2D-03

混合部分的正規分布の線形結合による手書き文字特徴量の分布推定に関する検討

鈴木雅人† 北越大輔† 松本章代‡
†東京工業高等専門学校情報工学科 ‡東北学院大学教養学部

1 はじめに

部分的正規分布および歪度成分分析を用いたマハラノビス距離 [1] は，手書き文字品質の低下に頑健な識別関数である．特徴量分布の非正規性を部分的正規分布を用いて吸収するのが 1 つの特徴であるが，歪みの大きい分布や多峰性を有する分布に十分対応することは困難である．著者らはこの問題に対応するため，部分的正規分布を線形結合した混合部分的正規分布によって分布の近似を行う手法 [2] を検討したが，正規分布の重ね合わせによる手法 [3] に比べて計算量が爆発的に増大するという問題を抱えている．そこで本稿では，歪度の大きな軸に対して変数変換を適用し，混合部分的正規分布をあてはめる軸の数をおさえることにより，識別精度を維持しつつ計算量を削減する手法について検討する．

2 識別関数の改良

低品質手書き文字認識において，d 次の特徴量に対する主成分分析結果のうち，中間層の主成分の歪度が認識精度に大きな影響を与えていることがわかっている．提案するアルゴリズムでは，これらの中間層の主成分に対して歪度成分分析 [1] を行い，得られた成分軸に対して変数変換を行う．そして，その中から正規分布で近似できない成分軸に対してのみ独立成分分析 (ICA)[4] を適用して混合部分的正規分布をあてはめ，識別関数を設計する．

以下では，混合部分的正規分布のあてはめ方法および，変数変換による計算量の削減方法について述べる．

2.1 混合部分的正規分布のあてはめ

特徴量の主成分分析結果のうち，中間層主成分に対して歪度成分分析を適用し，歪度の大きい成分軸 $x_1, \ldots, x_n$ を抽出する．一般に，特徴量の x_k 軸成分 $(k = 1, 2, \ldots, n)$ は，単に歪度が大きい偏った分布になっているだけでなく，多峰性を有するなど複雑な分布になっていることが多い．このような複雑な分布を表す確率密度関数を同定するのは困難であるため，ここでは単純な確率密度関数の線形結合によって複雑な分布を近似的に表現する方法を考える．特徴成分の混合モデルを式 (1) のように表すと，この問題は ICA を用いて解くことができるから，多峰性を有する特徴成分 x_k を，単峰性をなす特徴成分 $s_1, \ldots, s_n$ の線形結合で表すことができる．

$$\begin{aligned} &\vec{s} = A^{-1}\vec{x} \\ &\vec{s} = (s_1, \ldots, s_n)^t, \vec{x} = (x_1, \ldots, x_n)^t \end{aligned} \tag{1}$$

一方，特徴成分 s_j は単峰性をなす分布に従うが，歪度の大きな特徴成分を扱っているため，正規分布に歪みを加えた分布によって近似する必要がある．本稿では，識別関数設計の容易性などを考慮して，分散の異なる 2 つの正規分布を境界 $x = m$ でつなぎ合わせた部分的正規分布 (Asymmetric Partial Normal Distribution)[1] をあてはめる．図 1 のような部分的正規分布を $A(m, \sigma_1, \sigma_2)$ で表すものとすると，その確率密度関数 $p(x)$ は式 (2) のようになる．

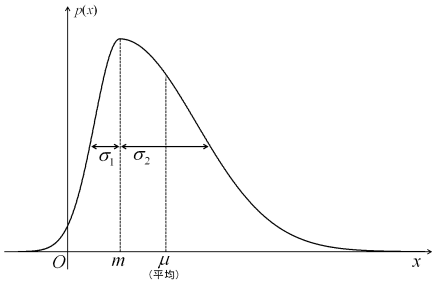

図 1: 部分的正規分布の確率密度関数

$$p(x) = \begin{cases} \kappa \exp\left(-\dfrac{(x-m)^2}{2\sigma_1^2}\right) & (x \le m) \\ \kappa \exp\left(-\dfrac{(x-m)^2}{2\sigma_2^2}\right) & (m \le x) \end{cases} \tag{2}$$

ただし，$\kappa = \dfrac{2}{\sqrt{2\pi}(\sigma_1 + \sigma_2)}$

3 つのパラメータ m,σ_1,σ_2 は, 平均・分散・歪度の関数として表すことができる．従って，与えられた学習デー

A Study of feature estimation method of handprinted character recognition using mixtures of partial normal distribution
†Masato SUZUKI †Daisuke KITAKOSHI ‡Akiyo MATSUMOTO
†Department of Computer Science, Tokyo National College of Technology
‡Faculty of Liveral Arts. Tohoku Gakuin University

タから標本平均，不偏分散，不偏歪度を計算し，それらを母数の推定量と考えて非線形連立方程式を解くことにより，パラメータを推定することができる．

本アルゴリズムを考慮したマハラノビス距離は式 (3) によって与えられる．

$$d(\vec{x}) = \sum_{k \notin \chi} \frac{\{(\vec{x} - \vec{m})^t \vec{e_k}\}^2}{\lambda_k} + \sum_{k \in \chi} \phi_k(\vec{x}) \quad (3)$$

ここに，χ は $x_1, \ldots, x_n$ のうち，本アルゴリズム適用軸番号の集合であり，ϕ_k は対象軸における距離である．ただし，対象軸 x_k では確率密度が多峰性を示すため，対象軸における学習データの平均を μ_k, 分散を σ_k^2 とするとき，

$$\left| \int_{\mu_k}^{x} q_k(x) dx \right| = \alpha \sigma_k \quad (4)$$

となる α を求め，α^2 を距離として使用する．

2.2 計算量の削減方法

一般に，数え上げによって得られる正数のみからなる特徴量の分布は，ガンマ分布で近似できることが多いと言われている．この性質は，主成分分析や歪度成分分析によって得られた軸に関しても同じである．このような分布を正規分布に近づける方法として，次のような変数変換を行う方法が知られている．

$$y = x^\beta \qquad (0 < \beta < 1) \quad (5)$$

そこで，歪度成分分析によって得られた各成分軸に対し，もっとも正規分布に近くなるパラメータ β を求め，変数変換後の分布が正規分布と見なせるかどうかを χ^2 検定によって判定する．正規分布と見なせる成分軸に混合部分的正規分布のあてはめを行わないことで，計算量を大幅に削減することが可能となる．

3 性能評価実験

提案手法による効果を検証するために，平仮名および教育漢字 1052 字種を対象とし，低品質文字データ 200 セット (学習用データ 180 セット，認識用データ 20 セット) を用いて認識実験を行った．表 1 の実験結果によると，歪度成分分析で得られた 15 軸に対して多峰性を検知して混合部分的正規分布をあてはめた軸は 2837 軸 (17.98%)，部分的正規分布をあてはめた軸は 12943 軸 (82.02%) であった．一方，式 (5) の β を 0.2 間隔で変化させて変数変換を行い，正規分布の適合度を有意水準 95%の χ^2 検定で調査したところ，多峰性を有する 2837 軸のうち 785 軸が正規分布で近似可能であり，混合部分的正規分布のあてはめに要する計算量は，約 27.67%を削減できた．同様に，多峰性のない 12943 軸のうち 4714 軸が正規分布で近似可能であり，部分的正規分布のあてはめに要する計算量は，約 36.42%を削減できた．また，従来の手法での認識精度は 84.46%であったのに対し，提案手法では誤認識文字が 6 文字増えて，認識精度は 84.43%とわずかに低下した．以上のことから，今回の検討により，認識精度をほぼ一定に保ったまま，計算量を大幅に削減できたといえる．

表 1: 多峰性が検出される軸数

	変数変換前	変数変換後
混合部分的正規分布あてはめ	2837 (17.98%)	2052 (13.00%)
部分的正規分布あてはめ	12943 (82.02%)	8229 (52.15%)

4 まとめ

歪度成分分析および混合部分的正規分布のあてはめによる低品質手書き文字の識別関数設計において，多峰性を有する成分軸にあてはめる混合部分的正規分布の計算は，多大な時間を要するため，実用化の観点から大幅な計算時間の削減が必要である．そのため本稿では，歪度成分分析によって得られた各成分軸に対して変数変換を行って正規分布への近似を試み，近似が難しい成分軸に対してのみ混合部分的正規分布をあてはめる方法を検討した．その結果，認識精度をほとんど低下させることなく，計算時間を大幅に削減することを実験により確認することができた．尚，本研究の一部は科学研究費補助金 (基盤研究 (C) 課題番号 22500170) の助成によるものである．

参考文献

[1] 鈴木，北越，松本，“歪度最大基準に基づく特徴選択法による低品質手書き文字認識法の検討”, 信学技報 PRMU2012-110, pp.251-256, Jan. 2013.

[2] 鈴木，北越，松本，“混合部分的正規分布による手書き文字識別関数設計に関する検討”, 2014 年信学全大，D-12-22, Mar, 2014.

[3] M.E. Tipping, et.al. “Mixtures of probabilistic principal component analyzers”, Neural Computation, vol.11, no.2, pp.443-482, 1999.

[4] Aapo Hyvärinen, “独立成分分析”, 東京電機大学出版局，2005.