文字認識における特徴空間分割に関する一考察

1L－10

早川祥史　米田政明　長谷博行　酒井充

富山大学工学部電子情報工学科

1 はじめに

特徴空間におけるカテゴリ分布は一般に入り組んだ状態にあると考えられる。そのため１カテゴリに複数の代表パターンを設けて認識を行う試みが多くなされた[1][2]。しかし、代表パターンを作るためには、カテゴリ内クラス分けをする必要がある。我々は、手書き数字を対象にして、他カテゴリのサンプルの存在を利用して自動的にカテゴリ内を分割することを試みた。実験は、原画像からメッシュ特徴を抽出したもの、原画像を正規化してメッシュ特徴を抽出したものに対してクラスタリングを行なった。その結果、生じるクラスタの数や大きさと認識率に関して検討した結果を報告する。

2 重心法によるクラスタリング

クラスタ間の距離として、各クラスタの重心間の距離を用いる。クラスタ C_f とクラスタ C_g が統合されクラスタ C_h となる時、他クラスタ C_l との距離 D_{hl} は

$$D_{hl}^2 = \frac{n_f}{n_h} D_{fl}^2 + \frac{n_g}{n_h} D_{gl}^2 - \frac{n_f n_g}{n_h^2} D_{fg}^2$$

と計算できる。ここに n_h, n_f, n_g はそれぞれクラスタ C_h, C_f, C_g を構成する要素の数である。

3 他カテゴリのサンプルを利用したクラスタリング

これまでカテゴリ内のサンプルの分布状態を考慮しないで標準パターンを作成していたが、カテゴリ間の分布が入り組んでくると認識誤りは避けられない。そこで分布形状にあった標準パターン構成ができれば認識に効果的に働くものと思われる。

まず、ｎ次元空間に学習サンプルが散在しているものとし、第ｉカテゴリに注目する。以下にアルゴリズムを記す。

① 第ｉカテゴリの全サンプルを初期クラスタとし、他にｉ以外の全サンプルを用意する。

② クラスタ間の距離がｉ以外のサンプルとの距離よりも小さい「近傍クラスタペア」を探す。

③ 第ｉカテゴリ内に最近傍クラスタペアがあるなら、重心法により統合して新クラスタ（統合されたサンプルの数の重みをもつ）とし、②へ。最近傍クラスタペアがなければ（第ｉカテゴリのどのクラスタも他カテゴリのサンプルが最近傍ならば）統合はせず、処理を終了する。

本方法の利点としては、通常のクラスタリングの停止条件となる最終クラスタ数あるいは最大クラスタ間距離が必要ない。

4 実験および考察

対象とするデータはなるべく変動が大きいものが望ましいので、郵政研究所から提供していただいた郵便番号枠内に記入された０～９までの数字を用いた。これは、８本／ｍｍの解像度で前処理は一切なく、サインペン、ペン、ボールペン、毛筆で書かれた文字からなる。また、１組の３つの数字はそれぞれ個別に扱った。

実験は次の条件により黒領域の占める割合をメッシュ特徴としたものについて行なった。

表 1: 条件と特徴抽出

条件１	原画像 (128 * 128) を均等分割 (8 * 8)
条件２	原画像を重心分割 (8 * 8)
条件３	原画像を縦横等倍し均等分割 (8 * 8)
条件４	原画像を枠一杯に拡大し均等分割 (8 * 8)

特徴を求めるのに、均等分割法と重心分割法を用いた。重心分割法は、文字の水平方向と垂直方向に投影したヒストグラムから求まる重心点で４つに領域を分割し、さらにそれぞれの領域で重心点を求め分割する方法である。以上の条件により抽出した特徴を用いて、カテゴリ内の平均値を辞書としたものと、クラスタリングを行ない、それによって得られた全クラスタを辞書としたものについて認識を行なった。また、クラスタ間の距離尺度にはユークリッド距離を用いた。

表２は各条件において生成されたクラスタ数と、カテゴリの平均値を辞書としたものの認識率（認識率１）、

A Consideration on Separation of feature Space for Character Recognition

Yoshifumi Hayakawa, Masaaki Yoneda, Hiroyuki Hase, Mitsuru Sakai

Toyama university

クラスタリング後の全クラスタを辞書として用いた時（多重辞書）の認識率（認識率２）、を各文字種毎に示した。また、最下段に平均値を示している。

図１は各条件における認識率１、認識率２の平均値を示している。

図２は各条件において各文字種で生成されたクラスタ数の割合（辞書に用いたサンプル数は文字種毎に違うので、サンプル数に対するクラスタ数の割合）と、全文字種の認識率の散布図である。

表２において、認識率１は条件１、２、３、４の順に認識率が上がっている事がわかる。ゆえに、抽出された特徴は条件１、２、３、４の順に良い特徴である事がいえる。また、図１、図２より、良い特徴抽出を行なうことによってクラスタ数は少なくなり、更に認識率が上がって行く事がわかる。また、条件１の結果から悪い特徴を選んでも本手法では、自動的にクラスタ数が増加し認識率を上げる効果があることがわかる。また、今回用いた変動の大きなデータに対しては、重みの大きいクラスタが数箇所に分布し、必ずしも１クラスタに集中することはなかった。。これより、本手法は良好な特徴抽出に対して特に適切なクラスタリングを行なっていることがわかる。

図３は本クラスタリング手法を原画像（128*128、２値）に適用して得られた主なクラスタの画像を文字種“１”と“６”について表示したものである。図より文字種“１”に関しては傾きの違いが表れ、文字種“６”については文字の大きさの違いが表れている事がわかる。

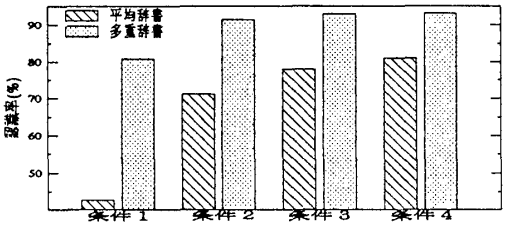

図 1: 平均辞書と多重辞書の認識率

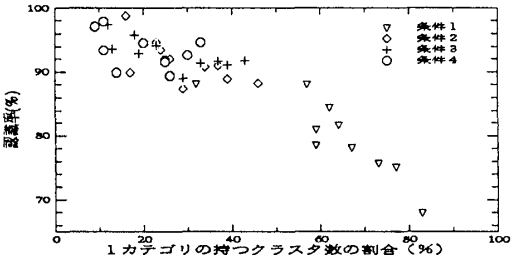

図 2: 全クラスタ数と認識率の分布

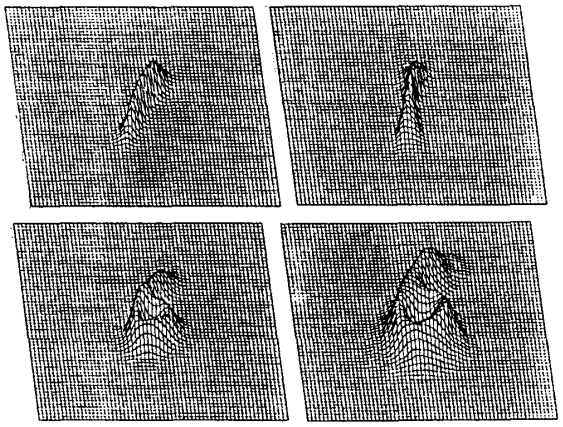
図 3: クラスタ“１”と“６”

5 まとめ

本稿では、他カテゴリのサンプルを利用したクラスタリングを提案した。本手法は、従来の様にクラスタリングの停止条件は必要なく、データに内在する分布に依存したクラスタが生成される。なお、本実験では本手法の効果を評価する事が目的なので、クラスタの大きさ（広がり）の情報は用いなかった。これを用いる事により更に良い認識率を得る事ができると思われる。

参考文献

[1] 大倉, 塩野, “カテゴリー内クラスタリングによる多重辞書類似度法の辞書パターン作成の一検討”, 信学論 D-II vol.J72-D-II No.4, (1989)

[2] 高橋, 佐野, “類似カテゴリを考慮したクラスタリングによる辞書設計”, 電子情報通信学会 D-325 vol.6-327, (1992)

[3] 河口至商著, “多変量解析入門 II”, 森北出版, (1978)

表 2: クラスタ数と認識率

	条件１			条件２			条件３			条件４		
文字種	クラスタ数	認識率１	認識率２	クラスタ数	認識率１	認識率２	クラスタ数	認識率１	認識率２	クラスタ数	認識率１	認識率２
0	300	35.2	78.5	90	81.2	89.8	66	81.1	93.5	58	89.6	93.4
1	119	76.9	88.2	59	98.7	98.7	44	87.4	97.3	33	86.1	97.1
2	197	46.0	81.7	121	58.0	88.9	101	68.8	91.3	79	77.1	91.8
3	270	32.2	81.0	133	64.8	87.3	87	85.0	92.8	65	89.9	89.9
4	147	48.6	75.8	93	75.8	88.2	79	73.5	91.0	50	79.8	91.5
5	172	27.3	68.0	76	69.5	90.9	89	65.8	91.7	69	77.8	94.6
6	149	37.8	88.1	64	70.2	93.4	47	83.6	95.7	31	82.7	97.8
7	241	39.5	75.2	73	84.0	94.6	118	73.9	91.6	95	74.1	92.6
8	289	52.3	78.2	149	62.1	90.7	126	75.7	89.0	112	77.5	89.3
9	410	34.6	84.5	177	57.6	92.0	152	75.7	94.1	136	72.8	94.5
平均	229	42.6	80.7	103	71.1	91.4	90	77.9	92.9	72	80.9	93.1