カーネル PCA による漢字部首の手書き識別

鈴木 日菜 (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時，観測されたデータをそれらの概念のうち一つに対応させることである．

パターン認識において，正当化された情報から特徴抽出を行う際に，何が重要かは明示的にはわかりづらい．そこで PCA (主成分分析) などの統計的な情報圧縮手法を用いる．それにより，情報を次元圧縮された特徴ベクトルへと変換して認識に必要な情報を抽出することが可能となる．

本研究では，手書き漢字の部首において，カーネル PCA を用いて，識別器を構成し，書き手の識別の可能性についての研究を行った．

2　先行研究

研究 [1] においては，6 人分のひらがな文字「す」「そ」「み」「さ」の文字データを集め，各文字の特徴点 (9 or 10 点) を取り，それぞれの座標を取得し，そのデータにカーネル PCA を施して，6 人のひらがな文字の識別を行なっている．

本研究では，これを漢字に応用し，漢字の部首のみで，文字の特徴点間の情報を用いた書き手の識別を行う．

3　カーネル PCA

3.1　PCA(主成分分析)

PCA とは，分散の大きい方向にデータを射影することで，多次元データの情報を，特性を保ちながらより低い次元に縮約させる方法である．

しかし，スケーリングフリーな識別ができる一方，線形データ解析手法のため，非線形なデータの構造が捉えにくいという欠点がある．

3.2　カーネル法

一般にカーネル法では，非線形変換を介して，データ x のいろいろな特徴量を取り出している．

$\phi_1, \ldots, \phi_d$ を特徴抽出するための非線形関数とし, 特徴ベクトルを $\overrightarrow{\boldsymbol{\phi}(x)} = (\phi_1(x), \ldots, \phi_d(x))^T$ と書く．このとき，カーネル関数は特徴抽出の内積に基づき，以下のように定義できる．

$$k(x, x') = \overrightarrow{\boldsymbol{\phi}(x)}^T \overrightarrow{\boldsymbol{\phi}(x')} = \sum_{m=1}^{d} \phi_m(x)\phi_m(x')$$

したがって，非線形に写像した空間での $\vec{\phi}(x)$ と $\vec{\phi}(x')$ の内積が，入力特徴 x と x' のみで計算でき，$k(x, x')$ から最適な非線形な非線形写像を構成することができる．このような関数 k をカーネルと呼び，このように高次元に写像しながらカーネルの計算のみで最適な識別関数を構成することを，一般に，カーネルトリックという．本研究では, *Gauss* カーネルを用いる．

$$k(x, x') = exp(-\beta \|x - x'\|^2)$$

パラメータは β とし, 本研究ではカーネルパラメータと呼ぶ.

3.3　カーネル PCA

カーネル PCA とは，高次元の特徴ベクトルに変換してから，通常の PCA を行い，低次元の線形部分空間を求める多変量解析手法である．以下がアルゴリズムである．

i 中心化されたデータ点の集合 $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$ から，グラム行列 $K = (k(x^{(i)}, x^{(j)}))$ を作る．

ii $K\alpha = \lambda\alpha$ という固有値問題を解く．

iii 上から M 個の固有値 $\lambda_1, \cdots, \lambda_M$, 固有値ベクトル $\alpha_1, \cdots, \alpha_M$ を用いて，M 次元 PCA プロットを行う．

本研究においては，M を 2 として，第 2 主成分まで用いて 2 次元 PCA プロットを行った．

4　提案方法

i 手書き漢字文字の書かれた紙をスキャンする．

ii 各文字の部首 (サンズイ, シンニョウ) の特徴点 (6 or 10 点), それぞれの座標を取得する．

iii 学習データにカーネル PCA を施し，いろいろな数値のカーネルパラメータ β を調べ、適切な数値を確定する．

iv カーネルパラメータ β を用いて，テストデータにカーネル PCA を施す.

5　文字データ

実データとして，4 人分の漢字文字「**海**」「**池**」「**決**」「**江**」「**返**」「**辺**」「**連**」「**述**」の文字データを用いる．

一人につき，各文字 10 字ずつ，80 文字のデータを集めた．サンズイ (「**海**」「**池**」「**決**」「**江**」) は各 6 点，シンニョウ (「**返**」「**辺**」「**連**」「**述**」) は各 10 点の特徴点を取り，それぞれ，サンズイに対して計 960 点，シンニョウに対して計 1600 点の座標データを用いる．「**海**」「**池**」「**決**」「**江**」「**返**」「**辺**」「**連**」「**述**」のうち，「**海**」「**池**」「**決**」「**返**」「**辺**」「**連**」を学習データ,「**江**」「**述**」をテストデータとする．

特徴点の取り方


6　実験概要

6.1　実験 1

目的　「**海**」「**池**」「**決**」「**江**」のサンズイ，「**返**」「**辺**」「**連**」「**述**」のシンニョウについて書き手の識別を目指して, カーネルパラメータ β を調整 (学習) する．

手法　学習データを用いて, いろいろな数値のカーネルパラメータ β を当てはめ, 図からカーネルパラメータ β を決定する. 決定したカーネルパラメータ β を用いて, テストデータにカーネル PCA を施す.

予想　サンズイの漢字では, 部首の特徴点が少ないため, 書き手の識別を行うことができないが, シンニョウでは特徴点が多いため, 書き手の識別を部首のみで行うことが可能.

結果　サンズイでは, 最適なカーネルパラメータ β を見つけることはできなかった (参照 図 1). シンニョウでは, 学習データで書き手の識別を行うことが可能 (参照 図 2) であった. 図 1 で決定したカーネルパラメータ β を用いて, テストデータにおける漢字の書き手の識別を行うことができた (参照 図 3).

考察　サンズイでは不可能だったが, シンニョウではカーネル PCA を用いて 4 人の文字を識別することは可能である. カーネル PCA では, 人間の目では見つけることの難しい特徴を用いて, 識別を行なっている.

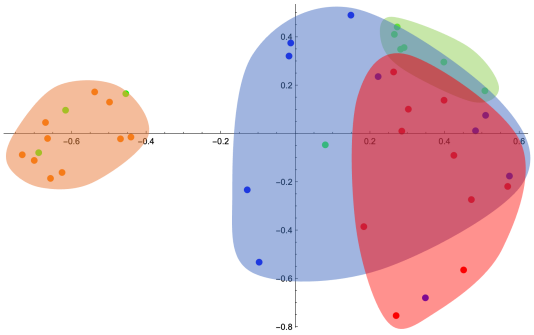

図 1 : サンズイ

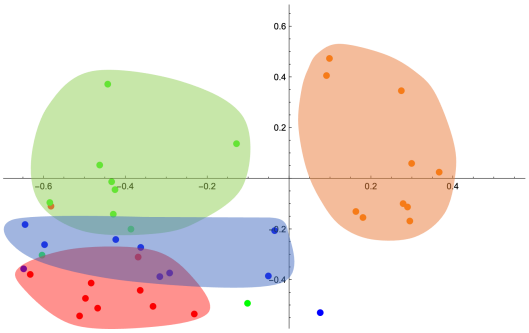

図 2 : シンニョウ (学習データ)

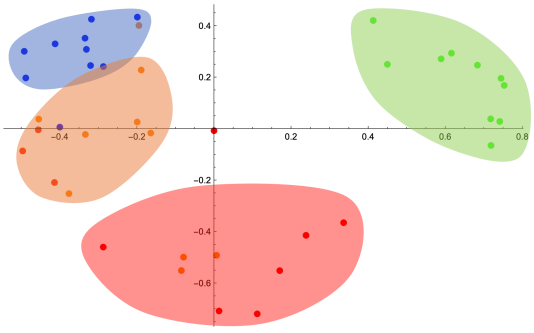

図 3 : シンニョウ (テストデータ)

6.2　実験 2

目的　「返」「辺」「連」「述」のシンニョウについて, カーネル PCA を用いて書き手の識別を行う際の特徴点を調べる.

手法　識別に用いる座標のデータの数を, 4 人が識別できる限界まで減らしていく.

予想　特徴点を 2,3 点減らしても, 書き手の識別を行うことができる.

結果　シンニョウの識別において, 10 点全てを用いれば書き手の識別を行うことが可能 (参照 図 2) であったが, 1 点でも識別に用いる座標データを減らす (参照 図 4) と, 書き手の識別を行うことが難しかった.

考察　人間には一見, 識別に必要がなさそうな点に見えても, 識別に必要な点が複数存在する. 識別を行う際には, 特徴が表れると考えられる点を全て取り, そこから一つずつ減らしていくことで, 様々な文字で識別を行うことが可能になると考えられる. 今回はサンズイにおいて 6 点を特徴点として抽出したが, 特徴点を増やすことで, 識別を行うことが可能になる可能性があると考えられる.

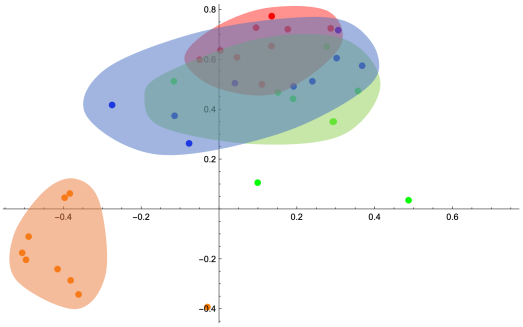

図 4 : シンニョウ (9 点) の例

7　本研究のまとめと今後の課題

カーネル PCA は, 手書き漢字文字の部首のみでのパターン識別の手法の一つとして有効であると考えられる. しかし, データを取る際に, 一度に 10 文字ずつ書いてもらったため, 途中から文字が崩れているように見えた. 今後はデータを取る日を細かく分けることで, 字崩れが起きず, より書き癖が正確に表れるように工夫したい.

今回はデータを 4 人に限定して行ったが, 人数を変えても結果が同様になるのか, についても検証していく必要がある. 今回はデータの特徴点の抽出を手作業で行っており, 本来の点との誤差が生じていると考えられるため, それらの誤差をどうなくしていくか, がより正確に研究を行うために考えるべき課題である.

参考文献

[1] 坂井佳帆 「カーネル PCA の手書き平仮名識別への応用」 (2018)

[2] 漢字データ https://kakijun.jp/m-s/