Vol.2015-MPS-103 No.31
Vol.2015-BIO-42 No.31
2015/6/24

文字列の集合上の Laplace 様混合モデルと EM アルゴリズムに基づく文字列クラスタリング

小谷野 仁[1,a)]　林田 守広[2,b)]　阿久津 達也[2,c)]

概要：本稿において, 我々は, [1], [2], [3] において作られてきた文字列の集合上の確率論を用いて, 文字列データに対する混合モデルと EM アルゴリズムの理論を展開することにより, 文字列データの教師無しクラスタリングの問題に取り組む. 我々は, まず, 実数の集合上の Laplace 分布をモチーフにして作られたパラメトリックな分布を文字列の集合上に導入し, その基本的な性質を調べる. この Laplace 様分布は, 位置を表す文字列のパラメーターと散らばりを表す正の実数のパラメーターを持つが, 一方のパラメーターが文字列であるため, 最尤推定量を陽に書くのが難しい. そこで, 我々は, 観測文字列の数が増加するに従って, 最尤推定量に概収束する推定量を構成し, 更にそれらによってパラメーターが強一致推定されることを示す. その後, 我々は, Laplace 様分布の混合モデルのパラメーターを推定するための反復アルゴリズムを構成し, 観測文字列の数とアルゴリズムの反復回数が増加するに従って, そのアルゴリズムが EM アルゴリズムに概収束すること, またそれにより Laplace 様混合モデルのパラメーターが強一致推定されることを証明する. 我々は, 最後に, この Laplace 様混合モデルから文字列の教師無しクラスタリング方式を導出し, それが正しい分類を行う事後確率が最大であるという意味で漸近的に最適であることを述べる.

キーワード：文字列, 教師無しクラスタリング, 確率論, 統計的漸近理論, Laplace 様分布, 混合モデル, EM アルゴリズム

String Clustering Based on a Laplace-like Mixture and EM Algorithm on a Set of Strings

Hitoshi Koyano[1,a)]　Morihiro Hayashida[2,b)]　Tatsuya Akutsu[2,c)]

Abstract: In this study, by developing a theory of a mixture model and EM algorithsm for string data on the basis of a probability theory on a set of strings developed in [1], [2], [3], we address the problem of clustering string data in an unsupervised manner. We first construct a parametric distribution on a set of strings in the motif of the Laplace distribution on a set of real numbers and reveal its basic properties. This Laplace-like distribution has two parameters one of which is a string that represents the location of the distribution and another is positive real number that represents the dispersion. It is difficult to explicitly write maximum likelihood estimators of the parameters because one parameter is a string. We construct estimators that almost surely converge to the maximum likelihood estimators as the number of observed strings increases and demonstrate that the estimators strongly consistently estimate the parameters. After that, we compose an iteration algorithm for estimating parameters of the mixture model of the Laplace-like distributions and demonstrate that the algorithm almost surely converges to the EM algorithm for the Laplace-like mixture and strongly consistently estimates its parameters as the numbers of observed strings and of iterations increase. We finally derive a procedure for unsupervised sring clustering from the Laplace-like mixture that is asymptotically optimal in the sense that the posterior probability of making correct classifications is maximized.

Keywords: Strings, unsupervised clustering, probability theory, statistical asymptotics, Laplace-like distributions, mixture models, EM algorithm

Vol.2015-MPS-103 No.31
Vol.2015-BIO-42 No.31
2015/6/24

1. はじめに

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, 文字列データの統計的な解析方法が様々な領域で求められている. 数や数ベクトルとして表されるデータに対する統計学は, データはある確率法則に従って観測された母集団の一部であるということを考慮してデータを解析し, 母集団に関して推測するために, 実数の集合や実ベクトル空間上の確率論に基づいて厳密に作られている. 同様に, 文字列データに対する統計的方法も文字列の集合上の確率論に基づいて作られるのが良いだろう. 本稿において, 我々は, [1], [2], [3] において作られてきた文字列の集合上の確率論を応用することにより, 文字列の集合上で混合モデルとそれに対する EM アルゴリズムの理論を展開し, 混合モデルに基づいた, 文字列データの教師無しクラスリング方法の開発に取り組む. 特に, 交差検証法を用いた数値的な仕方でではなく, 確率論を用いた理論的な仕方で提案方法の性能を評価することにこだわる.

2. 文字列の集合上の Laplace 様分布

$A = \{a_1, \cdots, a_{z-1}\}$ を $z-1$ 個の文字からなるアルファベットとする. 空文字 e に対して $a_z = e$ とおき, $\bar{A} = \{a_1, \cdots, a_z\} = A \cup \{e\}$ を拡張アルファベットと言うことにする. 空文字列を o によって表す. すなわち, $o = e\cdots$. 本稿では, A の元の有限列の末尾に空文字の無限列がつなげられたものとして, A 上の文字列を定義する. 文字列をこの仕方で定義すると, 様々な長さの文字列を実現し得るものとして, 確率文字列が自然に定義される. A 上の文字列の集合を A^* によって表す. 本稿における我々の対象は, A^* に値をとる離散確率過程である確率文字列, その分布, 及びその実現値である文字列データである. 以下では, [1], [2], [3] において特に DNA の解析のために作られてきた文字列の集合 A^* 上の確率論の基本的な枠組みといくつかの結果を使う.

A^* 上の距離関数として自然数値のもののみを考え, A^* 上の距離の集合を D によって表す. D の元として, Jaro–Winkler 距離, 最長共通部分列距離, Levenshtein 距離 (d_L と表す), Damerau–Levenshtein 距離などがある. 本稿では, 文字列の末尾の連続する文字の削除をそれらの文字の空文字への置換と見なし, 文字列の末尾への文字の挿入を文字列の末尾につなげられている空文字のそれらの文字への置換と見なす. ある文字列を他の文字列に変形するのに必要な, この意味での置換の最少回数を拡張 Hamming 距離と言い, $d_{H'}$ によって表す. 通常の Hamming 距離 (d_H と表す) は等しい長さの文字列の間でのみ定義されるから, 数学的な意味では A^* 上の距離でないが, $d_{H'}$ は数学的な意味での A^* 上の距離であって, 最も少ない種類の編集操作が許された距離と見なせる. $s \in A^*$ と $r \in \mathbb{N}$ (自然数の集合) に対して, $U(s,r) = \{t \in A^* : d(s,t) \leq r\}$ 及び $\partial U(s,r) = \{t \in A^* : d(s,t) = r\}$ と定める. 集合 S の元の数を $|S|$, べき集合を 2^S によって表す. $s \in A^*$ を構成する A の元の数を s の長さと言い, $|s|$ によって表す.

A^* 上にパラメトリックな分布を導入することから始める. 任意の $\lambda \in A^*$, $\rho \in (0,\infty)$ 及び $d \in D$ に対して, 関数 $q_d(\,\cdot\,;\lambda,\rho) : A^* \to [0,1]$ を

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

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.

$Q_d(\,\cdot\,;\lambda,\rho)$ 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 $q_d(s;\lambda,\rho)$ 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), λ に関して対称である. (iii) 特に, $q_d(s;\lambda,\rho)$ 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) 確率文字列 σ が $Q_d(\,\cdot\,;\lambda,\rho)$ に従って分布している時, $d = d_{H'}$ ならば, σ の中央文字列 ($\boldsymbol{M}(\sigma)$ と表す) は λ と一致し, 任意の $d \in D$ に対して, σ の $\boldsymbol{M}(\sigma)$ の周りの平均絶対偏差 ($\boldsymbol{\Upsilon}(\sigma)$ と表す) は ρ に一致する. (v) ある固定された文字列の周りの 1 次絶対モーメントが所与の正の実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. (vi) $d = d_{H'}$ ならば, 観測を繰り返す時, パラメーター λ と ρ の最尤推定量は, それぞれ標本の中央文字列とその周りの平均絶対偏差に漸近的に等しい. このように, $Q_d(\,\cdot\,;\lambda,\rho)$ は, $\mathbb{R}$ (実数の集合) 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し, $\mathrm{L}_{A^*}(\lambda,\rho)$ によって表す. 以下では, q と Q の添え字 d は省略する.

3. Laplace 様分布のパラメーターの推定

任意の言明 S に対して S a.s. は S が確率 1 で成り立つことを, $\xrightarrow{\text{a.s.}}$ は概収束を表す. $\mathcal{M}(\Omega, A^*)$ は, ある確率空間 $(\Omega, \mathfrak{F}, P)$ 上で定義されて, A^* に値をとる確率文字列の集合, $[\mathcal{M}(\Omega, A^*)^n]$ はコンセンサス配列が一意に定まる n 個の確率文字列の組の集合である. $d = d_{H'}$ の時, 確率文

[1] 京都大学大学院医学研究科医学統計生物情報学領域
Laboratory of Biostatistics and Bioinformatics, Graduate School of Medicine, Kyoto University, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan
[2] 京都大学化学研究所数理生物情報学領域
Laboratory of Mathematical Bioinformatics, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[a)] koyano@kuhp.kyoto-u.ac.jp
[b)] morihiro@kuicr.kyoto-u.ac.jp
[c)] takutsu@kuicr.kyoto-u.ac.jp

Vol.2015-MPS-103 No.31
Vol.2015-BIO-42 No.31
2015/6/24

字列 σ の中央文字列 $\boldsymbol{M}(\sigma)$ はコンセンサス配列 $\boldsymbol{M}_c(\sigma)$ と等しい. 従って, 前節の第 4 段落の (iv) より, 任意の $d \in D$ の下で, ρ を標本の平均絶対偏差 $\sum_{i=1}^n d(s_i,\lambda)/n$ によって (λ が未知であるならば, λ はその適当な推定値に置き換える), また $d = d_{H'}$ の時, λ を標本のコンセンサス配列 $\boldsymbol{m}_c(s_1,\cdots,s_n)$ によって推定するのは自然だろう. $\{\sigma_i = \{\alpha_{ij} : j \in \mathbb{Z}^+\} : i \in \mathbb{Z}^+\} \subset \mathcal{M}(\Omega, A^*)$ とし, 各 $h = 1,\cdots,z$ に対して

$$p(i,j,h) = P\big(\{\omega \in \Omega : \alpha_{ij}(\omega) = a_h\}\big),$$
$$\bar{p}(j,h,n) = \frac{1}{n}\sum_{i=1}^n p(i,j,h)$$

と定める. $p(i,j,h)$ は, i 番目の確率文字列の j 番目の文字が拡張アルファベット $\bar{A}$ の中の h 番目の文字を実現する確率であって, $\bar{p}(j,h,n)$ は, n 回の観測を行った時に $\bar{A}$ の中の h 番目の文字が観測される平均的な確率である. [3] の系 4.2 と 4.3 を使うと, 次の結果を得る.

命題 1. 確率文字列 $\sigma_1 = \{\alpha_{1j}\},\cdots,\sigma_n = \{\alpha_{nj}\}$ の実現値に基づいて, $d = d_{H'}$ を持つ $\mathrm{L}_{A^*}(\lambda,\rho)$ のパラメーター λ を推定する問題において, (i) 各 $j \in \mathbb{Z}^+$ に対して $\alpha_{1j},\cdots,\alpha_{nj}$ が独立で, (ii) 各 $n \in \mathbb{Z}^+$ に対して $(\sigma_1,\cdots,\sigma_n) \subset [\mathcal{M}(\Omega,A^*)^n]$ であって, (iii) $\iota(j) = \arg\max_{1\le h\le z}\bar{p}(j,h,n)$ が n に依らず一意に定まり, (iv) $\{a_{\iota(j)} : j \in \mathbb{Z}^+\} = \lambda$ が成り立つならば, $N_0 \in \mathbb{Z}^+$ が存在して, $n \ge N_0$ ならば, $\boldsymbol{m}_c(s_1,\cdots,s_n) = \lambda$ a.s.

命題 2. 確率文字列 $\sigma_1,\cdots,\sigma_n$ の実現値 $d = d_{H'}$ を持つ $\mathrm{L}_{A^*}(\lambda,\rho)$ のパラメーター ρ を推定する問題において, (i) $\sigma_1,\cdots,\sigma_n$ が独立で, (ii) 各 $n \in \mathbb{Z}^+$ に対して $(\sigma_1,\cdots,\sigma_n) \in [\mathcal{M}(\Omega,A^*)^n]$ であって, (iii) $\sigma_1,\cdots,\sigma_n \in [\mathcal{M}(\Omega,A^*)]$ が成り立ち, (iv) $\sigma_1,\cdots,\sigma_n$ が同一の有限次元分布族を持つならば, $n \longrightarrow \infty$ の時,

$$\frac{1}{n}\sum_{i=1}^n d_{H'}(s_i, \boldsymbol{m}_c(s_1,\cdots,s_n)) \xrightarrow{\text{a.s.}} \rho.$$

すなわち, $\sum_{i=1}^n d_{H'}(s_i,\boldsymbol{m}_c(s_1,\cdots,s_n))/n$ は ρ を強一致推定する.

よって, λ と ρ 強一致推定量が得られたが, これらの推定量は λ と ρ の最尤推定量だろうか. まず, 標本平均絶対偏差 $\sum_{i=1}^n d(s_i,\lambda)/n$ と ρ の最尤推定量の間には, 次の関係がある.

命題 3. 任意の $d \in D$ に対して, $\mathrm{L}_{A^*}(\lambda,\rho)$ のパラメーター ρ の最尤推定量は

$$\check{\rho}(s_1,\cdots,s_n) = \frac{1}{n}\sum_{i=1}^n d(s_i,\lambda)$$

によって与えられる. λ が未知である場合には, 上式の左辺の λ はその適当な推定量に置き換える.

次に, λ を最尤推定する問題を考える. $|\partial U(\lambda, d(s_i,\lambda))| \ge 1$ と $\rho/(\rho+1) < 1$ より, ρ が与えられている時,

$$F(\lambda,\rho) = -\sum_{i=1}^n \log\big|\partial U(\lambda, d(s_i,\lambda))\big| + \log\left(\frac{\rho}{\rho+1}\right)\sum_{i=1}^n d(s_i,\lambda). \quad (2)$$

を最小化する λ を求めたい. λ の最尤推定値を決定する関数 $F(\lambda,\rho)$ が ρ にも依存し, 命題 3 より, ρ の最尤推定量は λ に依存する. また, 文字列球の大きさの一般的な公式を求める問題は未解決問題であるから, 式 (2) の右辺の 2 つの項を同時に考慮して λ に関する最小化問題を解くのは難しい. しかし, $d = d_L$ の時には, $\sum_{i=1}^n d_L(s_i,\lambda)$ の最小解の近似アルゴリズムは知られている. 従って, $d = d_L$ ならば, 次のアルゴリズムによって式 (2) の最小化問題の近似解を探すのは自然だろう.

1 既存のアルゴリズム (例えば [4]) を用いて,

$$\check{\lambda}^{(0)} = \arg\min_{\lambda\in A^*}\sum_{i=1}^n d_L(s_i,\lambda)$$

を求める.

2

$$\check{\rho}^{(0)} = \frac{1}{n}\sum_{i=1}^n d_L(s_i,\check{\lambda}^{(0)}), \quad F_*^{(0)} = F(\check{\lambda}^{(0)},\check{\rho}^{(0)})$$

を計算する.

3 $t = 1,2,\cdots$ に対して,

3.1 $\{s_\gamma : \gamma \in \Gamma^{(t-1)}\} = \partial U(\check{\lambda}^{(t-1)},1)$ と定め, 各 $\gamma \in \Gamma^{(t-1)}$ に対して

$$v_\gamma = \frac{1}{n}\sum_{i=1}^n d_L(s_i,s_\gamma), \quad F(s_\gamma,v_\gamma)$$

を計算する.

3.2 $\gamma \in \Gamma^{(t-1)}$ が存在して, $F(s_\gamma,v_\gamma) < F_*^{(t-1)}$ が成り立つならば,

$$\gamma^* = \arg\min_{\gamma\in\Gamma^{(t-1)}} F(s_\gamma,v_\gamma), \quad \check{\lambda}^{(t)} = s_{\gamma^*},$$
$$\check{\rho}^{(t)} = v_{\gamma^*}, \quad F_*^{(t)} = F(\check{\lambda}^{(t)},\check{\rho}^{(t)}), \quad t = t+1$$

と定めて, ステップ 3.1 に戻る. そうでないならば, 反復を止めて,

$$\check{\lambda} = \check{\lambda}^{(t-1)}, \quad \check{\rho} = \check{\rho}^{(t-1)}$$

を返す.

$d = d_{H'}$ の時には, $m_c(s_1,\cdots,s_n)$ と λ の最尤推定量の間には, 複雑で興味深い関係がある. この場合には, 十分に多くの観測文字列が与えられているならば, λ と ρ の最尤推定量を陽に書ける.

命題 4. $d = d_{H'}$ を持つ $\mathrm{L}_{A^*}(\lambda,\rho)$ に従って独立

に分布する確率文字列 $\sigma_1,\cdots,\sigma_n$ の実現値 $s_1,\cdots,s_n$ に基づいて λ と ρ を推定する問題において, (i) 各 $n\in\mathbb{Z}^+$ に対して $(\sigma_1,\cdots,\sigma_n)\subset[\mathcal{M}(\Omega,A^*)^n]$ であって, (ii) $\iota(j)=\arg\max_{1\le h\le z}\bar{p}(j,h,n)$ が n に依らず一意に定まり, (iii) $\{a_{\iota(j)}:j\in\mathbb{Z}^+\}=\lambda$ が成り立つならば, $N_0\in\mathbb{Z}^+$ が存在して, $n\ge N_0$ ならば,

$$\check{\lambda}(s_1,\cdots,s_n)=\boldsymbol{m}_c(s_1,\cdots,s_n)\quad\text{a.s.},$$
$$\check{\rho}(s_1,\cdots,s_n)=\frac{1}{n}\sum_{i=1}^{n}d_{H'}(s_i,\boldsymbol{m}_c(s_1,\cdots,s_n))\quad\text{a.s.}$$

4. Laplace 様混合モデルの推定アルゴリズム

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

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

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. このモデルのパラメーター空間は $\Theta=(0,1)^k\times(A^*)^k\times(0,\infty)^k$ である. $s_1,\cdots,s_n$ に基づいて $\boldsymbol{\theta}$ を推定するための反復アルゴリズムを構成する.

各 $i=1,\cdots,n$ に対して i 番目の観測文字列を $s_i=\{x_{ij}\in\bar{A}:j\in\mathbb{Z}^+\}$ によって表す. s_i は確率文字列 σ_i の実現値であるとする. 各 $g=1,\cdots,k$ に対して, k 次元実ベクトル $\boldsymbol{w}_g=(w_{g1},\cdots,w_{gk})$ を, $w_{gg}=1$ 及び $g'\neq g$ に対して $w_{gg'}=0$ によって定義し, $W=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_k\}$ とおく. $\boldsymbol{Z}_i=(Z_{i1},\cdots,Z_{ik})$ を W に値をとる k 次元潜在確率ベクトルとする. 各 $g=1,\cdots,k$ に対して $\boldsymbol{Z}_i$ の分布の確率関数を

$$P(\boldsymbol{Z}_i=\boldsymbol{w}_g)=\prod_{g'=1}^{k}\pi_{g'}^{w_{gg'}}$$

によって定義する. $P(\boldsymbol{Z}_i=\boldsymbol{w}_g)=\pi_g$ より, $\boldsymbol{Z}_i=\boldsymbol{w}_g$ と $P(\boldsymbol{Z}_i=\boldsymbol{w}_g)$ は, それぞれ i 番目の文字列が g 番目の部分母集団から抽出されるという事象とその確率を表していると解釈される. $\sigma_1(\omega)=s_1,\cdots,\sigma_n(\omega)=s_n$ が与えられている時の $\boldsymbol{Z}_i$ の条件付き分布の確率関数は,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Z}_i=\boldsymbol{w}_g|\sigma_1(\omega)=s_1,\cdots,\sigma_n(\omega)=s_n)=\frac{\pi_g q(s_i|\lambda_g,\rho_g)}{\sum_{g'=1}^{k}\pi_{g'}q(s_i|\lambda_{g'},\rho_{g'})}$$

と計算される. $\boldsymbol{\theta}$ のある推定量 $\hat{\boldsymbol{\theta}}$ に対して,

$$\zeta_{ig}=\mathrm{E}_{\boldsymbol{\theta}}[Z_{ig}|\sigma_1(\omega)=s_1,\cdots,\sigma_n(\omega)=s_n],\tag{3}$$
$$\hat{\zeta}_{ig}=E_{\hat{\boldsymbol{\theta}}}(Z_{ig}|\sigma_1(\omega)=s_1,\cdots,\sigma_n(\omega)=s_n)\tag{4}$$

と定める. 各 $i=1,\cdots,n$ に対して,

$$z_{ig}=1\text{ 及び }g'\neq g\text{ に対して }z_{ig'}=0$$
$$\Longleftrightarrow\ s_i\text{ は }g\text{ 番目の部分母集団から抽出された}$$

によって定義される k 次元実ベクトル $\boldsymbol{z}_i=(z_{i1},\cdots,z_{ik})$ を導入する. $\boldsymbol{z}_i$ は i 番目の文字列が観測された後で定義される未知の定数ベクトルである.

命題 5. 任意の $d\in D$ に対して, A^* 上の Laplace 様混合モデルに対する EM アルゴリズムは次の形を持つ.

1 各 $g=1,\cdots,k$ に対してパラメーターの任意の初期値 $\hat{\pi}_g^{(0)},\hat{\lambda}_g^{(0)}$ 及び $\hat{\rho}_g^{(0)}$ を選ぶ.

2 $t=1,2,\cdots$ に対して,

2.1 各 $i=1,\cdots,n$ と $g=1,\cdots,k$ に対して

$$\hat{\zeta}_{ig}^{(t)}=\frac{\hat{\pi}_g^{(t-1)}q(s_i|\hat{\lambda}_g^{(t-1)},\hat{\rho}_g^{(t-1)})}{\sum_{g'=1}^{k}\hat{\pi}_{g'}^{(t-1)}q(s_i|\hat{\lambda}_{g'}^{(t-1)},\hat{\rho}_{g'}^{(t-1)})}\tag{5}$$

を計算する.

2.2 各 $g=1,\cdots,k$ に対して

$$\hat{\pi}_g^{(t)}=\frac{1}{n}\sum_{i=1}^{n}\hat{\zeta}_{ig}^{(t)},\tag{6}$$

$$\hat{\lambda}_g^{(t)}=\arg\min_{\lambda_g\in A^*}\sum_{i=1}^{n}\hat{\zeta}_{ig}^{(t)}\left\{-\log\left|\partial U(\lambda_g,d(s_i,\lambda_g))\right|+d(s_i,\lambda_g)\log\left(\frac{\hat{\rho}_g^{(t-1)}}{\hat{\rho}_g^{(t-1)}+1}\right)\right\},\tag{7}$$

$$\hat{\rho}_g^{(t)}=\frac{1}{\sum_{i=1}^{n}\hat{\zeta}_{ig}^{(t)}}\sum_{i=1}^{n}\hat{\zeta}_{ig}^{(t)}d_{H'}(s_i,\hat{\lambda}_g^{(t)})\tag{8}$$

を計算する.

2.3 全ての $g=1,\cdots,k$ に対して $\hat{\pi}_g^{(t)}$, $\hat{\lambda}_g^{(t)}$ 及び $\hat{\rho}_g^{(t)}$ がそれぞれ $\hat{\pi}_g^{(t-1)}$, $\hat{\lambda}_g^{(t-1)}$ 及び $\hat{\rho}_g^{(t-1)}$ に十分に近いならば, 反復を止めて, $\hat{\pi}_g^{(t)},\hat{\lambda}_g^{(t)}$ 及び $\hat{\rho}_g^{(t)}$ を返し, そうでないならば, t を 1 だけ増加させて, ステップ 2.1 に戻る.

$d=d_L$ の場合には, 前節において述べた λ の推定アルゴリズムを用いて, ステップ 2.2 における $\hat{\lambda}_g^{(t)}$ を更新するのが 1 つの自然な方法だろう.

次に, $d=d_{H'}$ の場合に, 命題 5 のアルゴリズムを詳しく検討する. $\ell=\max\{|s_1|,\cdots,|s_n|\}$ とおく. 各 $j=1,\cdots,\ell$ と $h=1,\cdots,z$ に対して

$$f_{gjh}=\frac{1}{n}\sum_{i\in\{i'\in\{1,\cdots,n\}:x_{i'j}=a_h\}}\hat{\zeta}_{ig}\tag{9}$$

とおく. f_{gjh} は, g 番目の部分母集団から抽出された文字列の j 番目の文字が, $\bar{A}$ の中の h 番目の文字と等しい確率の平均の推定量である. g 番目の部分母集団からの文字列の j 番目の位置に最も高い確率で出現すると推定される $\bar{A}$ の中の文字の添え字を h_{gj} によって表す. すなわち,

$$h_{gj}=\arg\max_{1\le h\le z}f_{gjh}.\tag{10}$$

g 番目の混合成分の位置パラメーター λ_g の推定量を

$$\hat{\lambda}_g=a_{h_{g1}}\cdots a_{h_{g\ell}}e\cdots\tag{11}$$

と定める. 各 $g=1,\cdots,k$ と $j\in\mathbb{Z}^+$ に対して h_{gj} 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 $\sum_{g=1}^{k}\sum_{i=1}^{n}\hat{\zeta}_{ig}d_{H'}(s_i,t_g)$ の最小化問題の解であることを示せる. また, コンセンサス配列は拡張 Hamming 距離の和を最小化する. 従って, 式 (11) によって与えられる $\hat{\lambda}_g$ は, 観測値が複数の母集団のうちのどれから抽出されたかが未知である場合のコンセンサス配列の確率的な拡張と見なせる. 式 (9) における $\hat{\zeta}_{ig}$ を $\hat{\zeta}_{ig}^{(t)}$ に置き換えることにより得られる $\hat{\lambda}_g$ を $\hat{\lambda}_g^{(t)}$ と書く.

以下では, $d=d_{H'}$ の時, λ_g のステップ t における推定値としてこの $\hat{\lambda}_g^{(t)}$ を用いる命題 5 のアルゴリズムをアルゴリズム H' と略記し, その漸近的性質を調べる. 我々は n と t に関する漸近理論を展開するから, 各 $i=1,\cdots,n$ と $g=1,\cdots,k$ に対して, アルゴリズム H' からの $\zeta_{ig},\pi_g,\lambda_g$ 及び ρ_g の推定値を, それぞれ $\hat{\zeta}_{ig}^{(n,t)},\hat{\pi}_g^{(n,t)},\hat{\lambda}_g^{(n,t)}$ 及び $\hat{\rho}_g^{(n,t)}$ によって表す. n_g を g 番目の部分母集団から抽出された文字列の数とする. 観測文字列 $s_1,\cdots,s_n$ のうち g 番目の部分母集団からのものを $s_{g1},\cdots,s_{gn_g}$ によって表す. 各 $i=1,\cdots,n_g$ に対して s_{gi} は確率文字列 σ_{gi} の実現値であるとする. アルゴリズム H' は次の漸近的性質を持つ.

命題 6. (a) $\hat{\pi}_g^{(n,t)}\xrightarrow{\text{a.s.}}\pi_g,\hat{\lambda}_g^{(n,t)}=\lambda_g$ a.s. であって, 任意の $g=1,\cdots,k$ に対して, $n_g,t\longrightarrow\infty$ の時, $\hat{\rho}_g^{(n,t)}\xrightarrow{\text{a.s.}}\rho_g$ が成り立つならば, 任意の $i=1,\cdots,n$ と $g=1,\cdots,k$ に対して, $n_g,t\longrightarrow\infty$ の時,

$$\hat{\zeta}_{ig}^{(n,t)}\xrightarrow{\text{a.s.}}\zeta_{ig}.$$

逆に, (b) (i) $\sigma_{g1},\cdots,\sigma_{gn_g}$ に対して命題 2 の条件が満されていて, (ii) $n_g,t\longrightarrow\infty$ の時, $\hat{\zeta}_{ig}^{(n,t)}\xrightarrow{\text{a.s.}}\zeta_{ig}$ が成り立つならば, $n_g,t\longrightarrow\infty$ の時,

$$\hat{\pi}_g^{(n,t)}\xrightarrow{\text{a.s.}}\pi_g,\quad \hat{\lambda}_g^{(n,t)}=\lambda_g\quad\text{a.s.},\quad \hat{\rho}_g^{(n,t)}\xrightarrow{\text{a.s.}}\rho_g.$$

次に, アルゴリズム H' によって, $d=d_{H'}$ を持つ A^* 上の Laplace 様混合モデルのパラメーター $\boldsymbol{\theta}$ を強一致推定するための条件を考える. 次の 2 つの条件を導入する. C_1: 各 $i=1,\cdots,n$ に対して $\sum_{g=1}^{k}z'_{ig}=1$ かつ $(z'_{11},\cdots,z'_{1k},\cdots,z'_{n1},\cdots,z'_{nk})\neq(z_{11},\cdots,z_{1k},\cdots,z_{n1},\cdots,z_{nk})$ となる任意の $(z'_{11},\cdots,z'_{1k},\cdots,z'_{n1},\cdots,z'_{nk})\in(0,1)^{kn}$ に対して, $n^*\longrightarrow\infty$ の時,

$$\max_{(\lambda_1,\cdots,\lambda_k,\rho_1,\cdots,\rho_k)}\frac{1}{n}\sum_{g=1}^{k}\sum_{i=1}^{n}z'_{ig}\log q(s_i;\lambda_g,\rho_g)$$
$$\leq\max_{(\lambda_1,\cdots,\lambda_k,\rho_1,\cdots,\rho_k)}\frac{1}{n}\sum_{g=1}^{k}\sum_{i=1}^{n}z_{ig}\log q(s_i;\lambda_g,\rho_g)\quad\text{a.s.}$$

が成り立つ. ここで, $n^*=\min\{n_1,\cdots,n_k\}$. 例えば, 十分に多くの観測文字列が与えられていて, g 番目の部分母集団から抽出された文字列のうちの一部のみに基づいて, または/かつ, 他の部分母集団から抽出された文字列も含めて, ℓ_g^* より大きい対数尤度を持つ g 番目の部分母集団の分布が存在しないならば, 条件 C_1 は満たされている. ここで, 各 $g=1,\cdots,k$ に対して, ℓ_g^* は, g 番目の部分母集団から抽出された全ての文字列に基づいて最大の対数尤度を持つ g 番目の部分母集団の分布の対数尤度を表す. 命題 10 の下では, C_1 は病的な状況を排除するための自然な仮定である.

C_2: 与えられた $s_1,\cdots,s_n\in A^*$ に対して, $(\zeta_{11},\cdots,\zeta_{nk},\lambda_1,\cdots,\lambda_k,\rho_1,\cdots,\rho_k)\in(0,1)^{nk}\times(A^*)^k\times(0,1)^k$ に関する

$$\frac{1}{n}\sum_{g=1}^{k}\sum_{i=1}^{n}\zeta_{ig}\log q(s_i;\lambda_g,\rho_g)\tag{12}$$

の最大化問題の解 $(\zeta_{11}^*,\cdots,\zeta_{nk}^*,\lambda_1^*,\cdots,\lambda_k^*,\rho_1^*,\cdots,\rho_k^*)$ は一意であるとし, 各 $n,t'\in\mathbb{Z}^+$ に対して

$$\hat{\boldsymbol{\theta}}^{(n,0,t')}=\arg\max_{\hat{\boldsymbol{\theta}}^{(n,0)}\in\Theta}\frac{1}{n}\sum_{g=1}^{k}\sum_{i=1}^{n}\hat{\zeta}_{ig}^{(n,t')}\times\log q(s_i;\hat{\lambda}_g^{(n,t')},\hat{\rho}_g^{(n,t')})\tag{13}$$

と定める. ここで, $\hat{\zeta}_{ig}^{(n,t')},\hat{\lambda}_g^{(n,t')}$ 及び $\hat{\rho}_g^{(n,t')}$ は, それぞれ, 初期値 $\hat{\boldsymbol{\theta}}^{(n,0)}$ を持つアルゴリズム H' が反復ステップ t' において返す ζ_{ig},λ_g 及び ρ_g の推定値である. $\hat{\boldsymbol{\theta}}^{(n,0,t')}$ は, 全ての初期値のうち, アルゴリズム H' が反復ステップ t' において関数 (12) を最大化する推定値を返す初期値である.

命題 7. 条件 C_1 と C_2 が成り立ち, $\sigma_{g1},\cdots,\sigma_{gn_g}$ に対して命題 2 の条件が満されているとする. 更に, $n,t\longrightarrow\infty$ の時, アルゴリズム H' からの推定値 $\tilde{\boldsymbol{\theta}}^{(n,t)}$ が $\boldsymbol{\theta}$ を強一致推定する初期値 $\tilde{\boldsymbol{\theta}}^{(n,0)}=(\tilde{\pi}_1^{(n,0)},\cdots,\tilde{\pi}_k^{(n,0)},\tilde{\lambda}_1^{(n,0)},\cdots,\tilde{\lambda}_k^{(n,0)},\tilde{\rho}_1^{(n,0)},\cdots,\tilde{\rho}_k^{(n,0)})\in\Theta$ が存在するならば, $n^*,t,t'\longrightarrow\infty$ の時, $\boldsymbol{\theta}$ は, 式 (13) によって与えられる初期値 $\hat{\boldsymbol{\theta}}^{(n,0,t')}$ を持つアルゴリズム H' からの推定量 $\hat{\boldsymbol{\theta}}^{(n,t,t')}$ によって強一致推定される.

命題 7 を使うと, A^* 上の Laplace 様混合モデルに対する EM アルゴリズムに関する次の命題を示すことができる.

命題 8. 命題 7 の条件の下で, 式 (12) を満たす初期値 $\hat{\boldsymbol{\theta}}^{(n,0,t')}$ を持つアルゴリズム H' は, $n,t,t'\longrightarrow\infty$ の時, A^* 上の Laplace 様混合分布に対する EM アルゴリズムに概収束する.

5. Laplace 様混合モデルに基づく文字列クラスタリング

最後に, これまでの節において得られた結果に基づいて, 文字列のクラスタリング方式を導出する. n 個の文字列 $s_1,\cdots,s_n\in A^*$ を k 個のクラスに分類する問題を考える. $s_1,\cdots,s_n$ を生成する機構のモデルとして, 混合係数 $\pi_1,\cdots,\pi_k$ を持つ $\mathrm{L}_{A^*}(\lambda_1,\rho_1),\cdots,\mathrm{L}_{A^*}(\lambda_k,\rho_k)$ の混合モデル

$$\sum_{g=1}^{k}\frac{\pi_g}{(\rho_g+1)\left|\partial U(\lambda_g,d(s,\lambda_g))\right|}\left(\frac{\rho_g}{\rho_g+1}\right)^{d(s,\lambda_g)}$$

を想定する. Bayes の定理から, 各 $i = 1, \cdots, n$ と $g = 1, \cdots, k$ に対して, $s_1, \cdots, s_n$ が与えられている時に, s_i が第 g クラスに属する事後確率は,

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{Z}_i = \boldsymbol{z}_g | s_1, \cdots, s_n) = \frac{\pi_g q(s_i; \lambda_g, \rho_g)}{\sum_{g'=1}^{k} \pi_{g'} q(s_i; \lambda_{g'}, \rho_{g'})}$$

によって与えられる.

命題 9. 定理 6 の (b) 部の条件が満されているとし, $\hat{\boldsymbol{\theta}}_{t'}^{(n,t)}$ は, 式 (13) を満たす初期値 $\hat{\boldsymbol{\theta}}_{t'}^{(n,0)} \in \Theta$ を持つアルゴリズム H' からの推定値であるとする. この時, 各 $i = 1, \cdots, n$ に対して

$g^* = \arg\max_{1 \le g \le k} \pi_{\hat{\boldsymbol{\theta}}_{t'}^{(n,t)}}(\boldsymbol{Z}_i = \boldsymbol{z}_g | s_1, \cdots, s_n)$ ならば,

s_i を第 g^*クラスに分類せよ

というクラスタリング方式は, $n, t \longrightarrow \infty$ の時, 正しく分類する事後確率が最大である意味で漸近的に最適である.

6. まとめ

本稿において, 我々は, 特に生物配列への応用を念頭において, 文字列の集合上で混合モデルとそれに対する EM アルゴリズムの理論を展開することにより, 混合モデルに基づいた, 文字列データの教師なしクラスタリング方法の開発に取り組んだ. 今後は, 提案した方法を生物配列の解析に応用することにより, 実際のデータ解析のおけるその有用性を検討したい.

補足: パラメーター空間が $A^* \times (0, \infty)$ の場合の最尤推定量の強一致性

本節では, これまでの節における結果のうちのいくつかを得るために使われた, パラメーター空間が $A^* \times (0, \infty)$ である場合の最尤推定量の強一致性に関する結果を述べる. パラメーター空間が $\mathbb{R}$ である場合の最尤推定量の強一致性に関する一般的な結果については, [5], [6] を参照. $\sigma \in \mathcal{M}(\Omega, A^*)$ とし, $q(s; \boldsymbol{\theta})$ を σ の分布の確率関数とする. ここで, $\boldsymbol{\theta} = (\theta_1, \theta_2) \in A^* \times (0, \infty)$. 任意の $\boldsymbol{\theta}' \in A^* \times (0, \infty)$ に対して

$$\eta(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_{s \in A^*} \log(q(s; \boldsymbol{\theta})) q(s; \boldsymbol{\theta}')$$

とおく. 任意の $\boldsymbol{\theta} \in A^* \times (0, \infty)$ に対して $\eta(\boldsymbol{\theta}', \boldsymbol{\theta}) \le \eta(\boldsymbol{\theta}', \boldsymbol{\theta}')$ が成り立つ. 次の 3 つの正則条件を導入する. これらの条件は, パラメーター空間が $\mathbb{R}$ である場合の最尤推定量の強一致性のための正則条件を少し修正したものである.

1. $|\theta_2 - \theta_2'| > 0$ ならば, $\eta(\boldsymbol{\theta}', \boldsymbol{\theta}') - \eta(\boldsymbol{\theta}', \boldsymbol{\theta}) > 0$.

2. $M > 0$ に対して

$$g_M(s) = \sup_{\substack{\theta_1 \ne \theta_1^* \text{or} \\ |\theta_2 - \theta_2^*| > M}} \log q(s; \boldsymbol{\theta})$$

とおくと, 十分に大きな M に対して

$$c_g = \mathrm{E}_{\boldsymbol{\theta}^*}[g_M(s)] < \eta(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*).$$

3. 任意の $s \in A^*$ に対して $q(s; \boldsymbol{\theta})$ は θ_2 に関して偏微分可能であって, 正則条件 2 を満たす M に対して

$$h_M(s) = \sup_{|\theta_2 - \theta_2^*| \ge M} \left| \frac{\partial}{\partial \theta_2} \log q(s; \boldsymbol{\theta}) \right|$$

とおくと,

$$c_h = \mathrm{E}_{\boldsymbol{\theta}^*}[h_M(s)] < \infty.$$

命題 10. $\sigma_1, \cdots, \sigma_n \in \mathcal{M}(\Omega, A^*)$ は (i) 独立で, (ii) 同一の確率関数 $q(s; \boldsymbol{\theta})$ を持ち, (iii) $q(s; \boldsymbol{\theta})$ は台 A^* を持つとする. $\boldsymbol{\theta} = (\theta_1, \theta_2) \in A^* \times (0, \infty)$ とし, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$ をパラメーターの真値とする. 各 $i = 1, \cdots, n$ に対して σ_i の実現値を s_i によって表し, $s_1, \cdots, s_n$ に基づく $\boldsymbol{\theta}$ の最尤推定量を $\check{\boldsymbol{\theta}} = (\check{\theta}_1, \check{\theta}_2)$ によって表す. 正則条件 1 から 3 が満たされているならば, $N_0 \in \mathbb{Z}^+$ が存在して, $n \ge N_0$ ならば, $\check{\theta}_1 = \theta_1^*$ a.s., かつ $n \longrightarrow \infty$ の時, $\check{\theta}_2 \xrightarrow{\text{a.s.}} \theta_2^*$.

正則条件 1 から 3 は非常に一般的な条件であり, A^* 上の Laplace 様分布は, 命題 10 の条件 (iii) はもちろんのこと, これらの条件を満している.

参考文献

[1] Koyano, H., Hayashida, M. and Akutsu, T.: Maximum margin classifier working in a set of strings. arXiv:1406.0597v2.

[2] Koyano, H. and Kishino, H.: Quantifying biodiversity and asymptotics for a sequence of random strings, *Physical Review E*, Vol. 81, No. 6, p. 061912 (2010).

[3] Koyano, H., Tsubouchi, T., Kishino, H. and Akutsu, T.: Archaeal β diversity patterns under the seafloor along geochemical gradients, *Journal of Geophysical Research G (Biogeosciences)*, Vol. 119, No. 9, pp. 1770–1788 (2014).

[4] Olivares-Rodríguez, C. and Oncina, J.: A stochastic approach to median string computation, *Structural, Syntactic, and Statistical Pattern Recognition* (da Vitoria Lobo, N., Kasparis, T., Roli, F. Kwok, J. T., Georgiopoulos, M., Anagnostopoulos, G. C. and Loog, M., eds.), Springer, Berlin, pp. 431–440 (2008).

[5] Perlman, M. D.: On the strong consistency of approximate maximum likelihood estimators, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Le Cam, L. M., Neyman, J. and Scott, E. L., eds.), Vol. 1, Berkeley, CA, University of California Press, pp. 263–281 (1972).

[6] Wald, A.: Note on the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics*, Vol. 29, pp. 595–601 (1949).