H-023

より大きな多カテゴリ認識問題の正解率推定に関する検討
Accuracy Estimation of a Classification Problem with More Categories

高橋 知生† 酒井 充† 丸山 博†
Tomoki TAKAHASHI Mitsuru SAKAI MARUYAMA Hiroshi

1. はじめに

パターン認識問題のカテゴリ数は、数字認識では 10 文字、英字では 52 文字であるが、日本の漢字であれば、教育漢字で 1006 文字、jis 第一水準では 2965 文字、第二水準では 3390 文字であり、物体認識のように数百万になる場合もある。地球上の全員の顔識別を考えようとすると、さらに多くなる。このように非常にカテゴリの多い認識問題（超多カテゴリ認識問題）では、認識システムの開発に膨大な時間が必要となる。そこで、実験時間の短縮のため、その部分集合を対象とする認識実験の結果を基に全体の認識率を精度良く推定する方法が望まれる。

超多カテゴリ認識問題では、母集合からランダムにカテゴリを選んで構成した部分集合では、母集合よりカテゴリ数をある程度少なくしても、統計的安定性を確保できる可能性が高いと考えられる。今回、このような状況下で、カテゴリ部分集合の 1 回の認識実験の結果を基に、母集合に対する正解率を推定する方法を提案する。しかし、1 つの正解率だけでは、精度の良い推定は出来ない。そこで、文献[1]で提案されている方法を用いることにより、より小さい同じ大きさの部分集合の正解率の期待値を計算できるので、それらの値を用いることにより、推定精度を上げることできると期待できる。提案手法の有効性を、人工データを用いた認識結果と、手書き漢字認識結果を用いて、評価する。実験では、10 倍のカテゴリ数に対する正解率の推定を行った。また、人工データの実験においては統計的安定性についても検討する。

2. カテゴリ部分集合に対する正解率の期待値

カテゴリ集合 $\boldsymbol{C}$ を認識対象とする認識システムを考える。このとき、大きさ m の部分集合の正解率を P_{Cm} とし、その期待正解率を $EP_C(m)$ と記述する。

$$EP_C(m) = \mathrm{E}[P_{Cm}] \tag{1}$$

文献[1]によると、全体の認識を 1 回行うことにより、任意の大きさ m の期待正解率 $EP_C(m)$ を、少ない計算量で求めることが出来る。ここで、$m < |\boldsymbol{C}|$ である。

$EP_C(m)$ には次の性質が成り立つ。

- $EP_C(m)$ は減少関数
- $EP_C(m)$ は下に凸な関数

3. 提案手法

推定のために与えられるカテゴリ集合の大きさを M とする。その認識を行うことにより、大きさ $m(<M)$ のカテゴリ部分集合の期待正解率を求めることが出来、それを P_{Cm}, $m=1,...,M$ と記述する。ただし、$P_{C1}=1$ である。

任意のカテゴリ数における正解率を、基底関数 $f(m;\alpha)$ の線形和で推定する。

$$P_C(m) = \sum_{i=1}^{n} w_i f(m;\alpha_i) \tag{2}$$

この重み係数 $w_i, i=1,...,n$ を P_{Cm}, $m=1,...,M$ を用いて求めることになる。ここでは重み係数は非負とした。今回、 α_i, $i=1,...,n$ には妥当な解が存在する範囲の値を与えた。

3.1 基底関数

今回使用する基底関数を次式に示す。

$$f(m;\alpha) = \frac{1}{1+\alpha(m-1)} \tag{3}$$

この関数は **2** 節で述べた期待正解率の性質を満たすので、式(2)の $P_C(m)$ も同じ性質を満たすことが分かる。

3.2 重み係数

式(2)の重み係数 w_i, $i=1,...,n$ の求め方を示す。次のようにベクトル $\mathbf{w}$、行列 A、ベクトル $\mathbf{b}$ を設定する。

$$\mathbf{w} = \begin{pmatrix} w_1 & \cdots & w_n \end{pmatrix}^t \tag{4}$$

$$A = \begin{pmatrix} f(1;\alpha_1) & \cdots & f(1;\alpha_n) \\ \vdots & \ddots & \vdots \\ f(M;\alpha_1) & \cdots & f(M;\alpha_n) \end{pmatrix} \tag{5}$$

$$\mathbf{b} = \begin{pmatrix} P_{C1} & \cdots & P_{CM} \end{pmatrix}^t \tag{6}$$

このとき $\mathbf{w}$ を未知ベクトルとする次の連立一次方程式が成り立つ。

$$A\mathbf{w} = \mathbf{b} \tag{7}$$

次式の誤差の二乗和が最小となる解を、解が非負かつ解の和が 1 となるように、求めた。

$$\left\| A\mathbf{w} - \mathbf{b} \right\|_2^2 \tag{8}$$

4. 実験

実験の認識対象として、人工データと手書き漢字データの 2 種類を用いた。まず、それぞれ、1000 カテゴリを対象とした認識を行い、それらの結果を基に大きさ $m<1000$ のカテゴリ部分集合の期待正解率 $EP_C(m)$ を文献[1]の手法で求めた。それらの結果を基に、M の値として 20 や 100 の値を設定し、$10M$ までのカテゴリ数の正解率を予想し、評価を行った。

† 富山大学, University of Toyama

4.1 人工データを用いた実験

正規分布を用いた単純な分布モデルを用いた。

4.1.1 認識カテゴリが少ない場合（M=20）

実験結果を図 1 に示す。

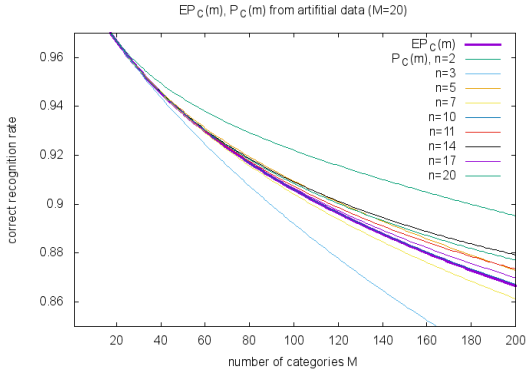

図 1　人工データを用いた推定（M=20）

M=20 までは、正しく近似できている。カテゴリ数 m が大きくなるにしたがい、真の値からのズレが大きくなることが分かる。特に、使用した重み係数が少ない n=2 や n=3 の場合は、真の値からのズレが大きいことが分かる。m=200 において予測精度が最も良かったのはこの中では n=10 の場合で、10M における予測誤差は 0.00067 であった。

4.1.2 認識カテゴリが多い場合（M=100）

実験結果を図 **2** に示す。

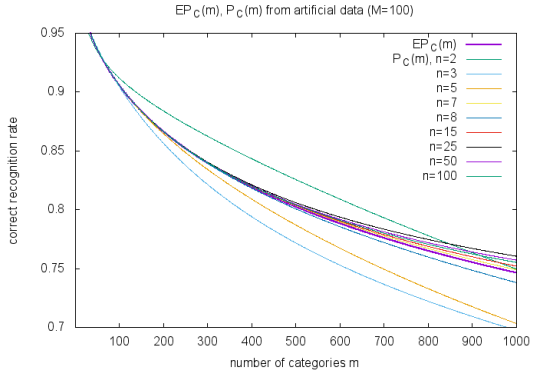

図 2　人工データを用いた推定（M=100）

M=20 の場合とほぼ同様である。m<M であっても、n=2 の場合はズレが大きいように見える。m>M では、真の値からのズレが大きいのは n<6 の場合である。m=1000 において予測精度が最も良かったのはこの中では n=7 の場合で、10M における予測誤差は 0.0028 であった。

4.2 手書き漢字データを用いた実験

現実の多カテゴリ認識問題である手書き漢字認識に適用した。実験結果を図 3 に示す。

今回の人工データの場合よりも、特に n が小さい場合、予測精度が落ちていることが分かる。m=1000 において予測精度が最も良かったのはこの中では n=30 の場合で、予測誤差は 0.012 であった。これを選択できれば、状況により必要とされる精度は異なるが、多くの場合では十分な精度であると考えられる。

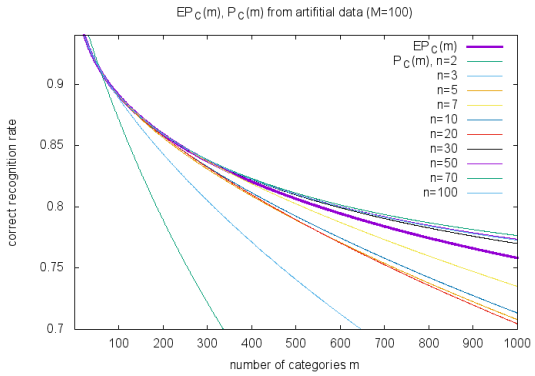

図 3　手書き漢字データを用いた推定（M=100）

5. カテゴリ部分集合の選び方による統計的安定性

大きさ M のカテゴリ部分集合の選び方による影響を人工データの M=100 の場合について調べた。M=100 における平均正解率は 0.905 であり、標準偏差は 3.15e-03 と小さい。ランダムに 5 つの部分集合（t=1,...,5）を発生させ、$P_C(m)$ を求めた。結果を図 4 に示す。

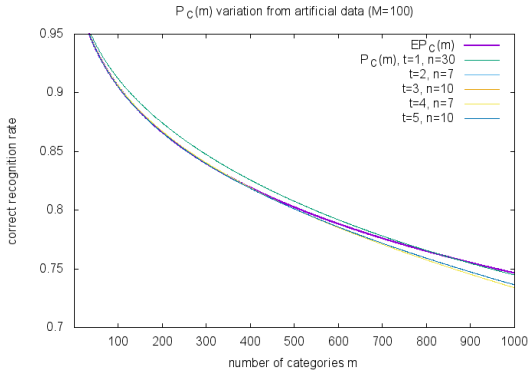

図 4　カテゴリ部分集合の選び方による影響（M=100）

今回は M=100 における正解率の標準偏差が比較的小さいことから、カテゴリ部分集合の変動の影響を少し受けてはいるが、あまり大きな影響を受けていないことが分かる。

6. まとめ

超多カテゴリ認識問題の設計にかかるコストを削減するため、認識対象の一部であるカテゴリ部分集合のみを認識することで、全体の正解率の推定する方法を提案した。また、人工データと手書き漢字データそれぞれに対する認識実験を行い、提案手法の有効性を示した。手書き漢字データでは用いた人工データに比べばらつきと推定誤差が比較的大きいことが分かったが、どちらにおいても適切な推定式が存在することが分かった。今後、最適な n を求める方法を検討したい。また、今回の実験で、部分集合を用いても統計的に安定で、提案手法が有効な認識問題が存在することを示したが、このことは認識対象で大きく異なると予想され、どのくらい大きさの部分集合であれば、統計的に安定するのか、検討していきたい。

参考文献

[1]酒井　充,米田政明, 長谷博行, “多カテゴリ認識問題の理論的考察－期待正解率の効率的計算式の導出－”，信学論, Vol.J79-A, No.11 (1996) .