1D-7

マルコフモデルによる自動分類に対する分類誤り確率の推定

小林　学 †　　後藤正幸 ‡　　松嶋敏泰 ‡　　平澤茂一 ††
† 湘南工科大学　　‡ 早稲田大学　　†† サイバー大学

1　はじめに

与えられた学習データを用いて，新規データがどのカテゴリに所属するかを自動的に判別する自動分類問題は，ベクトル空間モデルやサポートベクターマシン (SVM)[1] などの手法により大きく発展した．一方，J.Ziv らは無ひずみデータ圧縮法である LZ アルゴリズムを用いて文書分類を行う手法を提案した [3]．また圧縮性能の非常に優れた文脈木重みづけ (CTW) 法が提案されており，これを文書分類に適用する手法も提案されている [4]．

本研究ではマルコフモデルを用いた文書分類法を対象とする．ここで各カテゴリに対する文書が同一のマルコフモデルから生起するという仮定を置く．そのとき，新規テスト文書の文書長が有限の場合の分類誤り確率を評価する．具体的には新規テスト文書に対する理想符号長の平均と分散を評価し，中心極限定理を用いることにより各カテゴリの分類誤り確率を近似する手法を提案する．

2　FSMX 情報源

定義 1 整数 a, b に対して本稿では $[a,b] = \{a, a+1, \ldots, b\}$ と定義する．また，ある 2 つのシンボル c と d の連結を cd とつなげて表記する．さらに任意のシンボル系列 $\cdots y_{a-1} y_a \cdots y_b y_{b+1} \cdots$ に対して，$y_a^b = y_a y_{a+1} \cdots y_b$ と定義する．さらにシンボル系列 $y_a^b, a \le b$ に対する系列長を $|y_a^b| = b - a + 1$ と表記する．

次に 2 元情報源からの n ビットの出力を $x_1^n = x_1 x_2 \cdots x_n$ と表記する．ただし任意の $i \in [1,n]$ に対して $x_i \in [0,1]$ である．　□

例えば 2 次マルコフ情報源は，長さ 2 の文脈 $x_{i-2}x_{i-1}$ によりシンボル x_i の生起する確率が定まる．さて FSMX 情報源は，シンボルの生起する確率が文脈により異なり，さらにその文脈の長さも異なるマルコフ情報源である．今情報源を規定する文脈の集合を $\mathcal{S}$ と書く．

例 1 $\mathcal{S} = \{00, 10, 1\}$ とする．このとき情報源シンボル $x_i \in [0,1]$ が生起する確率は，パラメータ $\Theta_{\mathcal{S}} = \{\theta_{00}, \theta_{10}, \theta_1\}$ により次式で規定される．

$$\begin{aligned} P(x_i = 1 | x_{i-1} = 1) &= \theta_1, \\ P(x_i = 1 | x_{i-2}x_{i-1} = 10) &= \theta_{10}, \\ P(x_i = 1 | x_{i-2}x_{i-1} = 00) &= \theta_{00}. \end{aligned} \tag{1}$$

すなわち，文脈が 1 の場合に $x_i = 1$ となる確率は θ_1 であり，文脈が 10 の場合に $x_i = 1$ となる確率は θ_{10} などとなる．　□

このように x_i は 2 元ならば，$\mathcal{S}$ が与えられた下で，x_i に対する文脈が定まるとそのパラメータは 2 項パラメータとなる．

3　FSMX モデルによる文書分類

各文書はその文書が所属するカテゴリが存在するものと仮定する．簡単のためカテゴリには自然数の番号が付いており，カテゴリ数を C で表す．またカテゴリの集合を $\mathcal{C} = [1,C]$ と表す．学習文書はその所属するカテゴリが既知の文書であり，カテゴリ $c \in \mathcal{C}$ に所属する N_c 個の学習文書をそれぞれ $\tilde{x}_1^{(c)}, \tilde{x}_2^{(c)}, \ldots, \tilde{x}_{N_c}^{(c)}$ と表記する．文書分類問題はこれらの学習文書が与えられた下で，カテゴリが未知の新規テスト文書 $z_1^n \in [0,1]^n$ がどのカテゴリに所属するかを推定する問題である．

ここで，カテゴリ $c \in \mathcal{C}$ の文書は文脈の集合 $\mathcal{S}^{(c)}$ を持つ同一の FSMX 情報源から生成されるものと仮定する．また本稿では簡単のため，各 $\mathcal{S}^{(c)}, c \in \mathcal{C}$, は既知と仮定する．今カテゴリ $c \in \mathcal{C}$ に対して，連結した学習文書を $\tilde{x}^{(c)} = \tilde{x}_1^{(c)} \tilde{x}_2^{(c)} \cdots \tilde{x}_{N_c}^{(c)}$ とする．また $\tilde{x}^{(c)}$ 中の文脈 $s \in \mathcal{S}^{(c)}$ の下での 0 の出現回数を $a_s(\tilde{x}^{(c)})$，1 の出現回数を $b_s(\tilde{x}^{(c)})$ とする．このときカテゴリ c に対する FSMX モデルの文脈 $s \in \mathcal{S}^{(c)}$ に対するパラメータのラプラス型推定量は

$$\hat{\theta}_s = \frac{b_s(\tilde{x}^{(c)}) + \frac{1}{2}}{a_s(\tilde{x}^{(c)}) + b_s(\tilde{x}^{(c)}) + 1}, \quad s \in \mathcal{S}^{(c)}, \tag{2}$$

とできる．さて，カテゴリ $c \in \mathcal{C}$ の文書が生起する事前確率を $P(c)$ と表記すると，新規テスト文書 z_1^n の事後確率は式 (2) を用いて

$$P(z_1^n | \tilde{x}^{(c)}) = P(c) \prod_{s \in \mathcal{S}^{(c)}} \hat{\theta}_s^{b_s(z_1^n)} (1 - \hat{\theta}_s)^{a_s(z_1^n)}, \tag{3}$$

となる．結果的に式 (3) を最大とするカテゴリを z_1^n の推定カテゴリとすれば良い．

4　文書分類の誤り確率

さて本節では $\mathcal{C} = [1,2]$ として学習文書数がある程度大きな場合に，前節の文書分類法の分類誤り確率の

An Error Probability Estimation of the Text Classification Using Markov Model
†Manabu KOBAYASHI ‡Masayuki GOTO ‡Toshiyasu MATSUSHIMA ††Shigeichi HIRASAWA
†School of Engineering, Shonan Institute of Technology
‡Faculty of Science and Engineering, Waseda University
††Fuculty of Information Technology and Business, Cyber University

推定値を出力することを目的とする．さて

$$\Delta(z_1^n) \triangleq \log \frac{P(z_1^n|\tilde{x}^{(1)})}{P(z_1^n|\tilde{x}^{(2)})}, \tag{4}$$

と定義すると，新規テスト文書の真のカテゴリが 1 であるときの分類誤り確率 $P_{\mathrm{E}}^{(1)}$ は

$$P_{\mathrm{E}}^{(1)} = P\left(z_1^n \middle| \Delta(z_1^n) \le 0\right), \tag{5}$$

と表される．ただし上の確率はカテゴリ 1 の真のパラメータ集合 $\Theta_{\mathcal{S}^{(1)}}$ で規定される．ここで式 (3),(4) より

$$\begin{aligned}\Delta(z_1^n) &= \sum_{s\in\mathcal{S}^{(1)}} (a_s(z_1^n)\log(1-\hat{\theta}_s^{(1)}) + b_s(z_1^n)\log\hat{\theta}_s^{(1)}) \\ &\quad - \sum_{s\in\mathcal{S}^{(2)}} (a_s(z_1^n)\log(1-\hat{\theta}_s^{(2)}) + b_s(z_1^n)\log\hat{\theta}_s^{(2)}),\end{aligned} \tag{6}$$

が成り立つ．ただし注意しなければならないのは，確率変数は $a_s(z_1^n)$ 及び $b_s(z_1^n)$ である．

さて，式 (6) を用いて $\Delta(z_1^n)$ の平均と分散を求めることを考える．

$$\begin{aligned}E[\Delta(z_1^n)] &= \sum_{s\in\mathcal{S}^{(1)}} (E[(a_s(z_1^n)]\log(1-\hat{\theta}_s^{(1)}) + E[b_s(z_1^n)]\log\hat{\theta}_s^{(1)}) \\ &\quad - \sum_{s\in\mathcal{S}^{(2)}} (E[(a_s(z_1^n)]\log(1-\hat{\theta}_s^{(2)}) + E[b_s(z_1^n)]\log\hat{\theta}_s^{(2)}) \\ &\approx n\sum_{s\in\mathcal{S}^{(1)}} \left(\frac{a_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log(1-\hat{\theta}_s^{(1)}) + \frac{b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\hat{\theta}_s^{(1)}\right) \\ &\quad - n\sum_{s\in\mathcal{S}^{(2)}} \left(\frac{a_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log(1-\hat{\theta}_s^{(2)}) + \frac{b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\hat{\theta}_s^{(2)}\right) \\ &= \hat{E}[\Delta(z_1^n)],\end{aligned} \tag{7}$$

となる．次に 2 次のモーメントを求めるために，$\mathcal{S}^{(1)}, \mathcal{S}^{(2)}$ に対応する完全 2 分木をそれぞれ $M_{\mathcal{S}^{(1)}}, M_{\mathcal{S}^{(2)}}$ と表記し，それぞれのノード及び枝の和集合を取った木を $M_{\mathcal{S}^{(1,2)}} = M_{\mathcal{S}^{(1)}} \cup M_{\mathcal{S}^{(2)}}$ とする．結果的に $M_{\mathcal{S}^{(1,2)}}$ に対応する文脈の集合を $\mathcal{S}^{(1,2)}$ とする[1]．ここで $\forall s \in \mathcal{S}^{(1,2)}$ に対して $a_s(z_1^n)$ 及び $b_s(z_1^n)$ に対する分散・共分散を考える必要があるが，これを厳密に評価するのは大変である．そこで本稿ではこれらを多項分布で近似することを考える．このとき多項分布の分散・共分散を用いると

$$E[a_s(z_1^n)^2] \approx nq_s^{(1)}(1-\theta_s^{(1)})(1-q_s^{(1)}(1-\theta_s^{(1)})), \tag{8}$$

$$E[a_s(z_1^n)b_s(z_1^n)] \approx -n(q_s^{(1)})^2(1-\theta_s^{(1)})\theta_s^{(1)}, \tag{9}$$

$$E[a_s(z_1^n)a_{s'}(z_1^n)] \approx -nq_s^{(1)}(1-\theta_s^{(1)})q_{s'}^{(1)}(1-\theta_{s'}^{(1)}), \quad s \ne s', \tag{10}$$

などと簡単になる．ただし $q_s^{(1)}$ はカテゴリ 1 の情報源における文脈 s に対する定常確率である．従って

$$\begin{aligned}&E[(\Delta(z_1^n)^2] \\ &= E\left[\left\{\sum_{s\in\mathcal{S}^{(1,2)}} \left(a_s(z_1^n)\log\frac{1-\hat{\theta}_s^{(1)}}{1-\hat{\theta}_s^{(2)}} + b_s(z_1^n)\log\frac{\hat{\theta}_s^{(1)}}{\hat{\theta}_s^{(2)}}\right)\right\}^2\right] \\ &\approx n\sum_{s\in\mathcal{S}^{(1,2)}} \left\{\frac{a_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\left(1-\frac{a_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\right)\left(\log\frac{1-\hat{\theta}_s^{(1)}}{1-\hat{\theta}_s^{(2)}}\right)^2\right. \\ &\quad -2\frac{a_s(\tilde{x}^{(1)})b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|^2}\log\frac{1-\hat{\theta}_s^{(1)}}{1-\hat{\theta}_s^{(2)}}\log\frac{\hat{\theta}_s^{(1)}}{\hat{\theta}_s^{(2)}} \\ &\quad \left.+\frac{b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\left(1-\frac{b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\right)\left(\log\frac{\hat{\theta}_s^{(1)}}{\hat{\theta}_s^{(2)}}\right)^2\right\} \\ &\quad -n\sum_{s,s'\in\mathcal{S}^{(1,2)}|s\ne s'} \left\{\frac{a_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\frac{1-\hat{\theta}_s^{(1)}}{1-\hat{\theta}_s^{(2)}} + \frac{b_s(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\frac{\hat{\theta}_s^{(1)}}{\hat{\theta}_s^{(2)}}\right\} \\ &\quad \times\left\{\frac{a_{s'}(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\frac{1-\hat{\theta}_{s'}^{(1)}}{1-\hat{\theta}_{s'}^{(2)}} + \frac{b_{s'}(\tilde{x}^{(1)})}{|\tilde{x}^{(1)}|}\log\frac{\hat{\theta}_{s'}^{(1)}}{\hat{\theta}_{s'}^{(2)}}\right\} \\ &= \hat{E}[(\Delta(z_1^n)^2],\end{aligned} \tag{11}$$

と計算可能である．$\Delta = \Delta(z_1^n)$ となる確率密度は $\mu = \hat{E}[\Delta(z_1^n)]$, $\sigma^2 = \hat{E}[(\Delta(z_1^n)^2] - \hat{E}[\Delta(z_1^n)]^2$ と置いて，中心極限定理より正規分布を用いて $P(\Delta) \approx \mathcal{N}(\Delta|\mu,\sigma^2)$ と近似できる．従って式 (4),(5) より

$$P_{\mathrm{E}}^{(1)} \approx \int_{-\infty}^{0} \mathcal{N}(\Delta|\mu,\sigma^2)d\Delta, \tag{12}$$

と近似することが可能である．

さらに各カテゴリの学習文書に対するパラメータ推定において，$a_s(\tilde{x}^{(c)}), b_s(\tilde{x}^{(c)})$ 及び $\hat{\theta}_s^{(c)}$ に対して重複対数の法則を用いることもできる．そのとき，$P_{\mathrm{E}}^{(1)}$ の上界と下界の推定値を得ることも可能である．

5 まとめ

本研究ではマルコフモデルを用いた文書分類法を対象とし，2 値分類問題に対する分類誤り確率の推定法について検討を行った．このとき新規テスト文書に対する理想符号長の平均と分散を評価し，中心極限定理を用いることにより分類誤り確率を近似する手法を提案した．

参考文献

[1] C.M. ビショップ，パターン認識と機械学習，Springer，2008.

[2] H. Chen, P. Ti and X. Yao, “Probabilistic Classification Vector Machines,” IEEE Trans. On Neural Networks, vol.20, no.6, pp.902-914, June 2009.

[3] J. Ziv and N. Merhav,“A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification,” IEEE Trans. on Information Theory, vol.39, no.4, pp.1270–1279, July 1993.

[4] Z. Dawy, J. Hagenauer and A. Hoffmann,“Implementing the context tree weighting method for content recognition,” Proc. of the IEEE Data Compression Conference, p.536, March 2004.

[5] H. Cai, S. R. Kulkarni and S. Verdú,“Universal divergence estimation for finite-alphabet sources,” IEEE Trans. on Information Theory, vol.52, no.8, pp.3456–3475, Aug. 2006.

[6] D. Benedetto, E. Caglioti, and V. Loreto,“ Language Trees and Zipping, ” Physical Review Letters, vol. 88, no.4, p.28, Jan. 2002.

[1] すなわち $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}$ それぞれにおける長い方の文脈を取ったものの集合が $\mathcal{S}^{(1,2)}$ である．