最高位候補別パラメータによる候補列マッチング分類

2L－7

高橋邦夫，天沼博，小林賢一，熊谷憲二
神奈川大学工学部電気工学科

１．　まえがき

分類は認識の前処理として重要な意味をもっている。認識・分類法については多くの研究がなされている。分類に関しては，グループ化分類[2]，一定パラメータによる候補列マッチングが知られている。本文では一定パラメータを用いることなく候補ごとにパラメータを定める候補別パラメータによる候補列マッチング分類を提案する。

２．　分類法

2.1　候補列辞書の作成

学習候補列データから学習辞書を作成する。字種iの学習候補列$K_\ell(i,j)$の順位Lまでの学習サンプル番号ℓ=1,3,...,159に対し候補字種jが存在すれば１とし，存在しなければ０とする。これを$C^{(L)}(i,j)$とする。一般の候補列は$K(j)$とする。i，jは字種数881だけ存在する。

2.2　候補別パラメータによる候補列マッチング分類法

次のようにして候補別パラメータＴを決定する。

$$T=\max_n\left\{2\sum_{j=1}^{N}C^{(L)}(i(n),j)k(j)+P_n\right\}/2-P_c/2 \quad (1)$$

n=1,2,3...であり候補列順位を示し，$i(n)$は順位nの字種iを示す。字種iは候補列字種とする。

P_1=12，P_2=6，P_3=5，P_4=4，P_5=3，P_6=1...である。式（２）を満足する字種iを候補として採用する。

$$\left\{2\sum_{j=1}^{N}C^{(L)}(i(n),j)k(j)+P_n\right\}/2\geq T \quad (2)$$

ここで，$k(j)$は入力となる候補列であり字種jが存在すれば$k(i)$=1，存在しなければ$k(i)$=0である。

2.3　グループ化分類法

更に，グループ化を行って分類率の向上を図る。

次の条件を満足するとき，字種$i,k_1,\ldots,k_{n-1}$を候補として採用する。$k(j)$は入力となる候補列である。

$$\left\{2\sum_{j=1}^{N}C^{(L)}(i(n),j)k(j)+P_n\right\}/2$$
$$\geq \max_n\left\{2\sum_{j=1}^{N}C^{(L)}(i(n),j)k(j)+P_n\right\}/2-\delta T_i^{(n)} \quad (3)$$

字種$k_1,\ldots,k_{n-1}$は次の条件により決定される。

$$T_i^{(n)}=\min_{m,k,\ell,k\neq k_1,k_2,\ldots,k_{n-2}}\left\{\max_n\right.$$
$$\left.\left\{2\sum_{j=1}^{N}C^{(L)}(m(n),j)K_\ell(k,j)+P_n\right\}/2-C_{ik\ell}\right\} \quad (4)$$

$$C_{ik\ell}=\left\{2\sum_{j=1}^{N}C^{(L)}(i(n),j)K_\ell(k,j)+P_n\right\}/2 \quad (5)$$

３．　候補列パラメータによる分類結果

3.1　１次候補列データ

漢字候補列データとしてはＥＴＬ－８を用い構造化パターンマッチングＤ[1]により，グループ化分類したデータを第１のデータとして用いる。また，構造化パターンマッチングＤ，ストロークマッチングＳ[1]，ずらし類似度Ｚとし，Ｄ＋Ｓ＋Ｚ[1]により，グループ化分類したデータを第２のデータを用いる。更に，Ｄ＋Ｓ＋Ｚにより30位までを候補としたものを第３のデータとする。これらを表１に示す。

Classification by Candidate Series Matching - parameter per input character
Kunio Takahashi, Hiroshi Amanuma, Kenichi Kobayashi, Kenji Kumagai
Department of Electrical Engineering
KANAGAWA University
3-27-1 Rokkakubashi Kanagawa-ku Yokohama 221 Japan

3.2 分類結果

1次候補漢字を用い，2．2に述べた分類手法による結果を表2に示す。但し，L=N=10とする。

更に，第2のデータ自身と第3のデータによる結果の共通候補を候補として採用した結果を表3に示す。

更に，学習候補列を候補列として学習し，候補列と完全に一致したものを候補として採用した結果は表4に示される。

比較のために，第1のデータ及び第2のデータの順位と平均個数，分類率を表5，6に示す。第1のデータでは順位を用いる方法に比べ約30%～45%程度の平均個数の減少，第2のデータでは，約20%程度の平均個数の減少が達成された。

4．あとがき

本分類法による平均個数は順位を用いる手法に比べ大幅に減少し，分類に有効であることが示された。

ＥＴＬ－８を作成された電総研関係者に感謝いたします。

参照文献

（1）高橋邦夫，天沼　博，加藤弘之：”手書き漢字の学習パラメータによる分類－構造化パターンマッチング等による－”，信学論(D-II) J75-D-II, 5, pp.674-675(1992-03).

（2）高橋邦夫，天沼　博，加藤弘之：”学習グループ化による手書き漢字の分類”，信学論(D-II) J75-D-II, 9, pp.1626-1627(1992-09).

表1　1次候補列データ

	平均個数	分類率[%]	1個率[%]
第1のデータ	21.962	99.600	19.052
第2のデータ	3.114	99.459	57.443
第3のデータ	32.524	99.650	

表2　候補列マッチングによる結果

		Pc	平均個数	分類率
1	第1のデータ	20	14.416	99.581
2	〃	18	11.801	99.557
3	〃	16	9.221	99.489
4	第2のデータ	20	2.705	99.437
5	〃	18	2.557	99.420
6	第3のデータ	20	18.031	99.645
7	〃	16	10.980	99.613
8	〃	12	5.029	99.485

表3

		Pc	平均個数	分類率	1個率
1	第3のデータ	20	2.877	99.454	60.456
2	〃	16	2.588	99.437	61.592
3	〃	10	2.042	99.350	65.088

表4

		Pc	平均個数	分類率	1個率
1	第3のデータ	20	2.014	99.277	84.696
2	〃	16	1.868	99.277	84.366
3	〃	12	1.572	99.158	84.876

表5　第1のデータ

順位	平均個数	分類率
1	1.000	90.576
2	1.814	95.207
20	14.805	99.409
21	15.523	99.436
22	16.240	99.457
24	17.676	99.508
25	18.393	99.530
26	19.111	99.550
29	21.264	99.590
30	21.982	99.601

表6　第2のデータ

順位	平均個数	分類率
1	1.000	95.726
2	1.426	98.171
3	1.726	98.770
4	1.960	99.039
5	2.153	99.190
6	2.346	99.280
7	2.538	99.346
8	2.730	99.389
9	2.922	99.427
10	3.114	99.459