手書き文字認識における大分類のための決定木生成法*

4H－10

篠沢 佳久†　　大駒 誠一‡

慶應義塾大学 理工学部 管理工学科§

1 はじめに

大規模な文字認識を行なう場合まず始めに文字パターンをいくつかのグループに分類し，そして次に各グループ内で詳細に認識するという２段階による認識方法が一般的となっている．しかし大分類部において一回の処理で文字パターンをいくつものグループに分けることは複雑な分類問題となり結果的に分類率の低下につながる．そこで本研究では大分類部において複数の特徴の中から文字パターン群にとって最適な特徴を選択し，逐次分類していく決定木(分類木)の生成法を提案する．

2 文字パターンの分類

2.1 文字パターンのグループ分け

本研究ではまず文字パターンのグループ分けを決めて，次にそのグループ分けに従って文字パターンを分類できる識別機械を作成するという二段階の方法を用いた.[1] まず文字パターンを**Kohonenの自己組織化** [2] によって任意の特徴を用いて任意個のグループに分割する．ところが手書き文字のような変形の激しいパターンにおいては同じ文字が同じグループに分割されるとは限らない．そこで明確に分類できている文字だけを分類できたと判定する．判定方法は，

$$\frac{\text{最も多く分割された文字パターン数}}{\text{学習に使用した文字パターン数}} \geq \beta \qquad (1)$$

とする．βは分類判定のための係数であり，**判定係数**と呼ぶ．(1) 式を満たさない文字パターンは今回使用した特徴では分類できないと判断し，すべてのグループに属するものとする．

2.2 識別面の再学習

Kohonenの自己組織化による分割方法では同じ文字がすべて同じグループに分類できるわけではない．同じ文字でも違うグループとして分割されてしまうことがある．そのため前述の方法でグループ決めしたように，文字パターンを分類できる識別機械を新たに構築しなければならない．これはニューラルネットワークの一つである**バックプロパゲーションネットワーク** [3] を用いて作成した．(1) 式を満たした文字パターンだけを対象としてバックプロパゲージョンアルゴリズム (BP) によって学習を行なう．

3 特徴選択のための指標

逐次文字パターンを分類し決定木を作成していく際に，Kohonenの自己組織化によって文字パターンを同一の特徴を用いて分割し続けていくと同じ特徴を持った文字パターンが同一グループに集まることになり徐々に分類しづらくなる傾向がある．またこれまで多くの研究者によって文字認識で使用する特徴が提案されてきたが，分類すべき文字パターン群が変わった場合，その文字パターン群に有効な特徴も当然のことながら変わる．そのため各層ごとで使用する特徴を変えなければならないのだが，どの特徴をどういった基準で選択するかという問題が生じる．

1. 分類度

$$\sum_{i=1}^{P} X_i \times \frac{\text{文字パターン}\,i\,\text{における正当数}}{\text{使用した文字パターン}\,\mathrm{i}\,\text{の個数}}$$

$$X_i = \begin{cases} 1 & \text{(グループを決定した場合)} \\ 0 & \text{(未判定とした場合)} \end{cases}$$

2. バランス

$$1 - \frac{\text{分割したグループ中の最大カテゴリー数}}{\text{学習に使用した文字パターン数}} + 1 - \frac{\text{最大カテゴリー数} - \text{最小カテゴリー数}}{\text{学習に使用した文字パターン数}}$$

3. 線形分離度

$$\frac{\text{(1) 式を満たした文字パターン数}}{\text{学習に使用した文字パターン数}}$$

文字パターン群の分割を評価するのに分類度，バランス，線形分離度，以上３つの指標を導入した．こ

*The Construction of Decision Tree for Rough Classification in Handwritten Character Recognition.

†Yoshihisa SHINOZAWA
‡Seiichi OKOMA
§Faculty of Science and Technology , Keio University

れらの指標は複雑に関連し合っているので実際の評価にはこれらの指標を統合化した評価指標（**統合指標**）を使う．

統合指標 = 分類度 + 線形分離度 + バランス　(2)

この統合指標を最大にする特徴をその文字パターン群を分割するのに最適な特徴として選択する．

4 分類木生成アルゴリズム

ステップ 1

使用する特徴を $T_i (i = 1 \cdots N)$，分割数を $c (c = 1 \cdots C)$ とする時，

```
for ( i = 1 ; i <= N ; i++ ) {
   for ( c = 1 ; c <= C ; c++ ) {
      Kohonenの自己組織化 ( T_i , c ) ;
      統合指標 ( T_i , c ) ;
   }
}
```

というように特徴，分割数を逐次変え，Kohonen の自己組織化を行ない文字パターン群を分割し，(2) 式によって統合指標をそれぞれ求める．

ステップ 2

ステップ 1 で求めた統合指標を最大にする特徴 T_{max}，分割数 c_{max} をこの文字パターン群を分割するのに最適な特徴，分割数とする．

ステップ 3

決定した特徴 T_{max}，分割数 c_{max} で BP により再学習を行なう．

ステップ 4

分割した各グループの文字パターン数 M 個未満かどうかを調べ，M 個以上ならばその文字パターン群に対してステップ 1 からステップ 3 まで繰り返す．M 個未満ならば，そこで葉の生成を停止する．

5 評価実験

5.1 使用した手書き文字パターン

通産省電子技術総合研究所提供の手書き教育用文字データベース ETL8B の中の最初の一セット 320 文字を使用して評価実験をした．一文字につき 50 パターンを学習文字パターンとして手順に従って分類木を生成し，また別の 50 パターンを実験用の未学習文字パターンとして扱うことにした．細線化，正規化処理を施した二値画像からメッシュ特徴 (256 次元)，端点交点特徴 (256 次元)，方向線素特徴 (256 次元)，輪郭線特徴 (256 次元)，周辺分布特徴 (169 次元)，ペリフェラルパターン特徴 (192 次元) を抽出し，統合指標によって特徴を選択させ分類木を作成した．

5.2 バランスのとれた分類木の作成

逐次分割していくグループ数は 2 個と固定し，木の高さは 6 とし，分割するグループ数は 32 としたバランス木を作成した．比較として分類度のみ，そしてエントロピーを特徴選択の基準としてみた．

表 1: 木の高さを制限した分類木

指標	平均	σ^2	最大	最小	分類率
分類度	37.0	44.63	191	5	95.2%
entropy	66.6	26.34	106	23	95.2%
統合指標	58.8	25.18	97	26	95.3%

5.3 バランスを考慮しない分類木の作成

バランスのとれた木ではなく末端の葉で処理するカテゴリー数が M 個以下になるまで分類し続ける方式で分類木を生成した．末端の葉で処理するカテゴリー数の上限は 70 個とした．

表 2: 末端の葉のカテゴリー数の上限を決めた分類木

指標	最短ノード	最長ノード	分類率
分類度	3	16	94.2%
entropy	4	10	95.7%
統合指標	4	8	96.1%

6 まとめ

本研究では大分類部を木構造化する手法を提案した．その結果，分類度，バランス，線形分離度を考慮した統合指標を用いた場合，他の特徴選択の指標と比較して候補文字数を絞り込むことができ，分類率の高い木が生成できることを示した．

参考文献

[1] 篠沢 大駒：二分木を利用した大分類・詳細認識型文字認識ニューラルネットワークの作成，情報処理学会第 52 回（平成 8 年前期）全国大会，Vol2-233.

[2] T. コホネン：自己組織化と連想記憶，シュプリンガー・フェアラーク東京（1993）.

[3] 中野馨：入門と実習 ニューロコンピュータ，技術評論社（1989）.