# 複数の構成要素からなる文書データの分類を目的とした階層潜在クラスモデル

1X07C006-1　荒川貴紀
指導教員　後藤正幸

## 1　はじめに

文書データの効率的な管理技術として，文書の自動分類が重要性を増しており，様々な確率的文書モデルが提案されている．特に，文書の潜在クラスを仮定したモデルが近年注目を集めており，中でも多様な文書に対して汎用性が高く，優れた性能を示すモデルとして混合 Polya 分布に基づく文書モデル [1] がある．

一般的に文書の多くは，タイトルや本文といった複数の異なる種類のテキストデータの複合体として構成される．このような文書を分類する場合，本文だけでなくタイトルなどの付加的な構成要素を活用することで分類精度が向上する [2]．

本研究では，高精度な文書分類器を構築するため，潜在クラスモデルである混合 Polya 分布を用いて複数の構成要素からなる文書をモデル化することを考える．このような文書をモデル化する際には，構成要素間の独立性を仮定する方法 [3] がしばしば用いられるが，実際には構成要素同士は強い関係性を持つと考えられる．このような視座に立ち，本研究では構成要素同士の関係性を考慮するために，潜在クラスを階層的に導入した階層潜在クラスモデルを提案する．このモデルは，各構成要素の潜在クラスが，上位階層の一つの潜在クラスから生成されるという階層構造をなす．さらに，このモデルの学習方法として，EM アルゴリズムの内部で再び EM アルゴリズムを用いる二重 EM アルゴリズムを提案する．新聞記事データを用いた評価実験により，提案手法の有効性を示す．

## 2　従来研究

### 2.1　混合 Polya 分布

形態素解析により文書を単語単位に分割し，品詞による絞り込み等の適切な特徴語選択により $V$ 種類の単語が抽出されているものとする．このとき，文書 $\boldsymbol{x}$ の特徴ベクトルは単語の出現頻度 $x_v(v=1,2,\cdots,V)$ を用いて $\boldsymbol{x}=(x_1,x_2,\cdots,x_V)$ のように表される．その生成モデルとして，混合数 $M$，混合比 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_M)$ の混合 Polya 分布 $P_{PM}(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\alpha})$ は次式で定義される．

$$P_{PM}(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\alpha})=\sum_{m=1}^{M}\lambda_m P_{Polya}(\boldsymbol{x};\boldsymbol{\alpha}_m)$$
$$=\sum_{m=1}^{M}\lambda_m\frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m+x)}\prod_{v=1}^{V}\frac{\Gamma(x_v+\alpha_{mv})}{\Gamma(\alpha_{mv})},\quad(1)$$

ただし，$\sum_{m=1}^{M}\lambda_m=1$, $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_M)$，$\boldsymbol{\alpha}_m=(\alpha_{m1},\alpha_{m2},\cdots,\alpha_{mV})$, $\alpha_m=\sum_{v=1}^{V}\alpha_{mv}$, $x=\sum_{v=1}^{V}x_v$ である．$P_{Polya}(\boldsymbol{x};\boldsymbol{\alpha}_m)$ は第 $m$ 番目の Polya 分布を表す．

### 2.2　構成要素間の独立性を仮定したモデル

本研究では，入力文書が複数の構成要素からなる場合の文書分類問題を対象とする．入力文書 $\boldsymbol{x}$ が $J$ 個の構成要素からなる場合，入力文書 $\boldsymbol{x}$ は構成要素 $j$ の特徴ベクトル $\boldsymbol{x}^j$ を用いて，$\boldsymbol{x}=(\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^J)$ のように表される (以下では，構成要素の種類 $j$ を上付き添え字で表わし，それ以外の変数を下付き添え字で表わすものとする)．このとき，確率モデル $P(\boldsymbol{x})$ は，複数の構成要素データの同時確率 $P(\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^J)$ となる．ここで，各構成要素が独立に生成されると仮定することにより，同時確率は構成要素ごとの生成モデル $P(\boldsymbol{x}^j)$ の積に分解され，

$$P(\boldsymbol{x})=\prod_{j=1}^{J}P(\boldsymbol{x}^j),\quad(2)$$

のように表される [3]．

## 3　提案手法

### 3.1　階層潜在クラスモデル

本研究では，混合 Polya 分布を用いて複数の構成要素からなる文書をモデル化することを考える．構成要素 $j$ の単語の種類数を $V^j$ とする．各構成要素の独立性を仮定したもとで，構成要素 $j$ の生成モデルに混合数 $M^j$，混合比 $\boldsymbol{\lambda}^j=(\lambda_1^j,\lambda_2^j,\cdots,\lambda_{M^j}^j)$ の混合 Polya 分布を用いると，文書全体の確率モデルは

$$P(\boldsymbol{x};\boldsymbol{\lambda}^j\}_{j=1}^{J},\{\boldsymbol{\alpha}^j\}_{j=1}^{J})=\prod_{j=1}^{J}\left\{\sum_{m=1}^{M^j}\lambda_m^j P_{Polya}(\boldsymbol{x}^j;\boldsymbol{\alpha}_m^j)\right\},\quad(3)$$

のように表される．ただし，$\sum_{m=1}^{M^j}\lambda_m^j=1$，$\boldsymbol{\alpha}_m^j=(\alpha_{m1}^j,\alpha_{m2}^j,\cdots,\alpha_{mV^j}^j)$ である．実際の文書は，構成要素間に強い関係性が存在すると考えられる．構成要素間の関係性を考慮するために，個別の構成要素の潜在クラスに加え，図 1 に示すように全ての構成要素に共通する文書全体の潜在クラスを導入する．そして，構成要素ごとの個別の潜在クラスの選択確率は，共通する一つの潜在クラスによって決定されるという階層構造を仮定する．以下では，このモデルを階層潜在クラスモデルと呼ぶことにする．

階層潜在クラスモデルは，式 (3) のモデルを要素分布とした混合分布であり，その混合数を $L$，混合比を $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_L)$ とすると，次式のように表される．

$$P(\boldsymbol{x}\,;\boldsymbol{\mu},\{\boldsymbol{\lambda}^j\}_{j=1}^{J},\{\boldsymbol{\alpha}^j\}_{j=1}^{J})$$
$$=\sum_{l=1}^{L}\mu_l\prod_{j=1}^{J}\left\{\sum_{m=1}^{M^j}\lambda_{lm}^j P_{Polya}(\boldsymbol{x}^j;\boldsymbol{\alpha}_m^j)\right\},\quad(4)$$

ただし，$\sum_{l=1}^{L}\mu_l=1$，$\sum_{m=1}^{M^j}\lambda_{lm}^j=1$ である．

階層潜在クラスモデルにより，ある文書 $\boldsymbol{x}=(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^j,\cdots,\boldsymbol{x}^J)$ が生成される過程を図 1 に示す．階層潜在クラスモデルのもとでは，まず全構成要素に共通の潜在クラス $w\in\{1,\cdots,L\}$ が一つ選択され，$w$ に依存して各構成要素の潜在クラス $z^j\in\{1,\cdots,M^j\}(j=1,\cdots,J)$ が $J$ 個選択される．さらに，各 $z^j$ に依存して，各 $\boldsymbol{x}^j$ が生成される．ただし，各々の潜在クラスが選択される確率は $P(w=l)=\mu_l$，$P(z^j=m|w=l)=\lambda_{lm}^j$ である．

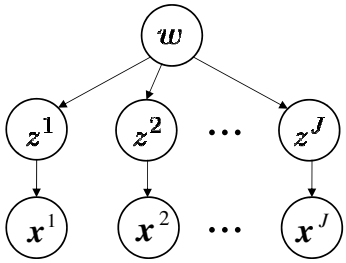


**図 1. 階層潜在クラスモデルのグラフィカルモデル**

### 3.2 二重 EM アルゴリズムによるパラメータ推定

多くの混合分布モデルは，EM アルゴリズムによって混合比と要素分布パラメータを交互に更新することでパラメータを推定することができる．ところが，式 (4) に示した階層潜在クラスモデルは，要素分布がさらに混合分布の積であるため，単純に EM アルゴリズムを適用した場合，要素分布パラメータの更新式を閉じた形で得ることができない．

そこで，階層潜在クラスモデルのパラメータを推定する際には，EM アルゴリズムの内部で再び EM アルゴリズムを用いる．この内側の EM アルゴリズムにより，要素分布パラメータを更新することができる．$N$ を学習文書数，$x_{iv}^j (v = 1, 2, \cdots, V^j)$ を $i$ 番目の文書の構成要素 $j$ に出現する単語の出現頻度とすると，各パラメータの更新式は以下のようになる．

$$\mu_l = \frac{1}{N} \sum_{i=1}^{N} P_{il}, \tag{5}$$

$$\lambda_{lm}^j = \frac{1}{N\bar{\mu}_l} \sum_{i=1}^{N} P_{ilm}^j, \tag{6}$$

$$\alpha_{mv}^j = \bar{\alpha}_{mv}^j \frac{\sum_{i=1}^{N} (\sum_{l=1}^{L} P_{il} P_{ilm}^j) \{x_{iv}^j / (x_{iv}^j - 1 + \bar{\alpha}_{mv}^j)\}}{\sum_{i=1}^{N} (\sum_{l=1}^{L} P_{il} P_{ilm}^j) \{x_i^j / (x_i^j - 1 + \bar{\alpha}_m^j)\}}, \tag{7}$$

ただし，$x_i^j = \sum_{v=1}^{V^j} x_{iv}^j$ であり，$\bar{\mu}_l$，$\bar{\alpha}_{mv}^j$ などの上付きバーは更新前の値を表す．また，$P_{il}$ は $i$ 番目の文書が属する潜在クラス $w_i$ の事後確率，$P_{ilm}^j$ は $i$ 番目の文書が属する潜在クラス $w_i$ が $l$ であるという条件のもとでの構成要素 $j$ が属する潜在クラス $z_i^j$ の事後確率であり，それぞれ以下のように定義する．

$$P_{il} = \frac{\bar{\mu}_l \prod_{j=1}^{J} \{\sum_{m=1}^{M^j} \bar{\lambda}_{lm}^j P_{Polya}(\boldsymbol{x}_i^j; \bar{\boldsymbol{\alpha}}_m^j)\}}{\sum_{l=1}^{L} \bar{\mu}_l \prod_{j=1}^{J} \{\sum_{m=1}^{M^j} \bar{\lambda}_{lm}^j P_{Polya}(\boldsymbol{x}_i^j; \bar{\boldsymbol{\alpha}}_m^j)\}}, \tag{8}$$

$$P_{ilm}^j = \frac{\bar{\lambda}_{lm}^j P_{Polya}(\boldsymbol{x}_i^j; \bar{\boldsymbol{\alpha}}_m^j)}{\sum_{m=1}^{M^j} \bar{\lambda}_{lm}^j P_{Polya}(\boldsymbol{x}_i^j; \bar{\boldsymbol{\alpha}}_m^j)}. \tag{9}$$

内側の EM アルゴリズムでは，式 (6) と式 (7) によって $\boldsymbol{\lambda}^j$ と $\boldsymbol{\alpha}^j$ を交互に反復更新する．さらに，この内側の EM アルゴリズムと，式 (5) による $\boldsymbol{\mu}$ の更新を交互に反復することで，全パラメータの推定値を得ることができる．以下に，アルゴリズムの流れをまとめる．

**[二重 EM アルゴリズム]**

**Step0)** 各パラメータの初期値を任意に定める．

**Step1)** 式 (5) により $\boldsymbol{\mu}$ を更新する．

**Step2)** 式 (6) により $\boldsymbol{\lambda}^j$ を更新する．

**Step3)** 式 (7) により $\boldsymbol{\alpha}^j$ を更新する．

**Step4)** $\boldsymbol{\lambda}^j$，$\boldsymbol{\alpha}^j$ の値が収束するまで Step2 から Step3 を繰り返す．

**Step5)** $\boldsymbol{\mu}$ の値が収束するまで Step1 から Step4 を繰り返す．収束したらアルゴリズムを終了する． □

## 4 評価実験

### 4.1 実験データ

提案手法の有効性を検証するために，毎日新聞 2005 年の記事データを用いて評価実験を行った．カテゴリは社説，国際，経済，家庭，芸能，スポーツ，社会の 7 つを使用した．各カテゴリ 1000 件ずつ，合計 7000 件の記事をランダムに抽出し，カテゴリごとにランダムに選択した 800 件を学習データ，200 件をテストデータとする．これを 3 回繰り返し，3 回の分類精度の平均で評価する．また，文書の構成要素として，本文とタイトルを用いた ($J = 2$)．評価尺度には，テストデータに対する分類精度を用いた．

### 4.2 実験結果及び考察

実験データに対し，以下の 3 つの手法を適用し，分類精度を比較した．

- 比較手法 1：構成要素ごとに独立な多項分布を使用したモデル．
- 比較手法 2：構成要素ごとに独立な混合 Polya 分布を使用した式 (3) のモデル．$M^j = 10 (j = 1, 2)$ とした．
- 提案手法：階層潜在クラスモデル．式 (4)．$L = 10$, $M^j = 10 (j = 1, 2)$ とした．

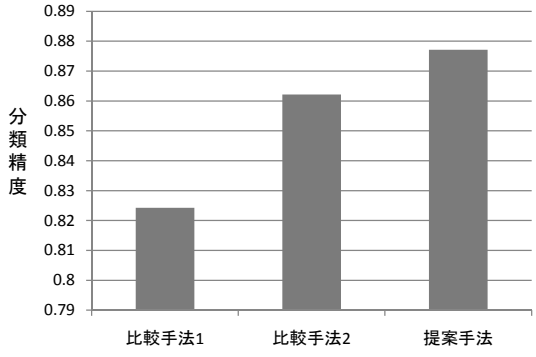


**図 2. 各手法の分類精度**

比較手法 1 と比較手法 2 を比べると，比較手法 2 の分類精度が優れることがわかる．これは，各々の構成要素に多項分布を仮定するのではなく，潜在クラスモデルである混合 Polya 分布を仮定することで，文書の潜在的な話題や内容といったものを適切にモデル化できているためであると考えられる．

また，提案手法は，両比較手法に比べて，分類精度が高いことがわかる．これは，全構成要素に共通の潜在クラスを導入することによって，構成要素同士の関係性を適切にモデル化できたためであると考えられる．

以上より，提案手法の有効性を示すことができた．

## 5 まとめと今後の課題

本研究では，複数の構成要素からなる文書の分類器として潜在クラスを階層的に仮定した階層潜在クラスモデルを提案し，その有効性を確認した．本研究では，新聞記事を用いた実験を行ったが，Web ページや論文など他の文書に対する有効性を検証することを今後の課題とする．

## 参考文献

[1] 貞光九月, 三品拓也, 山本幹雄, “混合ディリクレ分布を用いたトピックに基づく言語モデル” 電子情報通信学会論文誌, Vol.j88-D-Ⅱ, No.9, pp. 1771–1779, 2005.

[2] 藤野昭典, 上田修功, 斎藤和巳, “最大エントロピー原理に基づく付加情報の効果的な利用によるテキスト分類,” 情報処理学会論文誌, Vol.47, No.10, pp. 2929–2937, 2006.

[3] E. Brochu and N. Freitas, “Name That Song!: A Probabilistic Approach to Querying on Music and Text,” *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA, pp. 1505–1512, 2003.