞれ 50 個の計 100 個である．

カーネル関数による判別結果が表 3 である．円の人工データでの判別では，ガウスカーネルと多項式カーネルのパラメータはそれぞれ sigma=0.3,degree=2 である．そして，正三角形の人工データでの判別では，ガウスカーネルと多項式カーネルのパラメータはそれぞれ sigma=0.3,degree=4 である．

円の人工データも正三角形の人工データでもガウスカーネルと多項式カーネルを使用すれば，他の手法よりも判別率が良くなる結果となった．線形カーネルは主成分分析を用いた判別分析と似たような結果が得られるため，判別率の上昇は見られなかった．

表 3　2 群データでの判別率

	円	正三角形
判別分析	53.00	53.00
主成分	53.00	53.00
ガウス	(71.00)	58.00
線形	53.00	53.00
多項式	55.00	(61.00)

9　3 群における人工データの判別

前節の結果からそれぞれのカーネル関数の有効性が分かったが，それが多群の場合でもいえるのかを検証したい．そこで，先ほどと同様にデータを作成し，2 群ではなく 3 群のデータで判別を行う．

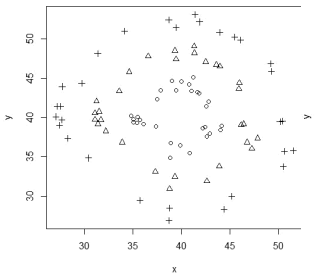

図 4　3 群の円

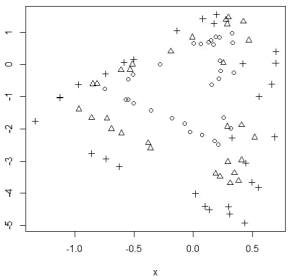

図 5　3 群の三角形

円の人工データと正三角形の人工データは，図 4，5 のように 3 群あると仮定し，データ数はそれぞれ 30 個の計 90 個である．

表 4　3 群データでの判別率

	円	正三角形
判別分析	35.55	39.99
主成分	35.55	39.99
ガウス	(61.11)	35.55
線形	35.55	39.99
多項式	35.55	(45.55)

各方法での判別結果が表 4 である．円の人工データでの判別では，ガウスカーネルと多項式カーネルのパラメータはそれぞれ sigma=0.15,degree=2 である．そして，正三角形の人工データでの判別では，ガウスカーネルと多項式カーネルのパラメータはそれぞれ sigma=0.3,degree=3 である．

3 群の場合でも 2 群と同じ結果が得られた．よって，3 群でも同様な特徴がみられることから，多群の場合でも円形のように丸くプロットされるようなデータではガウスカーネルが有効で，正三角形のように角のあるようなデータでは多項式カーネルが有効であるといえる．

10　筆跡データの適用法の模索

10.1　標準化した筆跡データによるカーネル関数の傾向

カーネル関数は標準化した筆跡データではどのような結果を得られるのかを検証する．なお，カーネル関数のパラメータはそれぞれ最適なパラメータを設定した．次元数は 18 次元に次元削減している．

その結果，判別分析と主成分分析を用いた判別分析では，標準化した筆跡データと標準化しなかった筆跡データによる違いはなかった．カーネル関数の場合，標準化しても良い結果は得られず，標準化しないで判別した方が良いことが分かった．

11　筆跡データ（2 次元）による判別

文字の形 (縦長の字，横長の字，縦と横の長さが同じ字) でカーネル関数の特徴に違いはあるのかを検証する．そこで筆跡データの「縦」と「横」の 2 次元のみで判別を行う．解析対象は「ほもとゆね」の「も」(縦長の字)，そして「みまあめさ」の「み」(横長の字),「め」(縦と横の長さが同じ字) である．

表 5　「ほもとゆね」の「も」による判別率

	標準化なし	標準化あり
判別分析	47.77	47.77
主成分	47.77	47.77
ガウス (0.1,0.3)	34.44	45.55
線形	47.77	47.77
多項式 (2 次,2 次)	48.88	38.88

表 6　「みまあめさ」の「み」による判別率

	標準化なし	標準化あり
判別分析	37.77	37.77
主成分	37.77	37.77
ガウス (0.01,0.01)	31.11	36.66
線形	37.77	37.77
多項式 (2 次,2 次)	36.66	22.22

表 7　「みまあめさ」の「め」による判別率

	標準化なし	標準化あり
判別分析	44.44	44.44
主成分	44.44	44.44
ガウス (0.01,0.01)	39.99	44.44
線形	44.44	44.44
多項式 (3 次,3 次)	51.11	33.33

表 5 から表 7 より判別分析，主成分分析を用いた判別分析，そして線形カーネルの判別率がどの文字も等しい結果が得られた．ガウスカーネルについては標準化した方が判別率が高くなる結果が得られた．多項式カーネルは「も」と「め」の文字が標準化しなかった場合，どの手法よりも高い判別率が得られた．しかし,「み」は多項式カーネルが他の手法よりもわずかに劣る結果となった．

ここで「みまあめさ」の「み」は文字の先頭に当たるが,「ほもとゆね」の「も」,「みまあめさ」の「め」は文字列の中央に入ることに気がついた．人が文字を書く上で一文字目の文字は不安定な文字であるのかと疑問に思った．そこで，先頭に当たる「あまめみさ」の「あ」,「ほもとゆね」の「ほ」,「とゆねもほ」の「と」，そして多項式カーネルが劣る結果となった文字と同じ「あまめみさ」の「み」を追加して比較する．

表 8 から表 11 より先頭に当たる文字の判別率をみると「あ」,「ほ」は多項式カーネルが他の手法よりも高く,「と」,「み」は他の手法よりも低い結果となった．しかし，先頭ではない字のすべてが多項式カーネルが最も良い結果となった．また，先頭の文字と先頭ではない字の判別率を比較すると先頭ではない字の方が全体的に高い判別率だということが分かる．よって，人が文字を書く上で，一文字目の文字は不安定な文字である可能性が高いといえる．

表 8 「ほもとゆね」の「ほ」による判別率

	標準化なし	標準化あり
判別分析	39.99	39.99
主成分	39.99	39.99
ガウス (0.01,0.01)	35.55	39.99
線形	39.99	39.99
多項式 (3 次,3 次)	42.22	36.66

表 9 「とゆねもほ」の「と」による判別率

	標準化なし	標準化あり
判別分析	32.22	32.22
主成分	32.22	32.22
ガウス (0.01,0.01)	26.66	32.22
線形	32.22	32.22
多項式 (2 次,2 次)	31.11	32.22

表 10 「あまめみさ」の「あ」による判別率

	標準化なし	標準化あり
判別分析	36.66	36.66
主成分	36.66	36.66
ガウス (0.01,0.01)	33.33	36.66
線形	36.66	36.66
多項式 (3 次,2 次)	38.88	18.88

表 11 「あまめみさ」の「み」による判別率

	標準化なし	標準化あり
判別分析	45.55	45.55
主成分	45.55	45.55
ガウス (0.015,0.01)	34.44	45.55
線形	45.55	45.55
多項式 (2 次,3 次)	46.66	37.77

12 まとめ

カーネル主成分分析を用いた判別分析では，5 文字による筆跡データでの判別は向かない結果となった．しかし，iris データでは，ガウスカーネルや線形カーネルが，他の手法よりも高い判別率となり，データによっては判別率を上げることが可能だということが分かった．

円形のように丸くプロットされるようなデータのときガウスカーネルが有効で，正三角形のように角のあるようなデータのときには，多項式カーネルが有効であると分かった．線形カーネルも実データの結果から，主成分分析を用いた判別分析と同じような結果が得られた．

筆跡データに関して適用法を探ったが，標準化せずに判別した方が良いことが分かった．また，変数を「縦」と「横」のみを使用することで多項式カーネルが有効であることが示せた．新しい発見として一文字目は判別するのに不安定なデータである可能性も見つけられた．

13 おわりに

研究を終えて，カーネル主成分分析を用いた判別分析ではまだまだ工夫が必要であると感じた．例えば，次元削減でどのような基準でどこまで減らすかや，パラメータの最適な設定の仕方などである．今後の課題としては，上記で述べたことも含め，筆跡データだけでなく次元数の多い実データでの検証や他のカーネル関数の特徴の検証などが考えられる．

参考文献

[1] 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.

[2] 荒切彰太：筆跡の交点を利用した矩形診断に関する統計的分析, 南山大学数理情報学部情報システム数理学科卒業論文要旨集,2012.

[3] 金明哲：R によるデータサイエンス, データ解析の基礎から最新手法まで, 森北出版株式会社,2007.

[4] 菅 民郎：初心者がらくらく読める, 多変量解析の実践, 上, 現代数学社,2000.

[5] 中村元樹：漢字情報を用いた筆跡の矩形診断に関する研究, 南山大学数理情報学部数理科学科卒業論文要旨集, 2006.

[6] 古橋あい・長谷川千津・伊藤志麻・浦末直樹：統計的解析による筆跡鑑定, 南山大学経営学部情報管理学科卒業論文要旨集,1997.