多カテゴリ認識における期待正解率の下界と近似

−あるカテゴリ数の期待正解率と期待大分類率が既知の場合−

4X−2

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

文字認識問題のように識別すべきカテゴリ数の多い認識問題が存在する。一般にこのような多カテゴリ認識問題は小数カテゴリの認識問題よりも難しい。われわれはこのような多カテゴリ認識問題におけるカテゴリ数と正解率の関係を調べている。今回は、一様分布モデルの正解率の期待値と分散を与える式を示し、次に、あるカテゴリ数の期待正解率の値を用いて異なるカテゴリ数の期待正解率の下界を与える。さらにそれを用いて、カテゴリ部分集合の認識実験をしてその正解率と大分類率を求めたときに、異なるカテゴリ数の集合の正解率を近似した結果を報告する。

2．期待正解率

対象とするカテゴリ母集合を CT、i 番目の大きさ M のカテゴリ部分集合を CT^M_i、その正解率を $Pc_i(M)$ とする。このカテゴリ部分集合における正解率の期待値が期待正解率 EPc(M) である。

$$EPc(M) \equiv E[Pc_i(M)] \quad \cdots(1)$$

ここで CT のパターン母集合をΩ、パターンを ω、パターン ω の属するカテゴリを CT(ω) とする。この認識系では入力パターン ω と任意のカテゴリ CT の間に得点 S(ω, CT) が一意に定まるとする。この得点が大きいカテゴリをその入力パターンの属するカテゴリとして判定する。このとき期待正解率は次のように求められる[1]。

$$EPc(M) = \int_{\Omega} (1-P(CT;S(\omega,CT)>S(\omega,CT(\omega))))^{M-1}\cdot p(\omega)d\omega \quad \cdots(2)$$

また、$u(\omega)=P(CT;S(\omega,CT)>S(\omega,CT(\omega)))$ と定義すると次の式になる[2]。

$$EPc(M) = \int_0^1 (1-u)^{M-1}\cdot p_u(u)du \quad \cdots(3)$$

この式で $p_u(u)$ は $p_u(u)du=P(\omega;u\leq u(\omega)<u+du)$ なる密度関数であり、カテゴリ母集合と特徴系と分類方法が決まれば一意に定まるが、これを上位率密度と呼ぶ。また、$(1-u)^{M-1}$ はカテゴリ数によってのみ変化し、認識系によらない。つまり、u を媒介としてカテゴリ数によらず認識系によって決まる要素と認識系によらずカテゴリ数によってのみ決まる要素に分離できているのがこの式の特徴である。

次に、入力パターンに対してその真のカテゴリが第 r 位に認識される確率を順位認識率と呼び、Pc, r(M) と書くこととする。この順位認識率の期待値である期待順位認識率は次式のようになる[1]。

$$EPc,r(M) = {}_{M-1}C_{r-1}\int_0^1 u^{r-1}\cdot(1-u)^{M-r}\cdot p_u(u)du \cdots(4)$$

3．一様分布モデル

一様分布モデルを定義し、その期待正解率の式を示し、特に一次元の場合の正解率の分散の式を示す。

カテゴリ内分布は各カテゴリすべて同じ一様分布と仮定する。つまり、各カテゴリ中心をもとに一定の半径内に一様に分布している。またカテゴリ間分布、つまり各カテゴリ中心の分布も一様分布と仮定する。一般性を失わずに、カテゴリ内分布の密度を 1 、カテゴリ間分布の密度を h とする。

分類手法としてユークリッド距離を使用する。このとき、その上位率密度は特徴の次元数によらず一様分布となり、次のように期待正解率が求まる[4]。

$$EPcR(M,h) = (1-(1-h)^M)/(M\cdot h) \quad (h<=1) \cdots(5)$$

一次元の場合、正解率の分散は次のように求まる。

$$VPcR1(M,h) = \{2\cdot(1-h)^{M+1}+(M-1)\cdot(1-2\cdot h)^{M+1}-(M+1)\cdot(1-h)^{2M}\}/\{M^2\cdot(M+1)\cdot h^2\} \cdots(6)$$

$$\fallingdotseq (1-1/M)\cdot h/3 \quad (M\cdot h<<1)$$

$$\fallingdotseq 2\cdot\exp(-M\cdot h)/(M^3\cdot h^2) \quad (M\cdot h>>1)$$

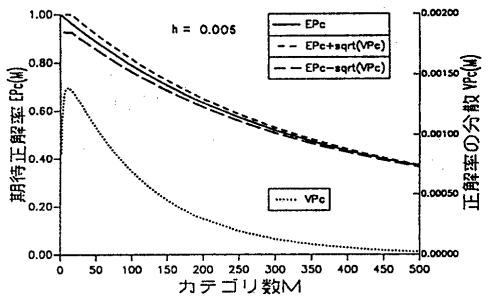

図1　一様分布モデルの期待正解率と分散

A Lower Bound and an Approximation of Expected Correct Probability of Multi-Class Recognition,
Mitsuru SAKAI , Masaaki YONEDA , Hiroyuki HASE
TOYAMA Univ.

図1は h=0.005 の場合の一様分布モデルの期待正解率(実線、(5) 式)と分散(細かい点線、(6) 式)の例である。

4．期待正解率の下界

期待正解率の下界についてはすでにいくつか示した[3][4]が、それらは EPc(2) を用いるものであった。今回はあるカテゴリ数 N の期待正解率 EPc(N) を用いて一般の場合の期待正解率 EPc(M) の下界の一つを示す。さらにこの下界を正解率の予測値としたときの誤差についても実験的に評価する。

4．1．下界の導出

期待正解率 EPc(N) と2位となる期待順位認識率 EPc,2(N) が既知の場合に EPc(M) の下界を導出する。

$$EPc(M) = \int_0^1 (1-u)^{M-1} \cdot p_u(u)\, du$$

ここで $f(u)=(1-u)^{N-2} \cdot p_u(u)/EPc(N-1)$ とすると

$$= EPc(N-1) \cdot \int_0^1 (1-u)^{M-N+1} \cdot f(u)\, du$$

$(1-u)^n$ は $(n \neq 0)$ において下に凸な関数なので

$$\geqq EPc(N-1) \cdot (1 - \int_0^1 u \cdot f(u)\, du)^{M-N+1}$$

ここで (4) 式の期待順位認識率を用いて変形すると

$$= EPc(N-1) \cdot \{1-EPc,2(N-1)/(N-1)/EPc(N-1)\}^{M-N+1}$$

$$= EPc(N) \cdot \{1-EPc,2(N)/\{(N-1) \cdot EPc(N)+EPc,2(N)\}\}^{M-N} \quad \cdots(7)$$

同様に期待順位認識率 EPc,r(M) の下界も求まる。

4．2．期待正解率の予測

上で求めた期待正解率 EPc(M) の下界を N=100 のときの期待正解率 EPc(100)=0.788 、期待順位認識率 EPc,2(100)=0.180 の場合について求め、図2に示す(実線、(7) 式による)。この図には 3. で求めた一様分布モデルの期待正解率のグラフ(図1の実線)を点線で重ねてある。これをみると、ここで求めた下界は良い下界になっていると考えられる。

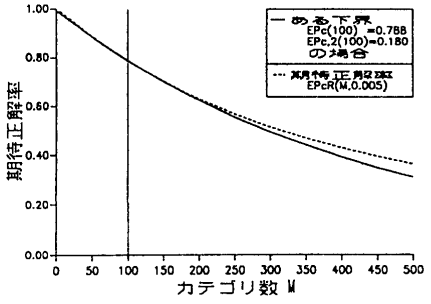

図2　ある下界と一様分布モデルの期待正解率

そこで、期待正解率の下界を正解率の近似式として用いる場合の精度を評価する。一次元の一様分布モデルで 100 カテゴリを発生させ、その正解率 Pc(100) と順位認識率 Pc,2(100) をもとに一般のカテゴリ数についての正解率 Pc(M) を(7)式を用いて近似する。図3にこの操作を 3000 回行ったときの平均とその偏差を示す(図には一様分布モデルの期待正解率(図1の実線)も細い実線で重ねてある)。

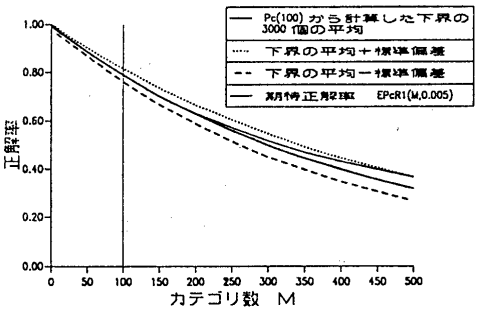

図3　実験データを用いた正解率の近似

図3を見ると、予測された正解率はカテゴリ数が大きくなるとそのバラつきが大きくなるが、このバラつきと誤り率との比は必ずしも大きくはなっていない。つまり、本質的なバラつきが大きくなるとは必ずしもいえない。このバラつきは主に最初に与えられた正解率 Pc(100) が期待正解率からバラつくことから生じたものである。

このようなバラつきを意味する正解率の分散は、同じ期待正解率を与える分布モデルでも一次元よりも多次元の方が小さくなり、また、一様分布モデルよりも正規分布モデルのような他の単峰性の分布モデルの方が小さくなると予想している。このことから、多次元の他の単峰性の分布モデルではさらに近似精度が良くなると予想される。

5．まとめ

一次元の一様分布モデルの正解率の分散を与える式を示し、次にあるカテゴリ数の期待正解率を用いて異なるカテゴリ数における期待正解率の下界を求めた。さらに一様分布モデルに適用しその有用性を示した。

参考文献

[1]酒井、米田：「多カテゴリ認識における誤り率について」、PRL81-104(1982-03)
[2]酒井、長谷、米田、吉田：「多カテゴリ認識における誤り率の解析－上位率を用いて－」PRL82-49(1982-11)
[3]酒井、米田、長谷、吉田：「カテゴリ数と誤り率の関係についての一考察」昭62信学総全大,Vol6,p193
[4]酒井、米田、長谷、吉田：「多カテゴリ認識の誤り率の上下界に関する一考察」昭62後期情処全大,pp.1971-1972