3B-04

文字列クラスタリングのための Laplace 様混合モデルに対する EM アルゴリズム *

小谷野 仁 †　　林田 守広 ‡
京都大学 大学院医学研究科　京都大学 化学研究所

近年, テキストデータや生物配列など, 生成される文字列データの量が飛躍的に増加し, 文字列データの解析法が様々な領域で求められている. 本研究において, 我々は, 特に生物配列への応用を念頭において, 混合モデルと EM アルゴリズムを用いた, 文字列データの教師無し多クラス分類法の開発に取り組んだ.

$A = \{a_1, \cdots, a_z\}$ をアルファベットとし, A 上の文字列の全体を A^* によって表す. 遺伝子配列を扱うならば, $A = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ である. $\mathfrak{D}$ を A^* 上の距離の集合とする. $\mathfrak{D}$ の元として, 拡張 Hamming 距離 (以下 $d_{H'}$ によって表す), Jaro–Winkler 距離, 最長共通部分列距離, Levenshtein 距離, Damerau–Levenshtein 距離などがあり, それぞれが A^* 上に距離位相を定める. また, A^* は連接を内部算法として半群をなし, 位相半群となる. このような位相構造と代数構造を持つ集合 A^* 上で確率論を展開して, ランダムに生成された文字列を扱うための理論を作ることを試みた研究として [2–4] がある. 本研究において, 我々は, まず A^* 上にパラメトリックな分布を導入して, その基本的な性質を調べ, 次にそのパラメーターの最尤推定量を求めて, 導入した分布の混合モデルに対する EM アルゴリズムを構成する. これらの準備の後, [2–4] において作られた確率文字列の理論の枠組みとその中で証明されたいくつかの極限定理を用いることにより, 構成した最尤推定量とアルゴリズムの精度を漸近理論の枠組みで研究し, 推定量とアルゴリズムの精度が保証されるための条件を述べる.

任意の $\lambda \in A^*$, $\rho \in (0, \infty)$ 及び $d \in \mathfrak{D}$ に対して, 関数 $q_d(\,\cdot\,; \lambda, \rho) : A^* \to [0, 1]$ を

$$q_d(s; \lambda, \rho) = \frac{1}{(\rho+1)\big|U(\lambda, d(s, \lambda))\big|} \left(\frac{\rho}{\rho+1}\right)^{d(s,\lambda)}$$

によって定め, 集合関数 $Q_d(\,\cdot\,; \lambda, \rho) : 2^{A^*} \to [0, 1]$ を $Q_d(E; \lambda, \rho) = \sum_{s \in E} q_d(s; \lambda, \rho)$ によって定義する. そうすると, $Q_d(\,\cdot\,; \lambda, \rho)$ は可測空間 $(A^*, 2^{A^*})$ 上の確率測度になることが確かめられる. 以下で述べるように, $Q_d(\,\cdot\,; \lambda, \rho)$ は, その確率関数は大きく異なる形を持つけれども, $\mathbb{R}$ 上の Laplace 分布と類似の性質を持つ. そこで, $Q_d(\,\cdot\,; \lambda, \rho)$ を A^* 上の Laplace 様分布と呼ぶことにする. 以下, q の添え字 d は省略する.

A^* 上の Laplace 様分布は次の性質を持つ: (i) 分布の位置と散らばりを表す 2 つのパラメーター λ と ρ を持つ. (ii) 確率関数 $q(s; \lambda, \rho)$ は λ において最大値をとり, $d(s, \lambda)$ が大きくなるに従って減少し (よって単峰), $d(s, \lambda)$ に関して対称である. (iii) 特に, $q(s; \lambda, \rho)$ は, $d(s, \lambda)$ が大きくなるに従って指数的に減少し, 正規分布と異なり, 変曲点を持たない. (iv) ある固定された文字列のまわりの 1 次絶対モーメントが所与の正の数と等しいという条件を満たす A^* 上の分布の族の中で最大のエントロピーを持つ.

更に, 次の結果が得らえる: (v) $s_1, \cdots, s_n \in A^*$ に対して, A^* 上の Laplace 様分布の位置と散らばりのパラメーター λ と ρ の最尤推定量は, それぞれ

$$\hat{\lambda}(s_1, \cdots, s_n) = \boldsymbol{m}(s_1, \cdots, s_n), \tag{1}$$

$$\hat{\rho}(s_1, \cdots, s_n) = \frac{1}{n} \sum_{i=1}^{n} d(s_i, \boldsymbol{m}(s_1, \cdots, s_n)) \tag{2}$$

によって与えられる. すなわち, 中央文字列と中央文字列からの絶対偏差の平均である. 特に $d = d_{H'}$ の時, $\hat{\lambda}(s_1, \cdots, s_n) = \boldsymbol{m}_c(s_1, \cdots, s_n)$ であって, コンセンサス配列となる. $\boldsymbol{m}$ と $\boldsymbol{m}_c$ の厳密な定義については, [3] の supplemental material を参照して欲しい.

*EM algorithm for a Laplace-like mixture for string clustering

†Hitoshi Koyano, Graduate School of Medicine, Kyoto University
‡Morihiro Hayashida, Institute for Chemical Research, Kyoto University

$s_1, \cdots, s_n$ を, 未知パラメーター $\theta = (\pi_1, \cdots, \pi_k, \lambda_1, \cdots, \lambda_k, \rho_1, \cdots, \rho_k)$ を持つ A^* 上の Laplace 様分布の混合モデル

$$q(s;\theta) = \sum_{g=1}^{k} \pi_g q_g(s;\lambda_g,\rho_g) \tag{3}$$

に従って分布する母集団からの n 個の観測文字列とする. 各 $i = 1, \cdots, n$ と $j \in \mathbb{Z}^+$ に対して, s_i の第 j 文字を x_{ij} とし, s_i の長さを $|s_i|$ によって表す. 各 $g = 1, \cdots, k$ に対して, s_i が第 g 部分母集団からのものである確率を z_{ig} によって表す.

部分母集団の番号 g を任意に選んで固定する. $\ell = \max\{|s_1|, \cdots, |s_n|\}$ とおき, 各 $i = 1, \cdots, n$ に対して $\hat{z}_{ig}$ を z_{ig} の何らかの推定値とし, 各 $j = 1, \cdots, \ell$ と $h = 1, \cdots, z$ に対して $f_{gh}(j) = \sum_{i \in \{i' \in \{1,\cdots,n\} : x_{i'j} = a_h\}} \hat{z}_{ig}$ とおく. $f_{gh}(j)$ は, 第 g 部分母集団から抽出された文字列の第 j 文字がアルファベット中の第 h 文字である確率の推定値である. 第 g 部分母集団の文字列の第 j 文字として最も高い確率で出現すると推定されるアルファベット中の文字の番号を $h_g(j)$ とする. すなわち, $h_g(j) = \arg\max_{1 \le h \le z} f_{gh}(j)$. そうして, 各 $g = 1, \cdots, k$ に対して, 第 g 部分母集団の分布の位置パラメーターの推定量を

$$\hat{\lambda}_g = a_{h_g(1)} \cdots a_{h_g(\ell)} e \cdots \tag{4}$$

とおく. ここで, e は空文字を表している. この時, 各 $g = 1, \cdots, k$ と $j \in \mathbb{Z}^+$ に対して, $h_g(j)$ が一意に定まるならば, $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ は, $t_1, \cdots, t_k \in A^*$ に関する拡張 Hamming 距離の加重和 $\sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} d_{H'}(s_i, t_g)$ の最小化問題の解であることを示せる. すなわち, $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ は拡張 Hamming 距離に関する中央文字列である.

この結果を使うと, $d = d_{H'}$ を持つ混合モデル (3) のパラメーターを推定するための EM アルゴリズムを導ける. 各 $i = 1, \cdots, n$, $g = 1, \cdots, k$ 及び $t \in \mathbb{N}$ に対して, $\hat{z}_{ig}^{(t)}, \hat{\pi}_g^{(t)}, \hat{\lambda}_g^{(t)}$ 及び $\hat{\rho}_g^{(t)}$ を, それぞれステップ t において得られた z_{ig}, π_g, λ_g 及び ρ_g の推定値とする. E ステップにおける $\hat{z}_{ig}^{(t)}$ の更新方法と M ステップにおける $\hat{\pi}_g^{(t)}$ の更新方法は, $\mathbb{R}^p$ 上の通常の EM アルゴリズムと変らない. 各 $g = 1, \cdots, k$ に対して, M ステップにおいて, 混合成分の位置パラメーター λ_g の推定量は, $\hat{z}_{ig}$ として $\hat{z}_{ig}^{(t)}$ を用い, 式 (4) に従って更新する. $\hat{z}_{ig}$ として $\hat{z}_{ig}^{(t)}$ を用いた $\hat{\lambda}_g$ を $\hat{\lambda}_g^{(t)}$ によって表す. また, 散らばりのパラメーター ρ_g の推定量は,

$$\hat{\rho}_g^{(t)} = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}^{(t)}} \sum_{i=1}^{n} \hat{z}_{ig}^{(t)} d_{H'}(s_i, \hat{\lambda}_g^{(t)}) \tag{5}$$

に従って更新する.

最尤推定量の漸近有効性を保証する正則条件には色々なものがあるが (有名なものとしては, 例えば [5, 6]), 推定量 (1) と (2) には適用できない. 本発表では, [2] において示された極限定理を用いることにより, 推定量 (1) と (2) の強一致性を保証する条件が, またそれと [7] の結果を組み合せることにより, 更新式 (4) と (5) によって得られる EM アルゴリズムからの推定値の列により, 混合モデル (3) の各混合成分のパラメーターが強一致推定されるための条件が得られることを述べる. また, [1] により, $\mathbb{R}$ 上の通常の中央値の素朴な拡張として, 平均絶対偏差の最小解として A^* 上に導入された中央文字列の定義を再考することを通して, (v) の結果を得るまでの経緯を述べる.

引用文献

[1] T. Kohonen. Median strings. *Pattern Recogn. Lett.*, 3(5):309–313, 1985.

[2] H. Koyano, M. Hayashida, and T. Akutsu. Maximum margin classifier working in a set of strings. *arXiv:1406.0597v2*, 2014. `http://jp.arxiv.org/abs/1406.0597`.

[3] H. Koyano and H. Kishino. Quantifying biodiversity and asymptotics for a sequence of random strings. *Phys. Rev. E*, 81(6):061912, 2010.

[4] H. Koyano, T. Tsubouchi, H. Kishino, and T. Akutsu. Archaeal β diversity patterns under the seafloor along geochemical gradients. *J. Geophys. Res. G*, 119(9):1770–1788, 2014.

[5] M. D. Perlman. On the strong consistency of approximate maximum likelihood estimators. In L. M. Le Cam, J. Neyman, and E. L. Scott, editors, *Proc. 6th Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 263–281, Berkeley, CA, 1972. University of California Press.

[6] A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.*, 29:595–601, 1949.

[7] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.