多カテゴリ認識問題における正解率の評価 － 教育パターンの効果 －

5Q－10

酒井 充　米田 政明　長谷 博行
富山大学工学部電子情報工学科

1 はじめに

我々は以前より多カテゴリ認識問題における正解率に関する理論を提案しているが[3]、今回、この理論を具体的な分布モデルに適用し、有効な結果を得ることができたので報告する。R.Y.Kain は G.F.Hughes の2カテゴリを対象としたモデル[1]を多カテゴリ認識問題に拡張し、分布が既知の場合の正解率を求めた[2]。我々はこの Kain の多カテゴリ認識問題のモデルにおいて、分布を教育パターンにより推定する場合の正解率を求めた。また、その結果をモンテカルロ法により確認した。

2 Hughes のモデルと Kain のモデル

まず、Hughes のモデルについて説明する。2つのカテゴリ C_1 と C_2 があり、パターン ω の特徴量 $x = ch(\omega)$ は Q 個の離散値 $\{x_1, x_2, \cdots, x_Q\}$ の1つをとるものとする。ここで、$\theta_j^{(i)} = P(x_j|C_i), i = 1,2$ とし、$\Theta = (\theta_1^{(1)}, \theta_2^{(1)}, \cdots, \theta_Q^{(1)}, \theta_1^{(2)}, \theta_2^{(2)}, \cdots, \theta_Q^{(2)})$ と表す。クラスの先験確率をここでは $P(C_1) = P(C_2) = 1/2$ とする。Θ が既知のとき、ベイズ決定による正解率 $Pc(\Theta)$ が定まる。ここで、Θを確率変数とし、どのΘも同等に確からしく生起するものとする[1, (8)]。この正解率の平均値 $\overline{Pc}(2,Q)$ は次式となる[1, (15)]。

$$\overline{Pc}(2,Q) = \frac{3 \cdot Q - 2}{4 \cdot Q - 2} \tag{1}$$

次に、教育パターンを用いて認識する場合を考える。教育パターンの頻度を $(s_1^{(1)}, s_2^{(1)}, \cdots, s_Q^{(1)}, s_1^{(2)}, s_2^{(2)}, \cdots, s_Q^{(2)})$ とし、$\sum_{j=1}^{Q} s_j^{(1)} = \sum_{j=1}^{Q} s_j^{(2)} = S$ とする。$x_j = ch(\omega)$ のとき、$s_j^{(i)}$ の大きいカテゴリに認識する。この正解率の平均値 $\overline{Pc}(2,Q,S)$ は次式となる[1, (19)]。

$$\overline{Pc}(2,Q,S) = \sum_{i=0}^{S}\sum_{j=0}^{S} \frac{\prod_{k=1}^{Q-2}(S-i+k) \cdot \prod_{k=1}^{Q-2}(S-j+k)}{(\prod_{k=1}^{Q-1}(S+k))^2} \cdot \frac{Q \cdot (Q-1)^2 (max(i,j)+1)}{2 \cdot (S+Q)} \tag{2}$$

Kain は Hughes のモデルを多カテゴリに拡張し、Θ が既知の場合の M カテゴリ認識問題の正解率の平均値 $\overline{Pc}(M,Q)$ を求めた[2, (11)]。

Evaluation of Correct Recognition Probability in a Multi-Class Recognition Problem – Effect of Design Data Set –
Mitsuru SAKAI , Masaaki YONEDA , Hiroyuki HASE
Toyama University
3190 Gofuku,Toyama,Toyama 930,Japan

$$\overline{Pc}(M,Q) = \frac{1}{M} \cdot (1 + \sum_{i=1}^{M-1} \frac{i!(Q-1)^i}{\prod_{j=1}^{i}(j \cdot Q + Q - j)}) \tag{3}$$

3 カテゴリ母集合と期待正解率

$\theta_j = P(x_j|C)$、$\theta = (\theta_1, \theta_2, \cdots, \theta_Q)$ とおくと、Kain の問題は次式の θ の分布をするカテゴリ母集合から、

$$dP(\theta) = (Q-1)!d\ \theta \tag{4}$$

M 個のカテゴリを取り出して構成したカテゴリ部分集合の正解率の期待値を求める問題と等価となる。我々はこのような問題を一般化し、カテゴリ母集合 $\mathbb{C}$ と分類方法 S が与えられているとき、大きさ M のカテゴリ部分集合の正解率の期待値 (期待正解率) を求める問題に一つの解答を与えた。すなわち、$\mathbb{C}$ のカテゴリ数が無限の場合の期待正解率 $EPc(M|\mathbb{C},S)$ は次式となる。

$$EPc(M|\mathbb{C},S) = \int_0^1 p_U(x) \cdot (1-x)^{M-1} dx \tag{5}$$

パターンωの真のカテゴリを $C(\omega)$ と表すとき、ω の入力に対し $\mathbb{C} - \{C(\omega)\}$ の中で $C(\omega)$ より得点の高いカテゴリの割合を規格化順位†と呼び、$U(\omega)$ と表す。規格化順位密度関数†† $p_U(x)$ はこの $U(\omega)$ を用いて次のように定義される。ここで全てのパターンの集合を Ω と表すものとする。

$$p_U(x)dx \equiv P(\omega; x \le U(\omega) < x + dx|\Omega) \tag{6}$$

4 Kain のモデルへの適用

まず、Kain のモデルに対して上式を適用し、Kain と同じ結果が得られることを示す。(4) 式において、θ_1 以外を積分すると次式となる。

$$dP(\theta_1) = (Q-1) \cdot (1-\theta_1)^{Q-2} d\theta_1 \tag{7}$$

よって、

$$\begin{aligned} p(t|x_j,\mathbb{C})dt &\equiv P(C; t \le P(x_j|C) < t + dt|\mathbb{C}) \\ &= (Q-1) \cdot (1-t)^{Q-2} dt \end{aligned} \tag{8}$$

となる。よって、

$$\begin{aligned} U(\omega) &= \int_0^1 p(t|x_j,\mathbb{C})dt \\ &= (1-t)^{Q-1} \end{aligned} \tag{9}$$

上式が全ての j について成り立つことから、

† 文献[3]では上位率と呼んでいたものである。
†† 文献[3]では上位率密度関数と呼んでいたものである。

$$p_U(x)dx = P(\omega; x \le U(\omega) < x+dx|\Omega)$$
$$= Q\cdot(1-x^{1/(Q-1)})dx \quad (10)$$

よって、期待正解率は上式と (5) 式より求まる。

$$EPc(M,Q) = Q\cdot(\frac{1}{M} - \frac{(M-1)!(Q-1)^M}{\prod_{i=1}^{M}(Q\cdot i - i + 1)}) \quad (11)$$

この式は Kain の得た (3) 式とは表現が異なるが、同じ式である。図 1、図 2 の中の $S=\infty$ の例はそれぞれ (10) 式と (11) 式で計算した結果と一致する。

5 教育パターンを用いる場合

次に、各カテゴリとも S 個の教育パターンを用いて認識する場合を考える。(5) 式の導出では教育パターンについて特に考慮していない。しかし、カテゴリ C とある S 個の教育パターンの組を新しいカテゴリ母集合 $\mathbb{C}'$ の一つのカテゴリとして登録する。全てのクラスと全ての可能なパターンの組み合わせに対してこの登録を行う。この $\mathbb{C}'$ を改めて $\mathbb{C}$ とすれば、$\mathbb{C}$ のカテゴリ数が無限の場合、(5) 式が成り立つ。

カテゴリ C において $t = P(x_j|C)$ とすると、S 個の教育パターンの中で x_j を値とするパターンが i 個発生する確率を $P(i|x_j,S,C)$ と表すと、これは ${}_SC_i\cdot t^i\cdot(1-t)^{S-i}$ となる。よって、この値の平均値は次式となる。

$$P(i|x_j,S,\mathbb{C}) \equiv P(C; i=s_j|x_j,S,\mathbb{C})$$
$$= \int_0^1 P(i|x_j,S,C)\cdot p(t|x_j,\mathbb{C})dt$$
$$= {}_{S-i+Q-2}C_{Q-2} / {}_{S+Q-1}C_{Q-1} \quad (12)$$

入力 ω によっては $C(\omega)$ と同点のカテゴリが存在することがある。同点カテゴリの存在しない ω ではその規格化順位の密度関数 $p_U(x|\ \omega)$ は $x=U(\omega)$ に単位インパルスを持つが、同点カテゴリがある場合には、同点カテゴリの割合を $V(\omega)$ とすると、次式となる。

$$p_U(x|\ \omega) = \begin{cases} 1/V(\omega) & , U(\omega)\le x<U(\omega)+V(\omega) \\ 0 & , \text{その他} \end{cases} \quad (13)$$

$x_j = ch(\omega)$ のとき、$U(\omega)$ と $V(\omega)$ は次式となる。

$$V(\omega) = P(i|x_j,S,\mathbb{C})$$
$$U(\omega) = \sum_{k=i+1}^{S} P(k|x_j,S,\mathbb{C}) \quad (14)$$

上式の値と $V(\omega'),U(\omega')$ がそれぞれ等しくなるようなパターン ω' の生起する確率は次式となる。

$$P(i|S,\Omega) \equiv P(\omega'; ch(\omega') = x_j, i=s_j, 1\le j\le Q|\Omega)$$
$$= \frac{Q}{S+Q}\cdot(i+1)\cdot\frac{{}_{S-i+Q-2}C_{Q-2}}{{}_{S+Q-1}C_{Q-1}} \quad (15)$$

ここで、u_i を次のように定義する。

$$u_i \equiv \sum_{k=i}^{S} P(k|S,\mathbb{C})$$
$$= \frac{{}_{S+Q-1-i}C_{Q-1}}{{}_{S+Q-1}C_{Q-1}} \quad (16)$$

よって、規格化順位密度関数は次式となる (図 1)。

$$p_U(x) = \frac{Q\cdot(i+1)}{S+Q},\ u_{i+1}\le x<u_i, 0\le i\le S \quad (17)$$

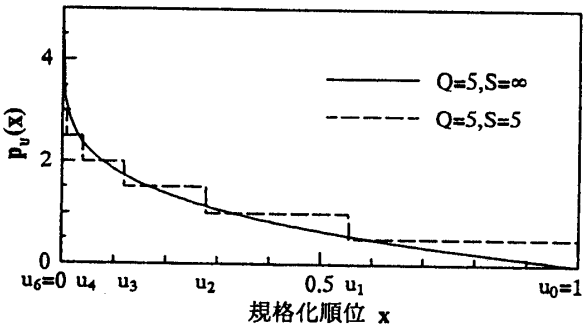

図 1 教育パターンを用いる場合の規格化順位密度関数

よって、期待正解率は以下のように求まる (図 2)。

$$EPc(M,Q,S) = \int_0^1 p_U(x)\cdot(1-x)^{M-1}dx$$
$$= \frac{Q}{M\cdot(S+Q)}\cdot(S+1-\sum_{i=1}^{S}(1-u_i)^M) \quad (18)$$

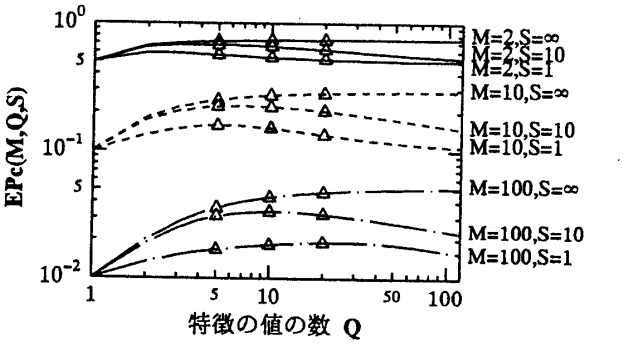

図 2 教育パターンを用いる場合の期待正解率

図中の △ は乱数を用いたシミュレーション結果であり、理論値とほとんど完全に一致した。

6 まとめ

我々は従来より多カテゴリ認識に関する研究を行っているが、その理論を適用することにより、Kain のモデルを拡張した教育パターンを用いる多カテゴリ認識問題を解いた。また、このことにより、我々の提案する理論が、同点カテゴリのある場合、教育パターンを用いる場合にも適用可能であることを示すとともに、本理論の有用性を示すことができた。

参考文献

[1] Hughes G.F. : *"On the mean accuracy of statistical pattern recognizers"*,**IT-14**,1,pp.55-63(1968).
[2] Kain R.Y. : *"The mean accuracy of pattern recognizers with many pattern classes"*,**IT-15**,3,pp.424-425(1969).
[3] 酒井 充 他 : "多カテゴリ認識問題の正解率の期待値に関する一考察",SITA'93,F22-2,pp633-636(1993-10).