

付加情報を用いるパターン認識

岩村 雅一<sup>†a)</sup> 内田 誠一<sup>††</sup> 大町真一郎<sup>†††</sup> 黄瀬 浩一<sup>†</sup>

Pattern Recognition with Supplementary Information

Masakazu IWAMURA<sup>†a)</sup>, Seiichi UCHIDA<sup>††</sup>, Shinichiro OMACHI<sup>†††</sup>, and Koichi KISE<sup>†</sup>

あらまし 本論文ではパターンが属するクラスの情報（付加情報）をパターンと同時に識別器に入力し、パターンと付加情報から矛盾のない答を導くことで誤認識を防ぐ方式を検討する。この方式では付加情報の情報量が増えれば増えるほど認識率は100%に近づく。そのため、従来のパターン認識のように、いかに認識性能を向上させるかではなく、ある認識率を達成するために必要な付加情報の情報量をいかに小さくできるかが課題となる。本論文では付加情報の割当方と認識性能の関係を導き、実験によりデモンストレーションする。

キーワード 識別器, 情報量, 混同行列, 付加情報

1. まえがき

パターン認識とは、一般にパターンから得られる特徴量を用いて、そのパターンをクラスの一つに対応づける処理である。ペイズ誤りとしてよく知られているように、この枠組みには原理的な限界がある。すなわち、一般にははずみ等によりパターンの特徴量はある分布をなして広がっており、その結果、2クラスの分布にはオーバーラップする範囲が存在する。この範囲に属するパターンは本質的にあいまいなものであり、その正しい認識は不可能である。

こうした限界のもと、オーバーラップがなるべく少なくなるような特徴、すなわち高い認識率が得られるような特徴の抽出が検討されてきた。その努力により実用的な精度をもつ認識器が多く開発されている。しかしながら、それらの認識対象ははずみが少ないなど、条件の良いものに限定されていることが多い。例えば市販OCRの対象は手書き文字ではなく活字が中心である。また、識別に有効な特徴量は必ずしも自明でなく、

開発者の経験に頼らざるを得ないという難点もある。

本論文では、図1(a)のような通常のパターン認識とは異なり、図1(b)のようにパターンから抽出される特徴量と、パターンとは別のチャンネルで与えられるクラス情報（以下、付加情報と呼ぶ。これは後述の記号の集合である）を併用することで、この限界を緩和する方法について検討する。この付加情報は、上述の

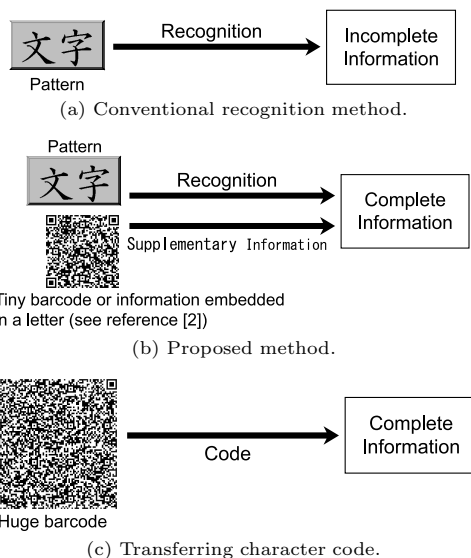


図1 認識と情報伝達のモデル (a) 通常のパターン認識 (b) 本論文で検討する認識モデル (c) コードをそのまま伝達

Fig.1 Recognition and data transferring models .

<sup>†</sup> 大阪府立大学大学院工学研究科, 堺市  
Graduate School of Engineering, Osaka Prefecture University, Sakai-shi, 599-8531 Japan

<sup>††</sup> 九州大学大学院システム情報科学研究院, 福岡市  
Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka-shi, 812-8581 Japan

<sup>†††</sup> 東北大学大学院工学研究科, 仙台市  
Graduate School of Engineering, Tohoku University, Sendai-shi, 980-8579 Japan

a) E-mail: masa@cs.osakafu-u.ac.jp

パターンのあいまい性を解消するための記号である。例えば、文字の「O」と「口」があいまいであり誤認識されやすいことが分かっているならば、文字「O」と「口」には異なる記号 1 と 2 をそれぞれ対応づけておく。そして、入力パターンが、特徴量による識別の結果「O」か「口」のどちらかであることが分かれば、あとは付加情報を用いて、その 2 クラスのどちらであるかを判定すればよい。例えば付加情報が 1 ならばその入力パターンは「O」と判定できる。

付加情報をパターンとは独立に認識系に与えることで認識性能の改善を図る研究には、最近ではアクティブ帳票 [1] があり、複比やアフィン不変量を用いた付加情報の埋込も検討されている [2], [3]。これらは付加情報を与えることによって認識候補を減らし、認識率の向上をねらった手法である。しかし、与えられた識別器に対してどのような情報を付加すれば認識性能がどの程度向上するのか等、付加情報と認識性能の関係については検討されていないため、本論文ではこの点について検討する。

なお、提案手法では付加情報を認識対象とは別に与えるのに対して、認識対象そのものを誤り訂正符号として置き換えてしまい、冗長性をもたせる方法も提案されている [4]。文献 [4] と提案手法の関連については 6. で考察する。

## 2. 認識性能の指標

### 2.1 識別器の認識性能と必要な情報量

付加情報をどれだけ与えれば、どれだけの認識性能が得られるのであろうか。最も単純な場合を例に検討してみよう。簡単のために、本論文では付加情報は誤らないものとする。 $N$  クラスに対して  $K$  種類の記号<sup>(注1)</sup> ( $K \leq N$ ) を割り当てるとすると、このときの付加情報の情報量は  $\log_2 K$  ビットである。もし  $N$  種類の記号 ( $\log_2 N$  ビットの情報) を用いるならば、これは図 1(c) のようにコードをそのまま伝達することに等しく、認識系は確実に正解クラスを言いあてることができる。したがって、認識率 100% が達成可能である。しかし、認識率 100% を達成するためには常に  $N$  種類の記号が必要なわけではなく、識別器の誤り傾向が既知であり、性能が良ければ必要な情報量は減少する。このように、ある認識率を達成するために必要な情報量は識別器の性能に依存するものと考えられる。そこで本論文では、識別器の誤り傾向をもとに、付加情報を用いるパターン認識の理論を展開する。

本論文では識別器の誤り傾向を表す指標として Confusion Matrix (混同行列、以下 CM と呼ぶ) を用いる。CM を用いて認識誤りに対処する研究としては、文書画像から OCR を用いてテキスト情報を抽出し、全文データベースを構築する際に CM を用いる方法 [5] がある。しかし、これは認識後に行う処理であり、CM の情報を認識に反映する提案方式とは異なる。

### 2.2 Confusion Matrix

CM は識別器に入力されるパターンが属するクラス (真のクラス) と識別器が出力するクラス (認識結果) の対応を表す行列である。通常、 $\omega_1, \omega_2, \dots, \omega_N$  をクラス、 $N \times N$  行列  $C$  を CM とすると、 $C$  の  $(i, j)$  要素  $c_{ij}$  はクラス  $\omega_i$  に属するパターンを認識によってクラス  $\omega_j$  に属すると判断した回数を表す。

CM の推定問題は確率分布の推定問題に帰着するため、CM の推定には十分な数のサンプルを用いるべきである。

次に、確率を要素とする、確率化した CM を定義する。クラス  $\omega_i$  に属するパターンをクラス  $\omega_j$  であると認識する確率  $P(\omega_j | \omega_i)$  を、 $(i, j)$  要素にもつ  $N \times N$  行列  $W$  とする。 $W$  は、 $C$  の各行の要素の和が 1 になるように正規化して得られる。すなわち、

$$C_i = \sum_{j=1}^N c_{ij} \quad (1)$$

とおいたとき、 $W$  の  $(i, j)$  要素  $w_{ij}$  は

$$w_{ij} = \frac{c_{ij}}{C_i} \quad (2)$$

で与えられる (図 2 参照)。

		Recognition Result				
		A	B	C	D	E
The True Class	A	0.6		0.4		
	B		0.8		0.1	0.1
	C	0.1		0.9		
	D		0.1		0.8	0.1
	E	0.2	0.1			0.7

図 2 確率化した Confusion Matrix の例。数字のない要素は値が 0 である。

Fig. 2 An example of a probabilistic confusion matrix. An empty element represents zero-valued one.

(注 1): 本論文で記号と呼ぶものは、情報理論では符号語と呼ばれる。

### 3. 認識率 100%かつリジェクト率 0%を実現する付加情報

図 1 (b) のような認識系にパターンと付加情報が与えられるモデルを考えたとき、認識率 100%かつリジェクト率 0%を実現する付加情報が満たすべき条件を導き、グラフを用いた定式化を行う。なお、本章の議論は CM が誤りなく得られるという理想的な状況考えたものである。そのため、より一般的かつ実用的な問題については 4. で議論する。

以後、認識系は記号とクラスの対応表と、CM (行列  $W$ ) をもっているとする。また、簡単のために各クラスのサンプルは等確率で生成されるとする。すなわち、クラス数を  $N$  としたとき、事前確率は  $P(\omega_i) = \frac{1}{N}$  である。

#### 3.1 行列 $W$ の分割

提案方式では CM の各行に付加情報となる記号を割り当てる。  $k$  番目の記号を割り当てた行の集合を

$$\mathcal{H}_k = \{l_1, \dots, l_{|\mathcal{H}_k|}\} \quad (3)$$

とおき、 $W$  を  $\mathcal{H}_1, \dots, \mathcal{H}_K$  に分割する。ここで  $|\mathcal{H}_k|$  は  $k$  番目の記号を割り当てられた行数である。例えば、図 3 は図 2 の CM に 3 種類の記号を割り当てた例であるが、この場合は  $\mathcal{H}_1 = \{1, 2\}$ ,  $\mathcal{H}_2 = \{3, 4\}$ ,  $\mathcal{H}_3 = \{5\}$  となる。ただし、 $\mathcal{H}_k$  の要素は連続している必要はない。

次に、 $k$  番目の記号が割り当てられた行の要素のうち  $j$  列目の要素を

$$B_{kj} = \{(l, j) \mid l \in \mathcal{H}_k\} \quad (4)$$

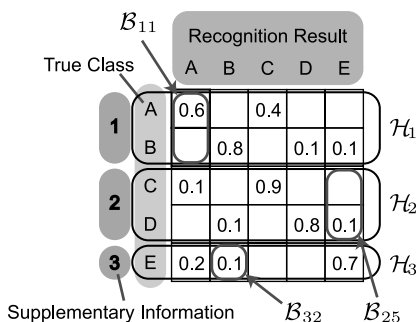


図 3 認識率 100%かつリジェクト率 0%を実現する付加情報 (記号の種類: 3, 認識率: 100%, リジェクト率: 0%)

Fig. 3 Symbols that achieves a 100% recognition rate with a 0% rejection rate. # of symbols:3, Recog. rate:100%, Reject. rate:0%.

とおく、 $B_{kj}$  の要素は CM の行と列を表す。例えば図 3 の場合、 $B_{11} = \{(1, 1), (2, 1)\}$ ,  $B_{25} = \{(3, 5), (4, 5)\}$ ,  $B_{32} = \{(5, 2)\}$  である。

#### 3.2 付加情報が満たすべき条件

認識率 100%かつリジェクト率 0%を実現するために付加情報が満たすべき条件を図 2 を例に考える。図 2 によると認識結果が A になるのは真のクラスが A, C, E のいずれかのときである。これらは当然、識別器では区別できない。もし識別器がそのまま A と出力してしまうと、真のクラスが C や E だった場合には誤認識が生じてしまう。そのため、真のクラスが A, C, E のいずれであったのかを区別する情報が必要となる。そこで A, C, E にそれぞれ別々の記号を割り当てると、誤認識が生じない代わりに少なくとも 3 種類の記号が必要となる。同様に、認識結果が B になるのは真のクラスが B, D, E のいずれかのときであるので、B, D, E にも別々の記号を割り当てる。

このように考えていくと、図 3 のように、A と B に一つ、C と D に一つ、E に一つ、と 3 種類の記号を付加することでリジェクトを行うことなく認識率 100%が達成できる。また、A と D に一つ、B と C に一つ、E に一つ、という割当でも同様である。つまり、認識率 100%かつリジェクト率 0%を実現するために付加情報が満たすべき条件は、すべての  $k, j$  について  $B_{kj}$  の要素が指し示す CM の要素に、値が 0 でないものが一つ以下しか存在しないことである(注2)。

#### 3.3 情報量が最小の付加情報を発見する問題

グラフを用いて情報量が最小の付加情報を発見する問題を定式化する。各クラスに対応する  $N$  個のノードからなる集合  $V$  と、 $V$  の要素間を結ぶエッジの集合  $E$  を考え、無向グラフ  $G = (V, E)$  を定義する。同じ記号を割り当てても誤認識が起こらない 2 クラスを、ノード間にエッジを張ることで表す。

該当するすべてのエッジを張り、最後にグラフ  $G$  をできるだけ少ない完全グラフに分割し、完全グラフごとに唯一の記号を割り当てる。このときの完全グラフの数を  $N_{\text{symmin}}$  とおくと、 $N_{\text{symmin}}$  種類の記号があれば認識率 100%かつリジェクト率 0%を実現できる。このとき必要な情報量は  $\log_2 N_{\text{symmin}}$  である。例として、図 3 を表すグラフを図 4 に示す。

$N_{\text{symmin}}$  は最も誤認識されやすいクラスの性質に

(注2): CM に 0 の要素が一つもない場合に認識率 100%を達成するには明らかに  $N$  種類の記号が必要になる。

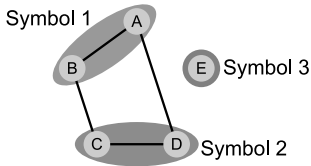


図 4 図 3 に対応するグラフ  $G$  の分割．一つの完全グラフが一つの記号に対応する．  
 Fig. 4 The division of the graph  $G$ , which corresponds to Fig. 3. Each complete graph corresponds to a symbol.

The True Class		Recognition Result				
		A	B	C	D	E
1	A	0.9		0.1		
2	B		0.9	0.1		
3	C			1		
4	D			0.1	0.9	
5	E			0.1		0.9

Supplementary Information

図 5 誤認識されやすいクラスを含む Confusion Matrix を確率化した例．どのクラスのサンプルもクラス C に間違えよう．  
 Fig. 5 An example of a probabilistic confusion matrix which includes the class into which patterns are often misclassified. The samples of all classes can be recognized as those of class C.

よって決まる値である．一般にパターン認識問題には認識が容易なクラスと困難なクラスが存在しているが、同様に誤認識されやすいクラスとされがたいクラスも存在する．誤認識されやすいクラスとは、CM を列ごとに見たとき、0 でない要素が多い列に対応するクラスである（例えば、図 5 のクラス C）．認識率 100% を達成するためには、識別器の認識結果が最も誤認識されやすいクラスになったとしても、認識系全体では正しい認識結果を出力できるだけの情報量を与える必要がある．例えば、図 2 の CM の場合は  $N_{\text{symmin}} = 3$  なので  $\log_2 3 \sim 1.58$  ビットの情報が必要となり、図 5 の CM の場合は  $N_{\text{symmin}} = 5$  なので  $\log_2 5 \sim 2.32$  ビットが必要である．図 2 と図 5 の CM から計算される認識率はそれぞれ 76% と 92% であるので、単純に認識率が高ければ付加情報が小さくて済むというものではない．このような現象が実際の CM でも起こることを 5. の実験で示す．

なお、情報量が最小の付加情報を見出す問題は NP

完全問題である PARTITION INTO CLIQUES [6] を最小化問題にしたものであり、NP 困難問題となる．したがって、現在のところ多項式時間で実行可能な、最適解を保証するアルゴリズムは知られていない．

#### 4. 付加する情報量と認識性能

3. では認識率 100%かつリジェクト率 0%を達成する付加情報の情報量について論じた．本章では 3. の議論を拡張し、認識率が 100%ではない場合や、リジェクト率が 0%でない場合について論じる．実は付加情報の情報量、達成可能な認識率、必要なリジェクト率は三つのうち二つの値を決めると残りの一つの値が決まる関係である．ここでは、(1) リジェクトを行わないときに付加する情報量と認識率の関係、(2) 誤認識率を 0%に保ったときに付加する情報量と必要なリジェクト率の関係、について考える．なお、これらの問題は 3. 3 の問題を含んでいるため、少なくとも 3. 3 の問題と同程度難しいと考えられる．

##### 4. 1 $B_{k_j}$ 内の 0 でない要素数

3. の議論を一般化するため、 $B_{k_j}$  内の 0 でない要素数を定義する．最初に、値が 0 であれば 0、さもなければ 1 を返す関数

$$z(x) = \begin{cases} 0, & \text{for } x = 0 \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

を定める．すると、 $B_{k_j}$  に含まれる要素のうち、 $w_{l_j} \neq 0$  を満たす要素の数  $q_{k_j}$  は関数  $z(x)$  を用いて次式で与えられる．

$$q_{k_j} = \sum_{(l,j) \in B_{k_j}} z(w_{l_j}) \quad (6)$$

##### 4. 2 リジェクトを行わないときに付加する情報量と認識率の関係

3. 2 で述べたように、 $B_{k_j}$  内に値が 0 でない要素が一つ以下しか存在しない場合 ( $q_{k_j} \leq 1$ ) は誤認識が生じず、二つ以上の 0 でない要素が存在した場合 ( $q_{k_j} \geq 2$ ) は誤認識が生じる．このとき、認識結果になり得る真のクラスは  $q_{k_j}$  通りの可能性があるが、認識結果として出力できるのはそのうち一つのクラスだけである．誤認識率が最も低くなる方法は、事後確率が最大になる要素を出力として採用することである．例えば、図 6 の場合、記号が 2 で認識結果が A であれば、C と出力するよりも E と出力した方が誤認識率が低くなるので、E と出力する．

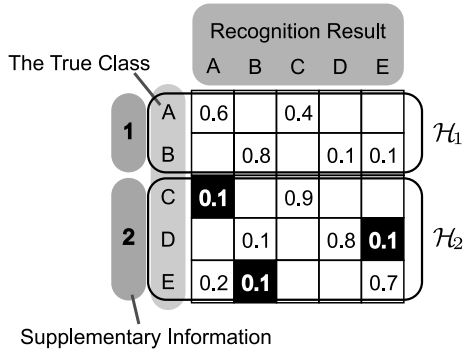


図 6 リジェクトを行わず、誤認識を許容する付加情報。色が反転している要素が誤認識が起こる状況を表す (記号の種類: 2, 誤認識率: 0.3/5 = 6%, 認識率: 94%, リジェクト率: 0%)。

Fig. 6 Symbols that accepts recognition errors without rejection. The reverse colored elements are causes of misclassification. # of symbols: 2, Misclassi. Rate: 0.3/5=6%, Recog. rate: 94%, Reject. rate: 0%.

以上より、記号の数を  $K$  としたとき、誤認識率関数  $R_{\text{error}}(K)$  は次式のように定式化できる。

$$R_{\text{error}}(K) = \frac{1}{N} \sum_j \sum_k \left\{ \sum_{(l,j) \in \mathcal{B}_{kj}} w_{lj} - \max_{(l,j) \in \mathcal{B}_{kj}} w_{lj} \right\} \quad (7)$$

式 (7) の右辺の括弧内は第 1 項が  $\mathcal{B}_{kj}$  内の要素の和、第 2 項が識別器の認識結果に対して認識系が出力するクラスに対応する要素を表す。

このように定式化された式 (7) を用いて、誤認識率を最小化する問題を解くわけだが、この問題の最適解を求めることは非常に困難である。実際、この問題が解ければ NP 困難問題である 3.3 で定式化した問題が解けるため、少なくともそれと同程度難しい問題であるといえる。そこで本論文では、誤認識率や次節で扱うリジェクト率を近似的に最低にするアルゴリズムとして、Algorithm 1 の欲張り法を用いる。Algorithm 1 では、記号の種類  $K$  は  $N$  から一つずつ減っていき、そのつど誤認識率やリジェクト率を計算する。 $K = 1$  は通常のパターン認識と等価である。Algorithm 1 を図 2 の CM に適用した場合の付加情報と誤認識率  $R_{\text{error}}$  の推移を図 7 に示す。

なお、提案手法では、CM を推定するためのサンプル数が少ない等の理由によって本来は 0 となるべきではない (しかし 0.001 等の 0 に非常に近い値となるべき) 要素が 0 となったとしても、言語モデルにおけ

Algorithm 1: 欲張り法のアルゴリズム。

- 1 すべてのクラスに別々の記号を割り当てる。すなわち、 $k = 1, \dots, N$  について、 $\mathcal{H}_k = \{k\}$  である。
- 2 誤認識率やリジェクト率を損失として扱い、損失を表す関数を  $L(K) = R_{\text{error}}(K)$  または  $L(K) = R_{\text{reject}}(K)$  と定義する。明らかに  $L(N) = 0$  である。
- 3 for  $K = N - 1$  to 1 do
- 4     同じ記号を割り当てたときの損失  $L(K)$  が最小となる行の集合を 1 組選び、 $\mathcal{H}_s, \mathcal{H}_t$  とする。
- 5      $\mathcal{H}_s$  と  $\mathcal{H}_t$  のクラスに同じ記号を割り当てて記号の数を 1 つ減らして  $K$  個にする。すなわち、
 
$$\mathcal{H}_s \leftarrow \mathcal{H}_s \cup \mathcal{H}_t \quad (8)$$

$$\mathcal{H}_t = \emptyset \quad (9)$$
 とする。
- 6     損失  $L(K)$  を計算する。
- 7 end

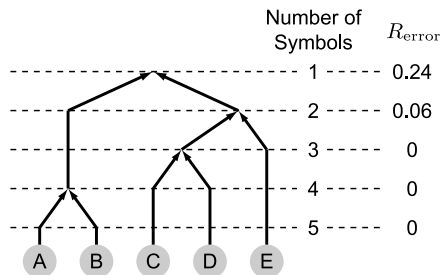


図 7 図 2 の Confusion Matrix に対して Algorithm 1 を用いたときの付加情報と誤認識率  $R_{\text{error}}(K)$  の推移

Fig. 7 Transition of symbols and the misclassification rate  $R_{\text{error}}(K)$  with Algorithm 1 for the confusion matrix in Fig. 2.

るゼロ頻度問題 [7] のような致命的な影響は生じない。なぜなら、ある  $w_{lj}$  が 0 であっても 0.001 であっても、式 (7) の計算結果にはほとんど影響を与えないからである。同様の議論は 4.3 のアルゴリズムに対しても成り立つ。

4.3 誤認識率を 0% に保ったときに付加する情報量と必要なリジェクト率の関係

4.2 同様、 $\mathcal{B}_{kj}$  内に値が 0 でない要素が一つ以下しか存在しない場合 ( $q_{kj} \leq 1$ ) は誤認識が生じず、二つ以上の 0 でない要素が存在する場合 ( $q_{kj} \geq 2$ ) は誤認識が生じる。認識結果になり得る真のクラスは  $q_{kj}$  通りの可能性があるが、誤認識率 0% を達成するためには、誤る可能性のある場合はすべてリジェクトするしかない。例えば、図 8 の場合、記号が 2 で認識結果が A であれば、C と出力しても E と出力しても誤認識の可能性が残る。そのため、この場合はいずれもリ

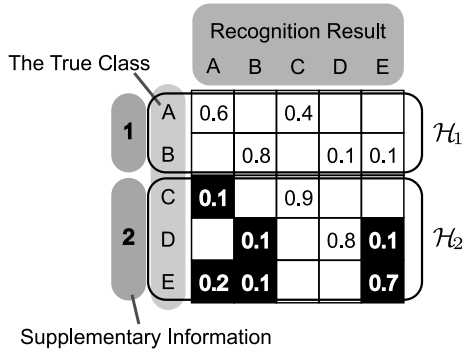


図 8 リジェクトによって誤認識を避ける付加情報。色が反転している要素がリジェクトされる状況を表す(記号の種類:2, 認識率:100%, リジェクト率:1.3/5 = 26%)。

Fig. 8 Symbols that achieves a 0% misclassification rate by rejection. The reverse colored elements are causes of rejection. # of symbols:2, Recog. rate:100%, Reject. rate:1.3/5 = 26%.

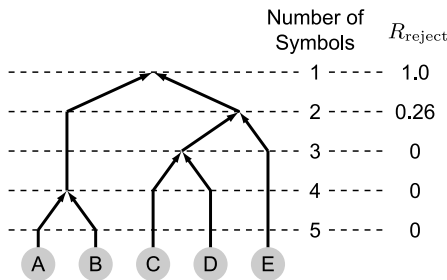


図 9 図 2 の Confusion Matrix に対して Algorithm 1 を用いたときの付加情報とリジェクト率  $R_{\text{reject}}(K)$  の推移

Fig. 9 Transition of symbols and the rejection rate  $R_{\text{reject}}(K)$  with Algorithm 1 for the confusion matrix in Fig. 2.

ジェクトする。

以上より、記号の数を  $K$  としたとき、リジェクト率関数  $R_{\text{reject}}(K)$  を次式のように定式化できる。

$$R_{\text{reject}}(K) = \frac{1}{N} \sum_j \sum_k s_{kj} \quad (10)$$

ただし、ここで  $s_{kj}$  は

$$s_{kj} = \begin{cases} 0, & \text{for } q_{kj} \leq 1 \\ \sum_{(l,j) \in B_{kj}} w_{lj}, & \text{otherwise} \end{cases} \quad (11)$$

である。式 (11) は、 $q_{kj} \leq 1$  であればリジェクトされず、それ以外であれば  $B_{kj}$  内の要素がすべてリジェクトされることを意味している。

4.2 と同様に、Algorithm 1 を図 2 の CM に適用した場合の付加情報とリジェクト率  $R_{\text{reject}}(K)$  の推移を図 9 に示す。

## 5. 実験

本論文の議論は特定の応用分野に特化しているわけではないが、例として文字認識における付加する情報量と認識率、リジェクト率の変化を観察する<sup>(注3)</sup>。以下、5.1 では付加情報を作成するために用いるテストサンプルの認識結果 (CM) が既知である場合、5.2 では CM が未知であるため、学習サンプルから推定する場合について実験を行う。

### 5.1 CM が既知である場合

CM が既知である場合について、記号の数  $K$  と  $R_{\text{error}}(K)$  の関係、 $K$  と  $R_{\text{reject}}(K)$  の関係を観察する。実験には実際の文字データをオープン実験で認識して得た 4 種類の CM を用いた。4 種類の CM は 2 種類の文字データと 2 種類の識別器の組合せによる。文字データには手書き文字と印刷文字を用いた。手書き文字は ETL9B (3036 字種, 200 セット) を  $64 \times 64$  の大きさに非線形正規化 [8] した後、196 次元の方向線素特徴量 [9] を抽出して用いた。印刷文字は 25 種類のフォントから ETL9B に含まれる 3036 字種の平仮名と漢字を抜き出し、ETL9B と同様の処理で特徴量を作成した。識別器としてはユークリッド距離と SQDF [10] を用いた。

実験手順は以下のとおりである。印刷データは leave-one-out 法で、手書きデータは 10-fold クロスバリデーションで行った。CM は既知であるとしたので、まずテストサンプルに対して認識を行い、その認識結果から CM を作成し、CM から誤認識率とリジェクト率の理論値を計算した。付加情報の与え方は 2 種類試し、Algorithm 1 による割当を “Greedy” とし、単純割当を “Simple” とした。単純割当とは、全クラスに 0 から始まる通し番号を付け、記号の数を  $K$  としたとき、クラス  $m$  には  $m \bmod K$  を付加情報として割り当てる方法である。

$K$  と  $R_{\text{error}}(K)$  の関係を図 10 に、 $K$  と  $R_{\text{reject}}(K)$  の関係を図 11 に示す。両図から、与える付加情報の情報量が増加するにつれて誤認識率とリジェクト率が減少していることが分かる。欲張り法と単純割当を比

(注3): ここでの目的はパターン認識に付加情報を導入したときの認識率やリジェクト率の変化を観察することであるので、パターンに付加情報を埋め込む具体的な方法については議論しない。

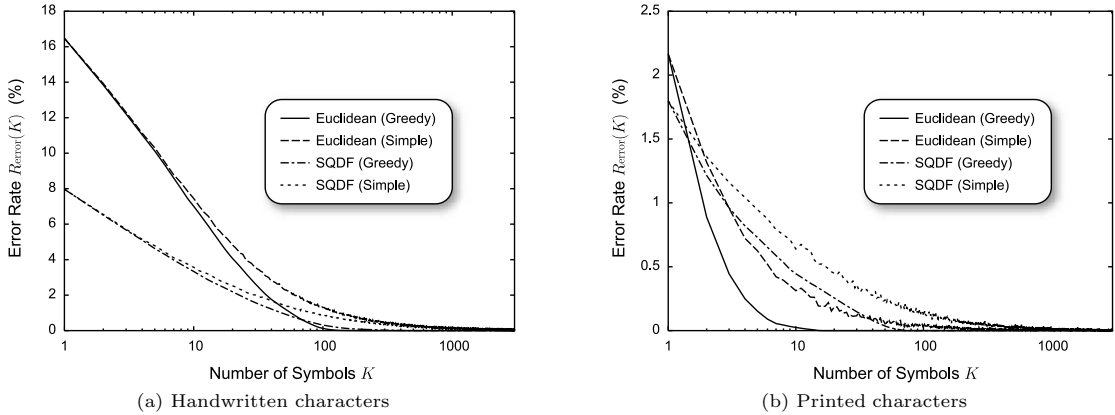


図 10 真の CM を用いて求めた付加する記号の数  $K$  と誤認識率  $R_{error}(K)$  の関係  
 Fig. 10 Relationship between the number of symbols  $K$  and the error rate without rejection  $R_{error}(K)$  calculated with the true CMs.

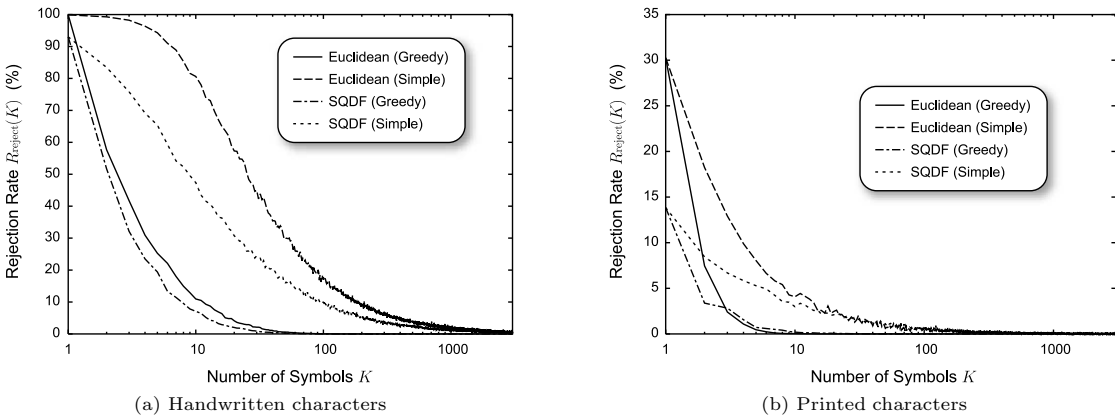


図 11 真の CM で求めた付加する記号の数  $K$  とリジェクト率  $R_{reject}(K)$  の関係  
 Fig. 11 Relationship between the number of symbols  $K$  and the rejection rate for the 100% recognition rate  $R_{reject}(K)$  calculated with the true CMs.

較すると、ほとんどすべての条件で欲張り法の方が優れていた。翻って、図 10 において  $K$  が小さいときには両者の差はほとんどなかった。このことから、認識性能の改善効果はおおむね付加情報の割当方法に依存すると考えられるが、割当方法によらない状況の存在も予想される。

SQDF とユークリッド距離を比較すると、識別器単体の認識率では SQDF が上回っているが、付加情報を与えた場合の認識率は必ずしも SQDF が優位ではなかった。例えば、図 10 (a) では  $K$  が 70 程度で、図 10 (b) では  $K = 2$  で両者の認識率が逆転している。これは 3.3 で述べた誤認識されやすいクラスの影響である。ユークリッド距離の認識率はそれほど高くないが、認識誤りの偏りが小さく、認識結果のクラスが

比較的分散していた。一方、SQDF は認識率が高い反面、多くのクラスが特定のクラスに誤認識される傾向が見られた。これが原因で、SQDF は識別器単体での認識率が高いにもかかわらず、同じ情報量の付加情報を加えた場合の認識性能の改善効果が低いという現象が見られた。このように付加情報を用いるパターン認識では、従来のように識別器単体で最高の認識率を目指すのではなく、付加情報を加味した上で、一定の認識性能を達成するために付加する情報量の最小化を目指すべきである。言い換えると、多少認識率は低くても誤認識されやすいクラスを生み出しにくい識別器が求められる。

### 5.2 CM が未知の場合

CM が未知である場合について実験する。この実験

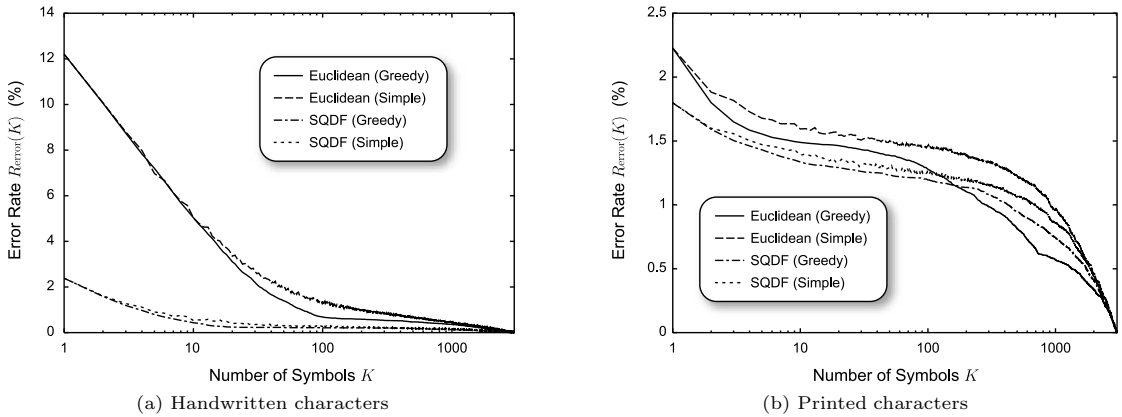


図 12 推定した CM で求めた付加する記号の数  $K$  と誤認識率  $R_{\text{error}}(K)$  の関係  
 Fig. 12 Relationship between the number of symbols  $K$  and the error rate without rejection  $R_{\text{error}}(K)$  calculated with estimated CMs.

では CM が既知であるとした 5.1 とは異なり、テストサンプルを認識したときの CM を学習サンプルのみから推定する必要がある。提案手法では、付加情報を仮に割り当てたときの誤認識率を CM と式 (7) を用いて計算し、付加情報の可能な割当の中からもなるべく誤認識率が小さくなるような付加情報の割当を選択する。そのため、推定した CM の値が真の値と異なるという問題は、誤認識率の推定値に影響を与え、最適な付加情報の割当が選択されることにより、結果的に誤認識率の上昇を引き起こす。ところが、CM の推定誤差のために付加情報の割当に最適性がない場合や、まして適当に割り当てた場合においても、付加情報を用いることにより誤認識率は確実に減少する。このことを以下の実験で確認する。

CM は以下の手順で作成した。印刷データに対しては leave-one-out 法を実行した。まず、25 種類のフォントを 1 種類のテスト用 (A) と 24 種類の学習用 (B) に分け、B を用いて学習と CM の作成を行う。以下、B を用いた CM の作成方法について述べる。B から CM を作成するときは、B を更にテスト用 1 種類 (BA) と残り 23 種類の学習用 (BB) に分けて認識実験を行い、CM を一つ求める。この処理を BA となるフォントを変えながら繰り返し、得られた 24 個の CM を足し合わせて CM を得る。ここまですべて学習である。この CM を用いて付加情報を計算し、A に対する認識実験を行う。この手順を、A となるフォントを変えながら 25 回繰り返し、平均の認識率とリジェクト率を計算する。

手書きデータに対してはカテゴリ当り 200 サンプ

ルを 10 分割する 10-fold クロスバリデーションを行った。印刷データのとときの変更点は A がクラスごとに 20 サンプル、B が 180 サンプル、BA が 18 サンプルになったことである。

4.2 と 4.3 で述べた関係を計算し、図 12 と図 13 に図示する。図中の“Greedy”と“Simple”の意味は 5.1 と同じである。図 12 より、欲張り法は単純割当よりも誤認識率が低いか、ほとんど同じだった。CM が既知である図 10 と比べると、CM の推定誤差のために記号の正しい割当ができず、欲張り法の優位性は低下している。一方、単純割当は記号の割当をランダムに行うため、CM が完全に分からない状況における認識性能の期待値を表していると考えられる。したがって、CM の推定誤差が極端に大きい場合であっても、大体この程度の誤認識率で済むことが期待できる。

記号の割当誤りについて詳しい考察を行うため、実際に生じた CM の推定誤差の一例を表 1 に示す。これは「あ」の例であるが、他の文字でも同様の傾向が見られる。このように、誤り傾向は似ているものの、辞書を作成するためのサンプル数が異なることにより、異なる CM が得られた。すなわち、5.2 の実験で用いた CM (180 サンプル使用) では、5.1 で用いた CM (200 サンプル使用) と比較して、正解となる(「あ」が「あ」と認識される)確率が減少し、誤認識する(正解以外の文字と認識される)確率が高くなった CM が得られた。したがって、Algorithm 1 で計算される損失  $L(K)$  が実際の値とは異なる値となり、その結果 5.1 の実験とは異なる符号割当が起こり、認識



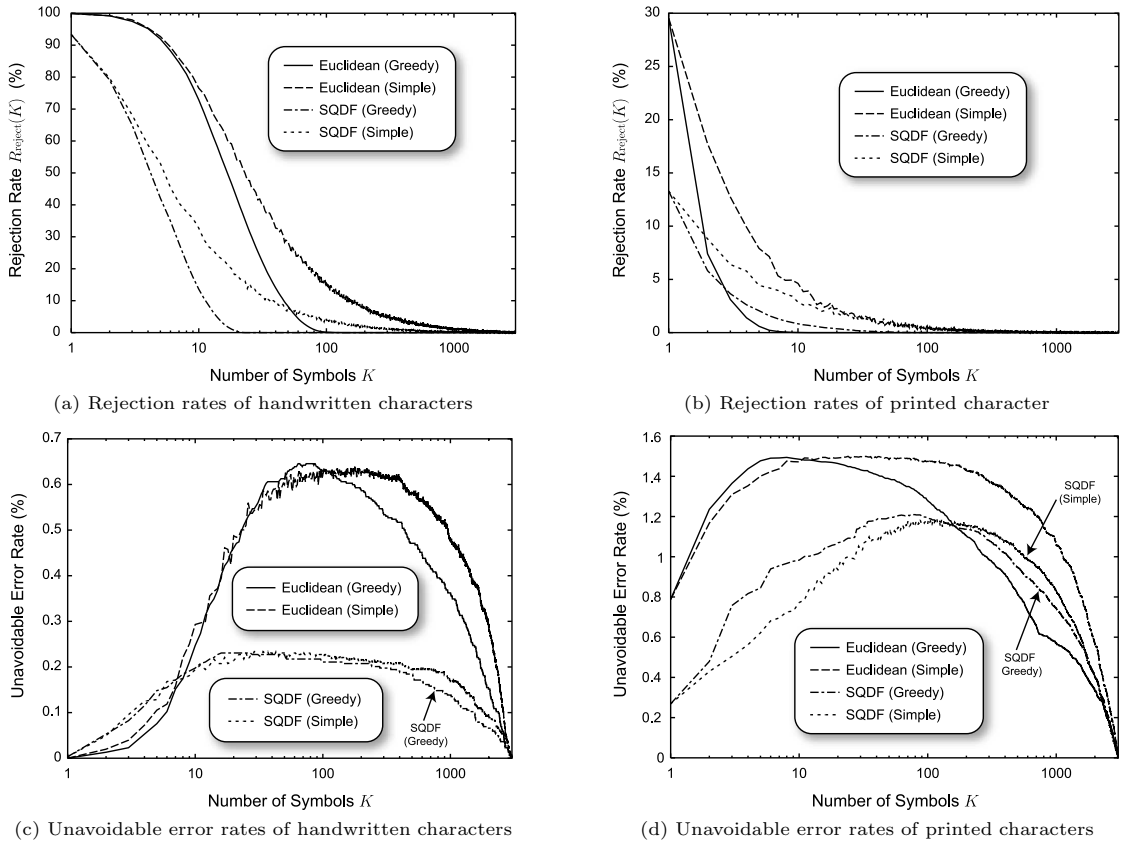


図 13 推定した CM で求めた付加する記号の数  $K$ 、リジェクト率  $R_{\text{reject}}(K)$  と、不可避な誤認識率の関係

Fig. 13 Relationship among the number of symbols  $K$ , the rejection rate  $R_{\text{reject}}(K)$  and the unavoidable error rate calculated with estimated CMs.

表 1 手書きデータの「あ」をユークリッド距離で認識した結果の詳細

Table 1 The detail of the recognition result of “あ” of handwritten characters using Euclidean distance.

認識結果	5.1 の「真の CM」での頻度	5.2 の「推定した CM」での頻度
あ	169	149
わ	6	5
お	3	3
乃	2	3
め	2	2
石	2	2
(その他)	16	16
計	200	180

率が低下している。

図 13 に結果を示すリジェクトを考慮した実験は誤認識率を 0% にする目的で行った。真の CM を用いたときとは異なり、図 13 (c) と図 13 (d) に示すように

誤認識が生じたが、誤認識率は 1.5% 以内に収まっている。これは実際の状況においても付加情報が誤認識の発生を回避するための一助になることを示している。欲張り法と単純割当を比べると、図 13 (c) と図 13 (d) では両者の誤認識率はほとんど同じであった。しかし、図 13 (a) と図 13 (b) における両者のリジェクト率を比べると、単純割当のリジェクト率は欲張り法に比べて高いため、この点においては欲張り法の方が優れているといえる。

## 6. 通信路モデルとの関連

最後に、本論文で提案する付加情報を用いるパターン認識と、情報理論における通信路モデルとの関連について述べておく。

通常の文字認識問題は、CM を通信路行列と考えれば、多元で非一様（不均一）な通信路における問題と

みなすことができる。実際、このような観点から文字認識の誤り訂正を扱った文献 [4] のような研究もある。この手法は誤り訂正のための情報を文字として表現し、1本の通信路で伝送するモデルである。それに対して、本論文で提案する付加情報を用いるパターン認識という枠組みは、文字認識の通信路とは別に付加情報を伝送するという2本の通信路を使ったモデルであると考えられる。このように、提案するモデルは通常の通信路行列を用いたモデルと少なくとも形式的には異なっている。

また、3. や 4. で述べたように、一般の CM に対して最適な誤り訂正符号を設計する問題は NP 困難に陥ると考えられるため、現実的には仮定や近似など、何らかの妥協が求められる。文献 [4] では、誤り訂正符号が設計しやすいように、対象とする CM に比較的強い仮定をもたせている。それに対して本論文では、CM に対する一切の仮定を行わない代わりに、最適な付加情報（文献 [4] の誤り訂正符号に対応）の割当を諦め、近似値を用いている。このような付加情報の割当に最適性が保証されない場合の認識性能については 5. で検証した。

このように、本論文で提案した付加情報を用いるパターン認識という枠組みは、情報理論における多元で非一様な通信路の問題と関連が深いと思われる。このような分野との関連を明らかにすることは重要な今後の課題の一つである。

## 7. む す び

本論文ではパターン認識の実用性を向上させるために、パターンが属するクラスの情報（付加情報）をパターンと同時に識別器に入力し、パターンと付加情報から矛盾のない答を導く方式を検討した。3. では、正しい CM が与えられたとき、認識率 100% を達成するために必要な情報量と付加情報の割当方を導いた。4. では 3. の議論を一般化して、付加情報の情報量、達成可能な認識率、必要なりジェクト率の三者の関係を導いた。

実験では文字認識結果を反映した 8 種類の CM を用いて、(1) 付加する情報量と認識率の関係と (2) 誤認識率を 0% に保ったときに付加する情報量と必要なりジェクト率の関係を観察した。その結果、識別器の誤り傾向 (CM) が既知であっても未知であっても付加情報を用いることにより、認識性能が向上することが確認できた。

本論文では更に、付加情報の情報量は、識別器の性能評価によく用いられている認識率とは別の性質をもつことを示し、付加情報の情報量を最小にする識別器の必要性を指摘した。そのような識別器を導入することにより、認識性能が更に改善される可能性がある。

## 文 献

- [1] 島村太郎, 朱 碧蘭, 櫻田武嗣, 中川正樹, “アクティブ帳票システムの設計と開発” 信学論 (D-II), vol.J87-D-II, no.12, pp.2091–2103, Dec. 2004.
- [2] 内田誠一, 岩村雅一, 大町真一郎, 黄瀬浩一, “カメラによる文字認識のためのカテゴリー情報の埋込に関する検討” 信学論 (D), vol.J89-D, no.2, pp.344–352, Feb. 2006.
- [3] 大町真一郎, 岩村雅一, 内田誠一, 黄瀬浩一, “面積比による付加情報を用いた実環境文字認識” 画像の認識・理解シンポジウム (MIRU2006) 論文集, pp.187–192, July 2006.
- [4] 稲葉宏幸, 寺島史雄, 若杉耕一郎, 笠原正雄, “誤り訂正符号を適用した文字認識システムに関する一考察” 信学論 (D-II), vol.J77-D-II, no.2, pp.353–361, Feb. 1994.
- [5] 太田 学, 高須淳宏, 安達 淳, “認識誤りを含む和文テキストにおける全文検索手法” 情処学論, vol.39, no.3, pp.625–635, March 1998.
- [6] M.R. Garey and D.S. Johnson, *Computers and Intractability*, W.H. Freeman and Company, New York, 1979.
- [7] 永田昌明, “文字類似度と統計的言語モデルを用いた日本語文字認識誤り訂正法” 信学論 (D-II), vol.J81-D-II, no.11, pp.2624–2634, Nov. 1998.
- [8] 山田博三, 斉藤泰一, 山本和彦, “線密度イコライゼーション—相関法のための非線形正規化法” 信学論 (D), vol.J67-D, no.11, pp.1379–1383, Nov. 1984.
- [9] 孫 寧, 田原 透, 阿曾弘具, 木村正行, “方向線素特徴量を用いた高精度文字認識” 信学論 (D-II), vol.J74-D-II, no.3, pp.330–339, March 1991.
- [10] S. Omachi, F. Sun, and H. Aso, “A new approximation method of the quadratic discriminant function,” *Lect. Notes Comput. Sci.*, vol.1876, pp.601–610, Sept. 2000.

(平成 18 年 5 月 8 日受付, 9 月 5 日再受付)



### 岩村 雅一 (正員)

平 10 東北大・工・通信卒．平 15 同大大学院博士課程了．同年同大学院工学研究科助手．平 16 阪府大大学院工学研究科助手．現在に至る．博士(工学)．パターン認識，コンピュータビジョンに関する研究に従事．IEEE，情報処理学会，Pattern Recognition Society 各会員．

Recognition Society 各会員．



### 内田 誠一 (正員)

平 2 九大・工・電子卒．平 4 同大大学院修士課程(情報)了．セコム(株)勤務を経て，現在，同大システム情報科学研究院知能システム学部門助教授．博士(工学)．画像パターン・時系列パターンの解析・認識に関する研究に従事．平 15 本会 PRMU

研究奨励賞，平 18 MIRU 長尾賞各受賞．IEEE，日本音響学会，情報処理学会，映像情報メディア学会各会員．



### 大町真一郎 (正員)

昭 63 東北大・工・情報卒．平 5 同大大学院博士課程了．同年同大情報処理教育センター助手．平 8 同大工学部助手．平 11 同大大学院工学研究科助教授．現在に至る．博士(工学)．その間，平 12~13 米国ブラウン大学客員助教授．パターン認識，コンピュータビジョン，並列処理，文字認識システムの開発などの

研究に従事．IEEE，情報処理学会，人工知能学会，Pattern Recognition Society 各会員．



### 黄瀬 浩一 (正員)

昭 61 阪大・工・通信卒．昭 63 同大大学院博士前期課程了．同年同大学院博士後期課程入学．平 2 阪府大・工・電気助手．平 5 同情報助手，平 8 講師，平 11 助教授を経て，平 17 同大大学院工学研究科教授．現在に至る．博士(工学)．その間，平 12

年~13 ドイツ人工知能研究センター客員教授．文書画像解析，情報検索などの研究に従事．情報処理学会，人工知能学会，電気学会，IEEE，ACM 等各会員．