H-049

k-部分空間法
k-Subspace method

藤井 康仁† 坂野 鋭‡
Yasuhito Fujii Hitosi Sakano

1. まえがき

本研究では部分空間法に基づくクラスタリング法 k-部分空間法，および，クラスごとに複数の部分空間を用いたパターン認識法である k-部分空間識別法を提案し，認識実験を通じて有効性を示す．

部分空間法は簡便で強力なパターン認識法であり，文字認識，顔認識など様々な分野で実用に供されている[1]．ただし，部分空間法は同じクラスのパターンであれば特徴間に同様な相関があることを想定している．

一方で，手書き文字に見られる異字体の様に，同じクラスに属するにも関わらず全く形状が異なる例がある．このような場合には一つのクラスについて複数の部分空間，つまりサブクラスの部分空間を用いることが自然である．

従来，k-平均法などのクラスタリング法はセントロイドと呼ばれる特徴空間上の点を基礎として構成されている．しかし，k-平均法は部分空間に関して，何ら拘束を持たないため，同じ部分空間に属するサブクラスに分割してしまうことは容易に想像される．このため，部分空間が異なるサブクラスになるようにデータを分割できる方法が必要である．

この問題を解決するために本研究では部分空間に基づくクラスタリング法である k-部分空間法，および，それに基づくパターン認識方法である，k-部分空間識別法を提案し，文字認識実験を通して有効性を検証する．

2. k-部分空間法

k-平均法はセントロイドを基礎としてクラスタリングを行うが，k-部分空間法ではセントロイドの代わりに部分空間を基礎としてクラスタリングを行う．

まずクラス c に属する学習データを k 組のサブセットに分割し，各サブセットにおいて部分空間を作成する．次に全学習データを各部分空間射影し，射影が最大となる部分空間が属するサブクラスに所属を変える．こうしてできた新たなサブセットで再び部分空間を構成し，全学習データを部分空間に射影する．

この操作を設定した収束条件を満たすまで繰り返すことによって，k 個のサブセットに対応した部分空間が作成される．

本アルゴリズムを 図 1 に示す．

3. 実験

実験には，手書き数字データセットである MNIST を使用した．MNIST は 0-9 の手書き数字 70000 枚で構成されており，各画像はピクセル数 784，輝度値 0-255 の 256 段階である．実験では 学習データ 60000 枚で学習を行い，テストデータ 10000 枚の識別精度によって評価を行った． 最初の分割では k-平均法 アルゴリズムで分割を 行い，収束条件は，m 回目に作られた部分空間に 存在していたサンプルが，m + 1 回目のある部分空 間に閾値 90%以上存在していることとした． 実験におけるパラメータとして，クラスタ数 k を 1 − 5, 部分空間を張る固有ベクトル数を 1 − 60 の範囲で変化させて精度 を測定した．実験結果を表 1，図 1 に示す．k=2- 5 のとき，全ての次元において k=1(つまり従来の CLAFIC) と比較し，識別率が向上した．特に k = 3，次元数 40 では約 1.8% の精度向上を達成した．

Algorithm 1 部分空間法に基づくクラスタリングアルゴリズム

Require: クラス c の全学習データ X, クラスター数 k, 固有ベクトルの数 n, 閾値 ϵ
Ensure: クラス c の k-部分空間
1: X を k 組のサブセット S に分割
2: **repeat**
3: **for all** S_i **in** S **do**
4: S_i の自己相関行列から固有値，固有ベクトルを計算
5: $P_i :=$ 固有値の大きさ上位 n 個の固有ベクトルを基底とした部分空間
6: **end for**
7: X を各部分空間に射影し，ノルムが最大となる空間へクラスタリング
S をクラスタリングされた集合で更新
8: 残存率 $r :=$ クラスタリング前の学習データが元の空間に留まった割合
9: **until** 収束条件: $r < \epsilon$
10: **return** k 個の部分空間

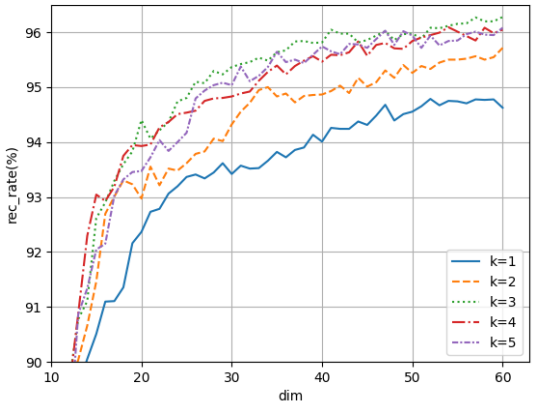

図 2 次元数(横軸)と識別精度(縦軸：%)

†島根大学総合理工学部知能情報デザイン学科
Interdisciplinary Faculty of Science and Engineering, Department of Information Systems Design and Data Science, Shimane University
‡島根大学大学院自然科学研究科 Graduate School of Natural Science and Technology, Shimane University

4. まとめと今後の課題

本研究で部分空間に基づくクラスタリング法である k-部分空間法及び，それに基づくパターン認識法である k-部分空間識別法を提案し実験的に有効性を示した．

手書き数字画像の識別において，各クラスに対して，複数の部分空間を用いることにより，単一の部分空間を用いるよりも高い識別精度が得られることを確認した．

しかしながら，k-部分空間法にはいくつか課題がある．まず，k-平均法と同様に初期値依存性が高いため，最初の学習データの分割方法を検討する必要がある．次に，最適な部分空間の数をどのように決定するかである．k=2-5 のときの識別率は k=1 よりも向上したが，実験結果のとおり k の値を大きくすれば識別制度が向上するわけではない．また，収束条件の設定方法も課題である．今回は m 回目に作られた部分空間に存在していたサンプルが m + 1 回目のある部分空間に存在する割合を収束条件としたが，閾値の値を増加させると部分空間が収束しなくなる状況に陥るため，別の収束条件を設定することを考慮する必要があると考える．

また，大きな問題としては k-平均法が EM アルゴリズムによる最尤推定であるという理論的な根拠が明確であるのに対し[3]，k-部分空間法にはそのような理論的根拠が無い．これを明確にしなくては，k-部分空間識別法が強力である理由が必ずしも明確にならない．

今後の課題として以上の問題を解決し，最適な k-部分空間法アルゴリズムを作成することを目指す．

次元数	40	50	60
k=1	94.00	94.55	94.62
k=2	94.86 (+0.86)	95.25 (+0.70)	95.71 (+1.09)
k=3	95.81 (+1.81)	95.97 (+1.42)	96.27 (+1.65)
k=4	95.46 (+1.46)	95.83 (+1.28)	96.04 (+1.42)
k=5	95.74 (+1.74)	95.93 (+1.38)	96.08 (+1.46)

表 1　識別率 (%)

参考文献

[1] 坂野 鋭, “部分空間法の最近の発展”, 電子情報通信学会誌, Vol.95, No.4 (2012).

[2] 石井 健一郎 ， 上田 修功 , 続わかりやすいパターン認識, オーム社, (2014)