社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－172（7）
2006／3／27

無限混合ディリクレ文書モデル

持橋大地　菊井玄一郎
ATR 音声言語コミュニケーション研究所 音声言語処理研究室
daichi.mochihashi@atr.jp, genichiro.kikui@atr.jp

概要

文書があるトピックの持つ確率分布から生成されたと仮定し, その確率分布パラメータと文書のトピックへの帰属確率を求めるモデルに, ナイーブベイズ法を Polya 分布を用いてベイズ的に精密にとらえ直した混合ディリクレモデル (DM) があるが, この方法はトピック数を事前に与える必要があるという欠点があった.
これに対し, 本論文では可算無限個の混合比にディリクレ過程事前分布を与えることにより, データの複雑さに合わせて混合数を自動推定するディリクレ過程混合モデルによる方法を検討する. モデル選択により混合数を決定する方法と異なり, この方法は混合数の事後分布をパラメータと同時に推定し, 期待値を取ることで予測を行うことができる. 実験の結果, 必要な混合数の上限を推測することができ, 特に小規模データに対しては性能がさらに上昇することがわかった.
キーワード: DM, ディリクレ過程, 混合モデル, K 平均法, 変分ベイズ法

Infinite Dirichlet Mixtures in Text Modeling

Daichi Mochihashi　Genichiro Kikui
ATR Spoken Language Communication Research Laboratories
daichi.mochihashi@atr.jp, genichiro.kikui@atr.jp

Abstract

This paper proposes a Dirichlet process mixture modeling approach to Dirichlet Mixtures (DM). Endowing a prior distribution on an infinite number of mixture components, this approach yields an appropriate number of components as well as their parameters at the same time. Experimental results on amino acid distributions and text corpora confirmed this effect and show comparative performance on large datasets and better performance on small datasets avoiding overfitting.
Keywords: Dirichlet Mixtures, Dirichlet processes, Mixture models, K-Means algorithm, variational Bayes

1　はじめに

文書をモデル化する方法として, ある文書全体がトピックと呼ばれる一つの確率分布から生まれたと仮定し, そのトピック確率分布群のパラメータ, および文書の各トピックへの帰属確率を計算する方法がある. この方法はトピックが既知の場合にはナイーブベイズ, 未知の場合には Unigram Mixtures [1] と呼ばれているが, UM の拡張として, 混合 Polya 分布を用いた Dirichlet Mixtures [2] はこの問題をベイズ的にとらえ, 文書モデルや文脈モデルにおいて高い性能を持つことが知られている [3][4].

この方法は単語単体上に, 一つ一つがトピックに対応する混合ディリクレ分布をユニグラムの事前分布として仮定し, 合成することで得られる混合 Polya 分布から, 観測データをもとに事前分布を推測する方法であるが, その際のトピック数は事前に与えなければならないという欠点があった. トピック数が少ないと単体上の真の分布をうまくモデル化できず, 多すぎると過適応を起こすため, データの複雑さに応じたトピック数の推定は重要な問題であると考えられる.

この問題は, 音声認識等で使われる混合正規分布の混合数推定とも同様の問題である. 従来このために, 混合数を既知としたモデルを計算し, その尤度を比較することでモデル選択を行う方法 (たとえば [5]) や, モデルの分割/マージを繰り返すことで最適混合数を見出だす方法 [6][7] などが行われてきたが, 高次元かつ大規模な文書データの場合には, 予想される混合数は非常に大きく, このような方法でモデル選択を行うことは計算量的に現実的でない. また, モデル選択の際に使われる尤度についても, 最大の尤度とそれ以外の間に大きな違いがないことが指摘されており [8], 選択を行って一意に決めるのではなく, 最適混合数の事後分布から期待値を取ることで行う予測方法が求められているといえる.

このような混合モデルの構造推定法として, 近年, ディリクレ過程 (Dirichlet process, DP) に基づくディリクレ過程混合モデルが注目されている [9]. この方法はすでに, 文書の中に含まれる単語ごとにトピック

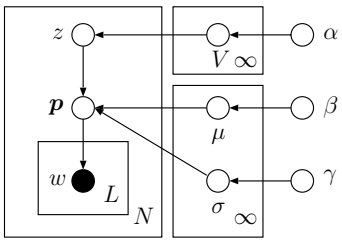


図 1: 無限 DM のグラフィカルモデル.

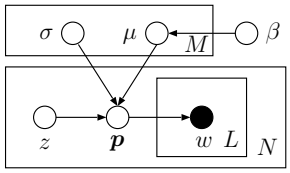


図 2: Smoothed DM のグラフィカルモデル.

事後分布を求める LDA [10] について適用され, 最適なトピック数を同時に推定できることが示されている [11] が, 本論文ではこの方法を, 文書レベルの混合モデルである DM に対して適用し, その結果を報告する. 図 1 が提案法のグラフィカルモデル, 図 2 が従来法のグラフィカルモデルである.

無限 (ディリクレ過程) 混合モデルはすでに, MCMC 法を用いて混合正規分布の場合に良い推定を与えることが報告されており [12], 指数分布族の場合にも一般化されているが, 指数分布族でない Polya 分布の場合には MCMC 法の適用が現実的でないため, 近年提案されたディリクレ過程に対する変分近似法 [13] によって事後分布を推定する.

2 章では, ディリクレ過程とディリクレ過程混合モデルについて簡単に説明し, その Stick-breaking 表現と有限精度での切り捨てについて述べる. 3 章で Polya 分布と DM について説明し, 変分ベイズ法を用いた無限 DM の変分ベイズ EM アルゴリズムを示す. 4 章で, DM の動機となったアミノ酸配列データ, および Cranfield コーパスと Reuters-21578 コーパスを用いた実験結果を示す. 5 章で結論と, 今後の展望について述べる.

2 ディリクレ過程混合モデル

2.1 Dirichlet process とは

Dirichlet process [14] とは, 「凝集」を表現する確率過程であり, 混合分布における各コンポーネント分布の無限個の生成モデルになっている. すなわち, 各コンポーネント分布の事前分布[1]がディリクレ過程 $G \sim \mathrm{DP}(\alpha, G_0)$ に従うとき, あるデータ x_{n+1} がコンポーネント分布 ψ_k に属する事前確率はこれまでに生成されたデータ $x_1, x_2, \ldots, x_n$ に依存し,

$$p(x_{n+1} \sim \psi_k) = \begin{cases} \dfrac{f_k}{n+\alpha} & (\psi_k \text{ がすでに出現}) \quad (1) \\ \dfrac{\alpha}{n+\alpha} & (\psi_k = \psi\,; \psi \sim G_0) \quad (2) \end{cases}$$

[1]したがって, ディリクレ過程とは確率分布の確率分布である.

である. ここで, f_k は $x_1 \ldots x_n$ の中で $x \sim \psi_k$ であった回数. すなわち, 新しいデータ x_{n+1} はこれまでに出現したコンポーネント ψ_k にその頻度 f_k に比例した確率で属するが, 小さい値 $\alpha/(n+\alpha)$ に比例して新しいコンポーネント ψ に属する可能性も持つ. このとき, ψ はそれ自体の事前分布 G_0 から新しく生成される.

$n = 0$ のとき, f_k は必ず 0 であるから, 式 (1) となることはなく, まず式 (2) から最初のコンポーネント $\psi_1 \sim G_0$ が生成され, ψ_1 から x_1 が生成される. x_2 は ψ_1 と $\psi_2 \sim G_0$ から生成される可能性をそれぞれ $1/(1+\alpha) : \alpha/(1+\alpha)$ の確率で持ち, 以下式 (2) が選ばれるごとにコンポーネント分布が増えていき, その増大速度はデータ数 n に対して $O(\log n)$ であることが知られている [15].

実際には, 上の過程はデータ $\mathbf{x} = x_1, x_2, \ldots, x_{n+1}$ の並び替えに依存しない. したがって, データ $x \in \mathbf{x}$ がコンポーネント ψ_k に属する事後確率は, x が最後になるようにデータを並び替えて,

$$p(\psi_k|x) \propto p(x|\psi_k)\, p(x \sim \psi_k) \quad (3)$$

$$= \begin{cases} \psi_k(x)\, \dfrac{f_k}{n+\alpha} & (\psi_k \text{ がすでに出現}) \\ \psi(x)\, \dfrac{\alpha}{n+\alpha} & (\psi \sim G_0) \end{cases} \quad (4)$$

として求まる. ここで, f_k は $\mathbf{x}$ から x を除いた中で ψ_k からデータが生成された回数, $\psi_k(x)$ はコンポーネント分布 ψ_k における x の確率密度である.

式 (4) はディリクレ過程混合分布に従うデータの事後分布を表す式になっているが, 実際に必要な情報は式 (3) からわかるように, コンポーネント番号を表す変数 θ_k $(k = 1, 2, \cdots, \infty)$ と対応するコンポーネント分布 ψ_i の二つに分けて考えることができ, 前者のみが自然数上の 1 次元のディリクレ過程に従うと考えてもよい.

2.2 Stick-breaking 表現

このとき, θ_k $(k = 1, 2, \cdots, \infty)$ の事前分布は, Stick-breaking 表現 Stick(α) とよばれる以下の式に従うことが知られている [16]$_{k-1}$

$$\theta_k = V_k \prod_{i=1} (1 - V_i) \quad (5)$$

$$V_i \sim \mathrm{Be}(1, \alpha)\,. \quad (6)$$

これは幾何分布をソフト化した分布であり[2], θ_k が選ばれる確率は, (θ_1 で止まらない確率)×(θ_2 で止まらない確率) ×$\cdots$× (θ_k で止まる確率) となって, 指数的に減少する事前分布を持つ. [3]

従って, 充分大きな k 以降の値は急速に小さくなるため, 閾値 K に対して $V_K = 1$ とすれば $\theta_k = 0$ $(k >$

[2]幾何分布の場合は, V_i は定数になる.

[3]この減衰の度合いは (6) 式より, DP のハイパーパラメータ α によって支配される. 後に述べるように, 実際には α について一様で, 減衰のほとんどない事前分布から始め, α の事後分布もデータから推定して用いる.

K) となり, 無限個の θ_k に対する近似を構成することができる.

このように切り捨てを行った分布を変分事後分布として用いることにより, ディリクレ過程混合モデルに対して変分ベイズ法を適用することができ, 一般に次の変分ベイズ EM アルゴリズムをもつ.

初期化:

- For $k = 1 \ldots K$,
  $\pi_{k1} = 1,\ \pi_{k2} = \alpha$.
- $\alpha = s_1 \,/\, s_2$.

E step:

$$\phi_{nk} \propto p(x_n|\psi_k) \cdot \exp\Big\{\Psi(\pi_{k1}) - \Psi(\pi_{k1}+\pi_{k2}) - \textstyle\sum_{i=k+1}^{K}\left[\Psi(\pi_{i2}) - \Psi(\pi_{i1}+\pi_{i2})\right]\Big\}. \quad (7)$$

M step:

$$(1)\quad \pi_{k1} = 1 + \textstyle\sum_{n=1}^{N} \phi_{nk}\,, \quad (8)$$

$$\pi_{k2} = \alpha + \textstyle\sum_{n=1}^{N}\left(\sum_{i=k+1}^{K}\phi_{ni}\right). \quad (9)$$

$$(2)\quad w_1 = s_1 + K\,,$$

$$w_2 = s_2 - \textstyle\sum_{k=1}^{K}\left[\Psi(\pi_{k2}) - \Psi(\pi_{k1}+\pi_{k2})\right],$$

$$\alpha = w_1/w_2\,. \quad (10)$$

$$(3)\quad \psi_k \text{ のパラメータを更新.} \quad (11)$$

導出については [13] を参照のこと. 指数分布族の混合モデルについては, 上記 (3) は自然パラメータの期待値を用いた単純な更新に帰着するが [13], Polya 分布は指数分布族でないため, ここでは [4] に従い, Reversing EM 法によりパラメータ更新を行った. これについて次に述べる.

3 無限混合ディリクレ文書モデル

ディリクレ過程混合モデルは, 混合コンポーネントを無限個生成する確率モデルであるため, コンポーネントの事前分布 G_0 による正則化が不可欠であり, 階層モデルを必要とする.[4]

図 1 のグラフィカルモデルに従う無限混合ディリクレ文書モデルは, 以下のプロセスによってコーパス

$$\mathbf{w} = \mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N \quad (12)$$

$$\mathbf{w}_n = w_1 w_2 \cdots w_{L_n} \quad (13)$$

を生成する.

1. For $k = 1 \ldots \infty$,
   (a) Draw $\mu_k \sim \mathrm{Dir}(\beta)$.
   (b) Draw $\sigma_k \sim \mathrm{Ga}(\gamma)$.
2. For $n = 1 \ldots N$,
   (a) Draw $z \sim \mathrm{Stick}(\alpha)$.
   (b) Draw $\mathbf{p} \sim \mathrm{Dir}(\sigma_z \mu_z)$.
   (c) Draw $\mathbf{w}_n \sim \mathrm{Mult}(\mathbf{p}, L_n)$.

[4] すなわち, G_0 が一様分布ではモデルが求まらない.

ここで Dir, Ga, Mult はそれぞれディリクレ分布, ガンマ分布, 多項分布である.

このとき, 上記 M ステップにおいて π から求めたコンポーネントの変分事後分布 $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_K$ を

$$\lambda_k = \frac{\pi_{k1}}{\pi_{k1}+\pi_{k2}} \prod_{i=1}^{k-1} \frac{\pi_{i2}}{\pi_{i1}+\pi_{i2}} \quad (14)$$

とすれば,

$$L = \log p(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}|\alpha, \beta, \gamma) \quad (15)$$

$$= \log\left[\prod_n \sum_k \lambda_k \mathrm{PL}(\mathbf{w}_n|\mu_k, \sigma_k) \cdot \prod_k \mathrm{Dir}(\mu_k|\beta) \cdot \prod_k \mathrm{Ga}(\sigma_k|\gamma)\right] \quad (16)$$

$$\geq \log\left[\prod_n \prod_k (\lambda_k \mathrm{PL}(\mathbf{w}_n|\mu_k, \sigma_k)/\phi_{nk})^{\phi_{nk}} \cdot \prod_k \mathrm{Dir}(\mu_k|\beta) \cdot \prod_k \mathrm{Ga}(\sigma_k|\gamma)\right] \quad (17)$$

$$= \sum_n \sum_k \phi_{nk}\big[\log \lambda_k/\phi_{nk} + \log \mathrm{PL}(\mathbf{w}_n|\mu_k, \sigma_k)\big] + \sum_k \log \frac{\Gamma(\sum_v \beta_v)}{\prod_v \Gamma(\beta_v)} \prod_v \mu_{kv}^{\beta_v - 1} + \sum_k \log \frac{\gamma_2^{\gamma_1}}{\Gamma(\gamma_1)} \sigma_k^{\gamma_1 - 1} \exp(-\gamma_2 \sigma_k) \quad (18)$$

$$\geq \sum_n \sum_k \phi_{nk}\left[\log \frac{\Gamma(\hat{\sigma}_k)\exp(\hat{\sigma}_k - \sigma_k) b_{nk}}{\Gamma(\hat{\sigma}_k + y_n)} \prod_v c_{nkv} (\sigma_k \mu_{kv})^{a_{nkv}}\right] + \sum_k \sum_v (\beta_v - 1)\log \mu_{kv} + \sum_k \{(\gamma_1 - 1)\log \sigma_k - \gamma_2 \sigma_k\} \quad (19)$$

ここで, [4] と同様に

$$a_{nkv} = [\Psi(\hat{\sigma}_k \hat{\mu}_{kv} + y_{nv}) - \Psi(\hat{\sigma}_k \hat{\mu}_{kv})] \cdot \hat{\sigma}_k \hat{\mu}_{kv}\,, \quad (20)$$

$$b_{nk} = \exp\left[\Psi(w_n + \hat{\sigma}_k) - \Psi(\hat{\sigma}_k)\right], \quad (21)$$

$$c_{nkv} = \frac{\Gamma(\hat{\sigma}_k \hat{\mu}_{kv} + w_{nv})}{\Gamma(\hat{\sigma}_k \hat{\mu}_{kv})} (\hat{\sigma}_k \hat{\mu}_{kv})^{-a_{nkv}} \quad (22)$$

である.

(19) 式を整理すると, μ については

$$p(\mu_k) \sim \mathrm{Dir}(\beta_v + \sum_n \sum_v \phi_{nk} a_{nkv}) \quad (23)$$

であり, 一方, σ については

$$L(\sigma_k) \propto -\textstyle\sum_n \phi_{nk}\sigma_k - \gamma_2 \sigma_k + \sum_n \phi_{nk} \sum_v a_{nkv} \log \sigma_k + (\gamma_1 - 1)\log \sigma_k \quad (24)$$

であるから,

$$p(\sigma_k) \sim \mathrm{Ga}\big(\gamma_1 + \sum_n \sum_v \phi_{nk} a_{nkv}\,, \gamma_2 + \sum_n \phi_{nk}\big) \quad (25)$$

となる.

事前分布パラメータ β については, LDA[10] と同じ Newton 法を用いた. γ についても, 事後分布をもとに Newton 法を以下のように導出することができる.

3.1 Γ 事前分布の Newton-Raphson 法

σ_k の事後分布である (25) 式を $q(\sigma_k) \sim \mathrm{Ga}(p_k, q_k)$ とおくと, 変分ベイズ法においては Γ 事前分布のパラメータ $\gamma = (\gamma_1, \gamma_2)$ は以下の Newton 法により求めることができる.

$$
\begin{aligned}
L &= \left\langle \log p(\boldsymbol{\sigma}|\gamma) \right\rangle_{q(\boldsymbol{\sigma})} = \sum_k \left\langle \log p(\sigma_k|\gamma) \right\rangle_{q(\sigma_k)} \\
&= \sum_k \left\langle \log \frac{\gamma_2^{\gamma_1}}{\Gamma(\gamma_1)} \sigma_k^{\gamma_1 - 1} \exp(-\gamma_2 \sigma_k) \right\rangle_{q(\sigma_k)} \qquad (26) \\
&= K\big(\gamma_1 \log \gamma_2 - \log \Gamma(\gamma_1)\big) \\
&\quad + (\gamma_1 - 1) \sum_k \big(\Psi(p_k) - \log q_k\big) - \gamma_2 \sum_k \frac{p_k}{q_k} \qquad (27)
\end{aligned}
$$

ここで, $\langle \log \sigma_k \rangle = \Psi(p_k) - \log q_k$ であることを用いた. 導出については, 付録 A を参照のこと. よって,

$$
\begin{aligned}
\frac{\partial L}{\partial \gamma_1} &= K\left(\log \gamma_2 - \Psi(\gamma_1)\right) + \sum_k \big(\Psi(p_k) - \log q_k\big)\,, \qquad (28) \\
\frac{\partial L}{\partial \gamma_2} &= K\gamma_1/\gamma_2 - \sum_k p_k/q_k\,, \qquad (29) \\
\frac{\partial^2 L}{\partial \gamma_1^2} &= -K\Psi'(\gamma_1)\,, \quad \frac{\partial^2 L}{\partial \gamma_2^2} = -K\gamma_1/\gamma_2^2\,, \qquad (30) \\
\frac{\partial^2 L}{\partial \gamma_1 \partial \gamma_2} &= \frac{\partial^2 L}{\partial \gamma_2 \partial \gamma_1} = K/\gamma_2 \quad \text{であるから,} \qquad (31)
\end{aligned}
$$

$$
\begin{aligned}
\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}^{new} &= \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} - \begin{bmatrix} -K\Psi'(\gamma_1) & K/\gamma_2 \\ K/\gamma_2 & -K\gamma_1/\gamma_2^2 \end{bmatrix}^{-1} \cdot \\
&\quad \begin{bmatrix} K\big(\log \gamma_2 - \Psi(\gamma_1)\big) + K\sum_k \big(\Psi(q_k) - \log p_k\big) \\ K\gamma_1/\gamma_2 - K\sum_k p_k/q_k \end{bmatrix}. \qquad (32)
\end{aligned}
$$

ゆえに, Newton 更新式は

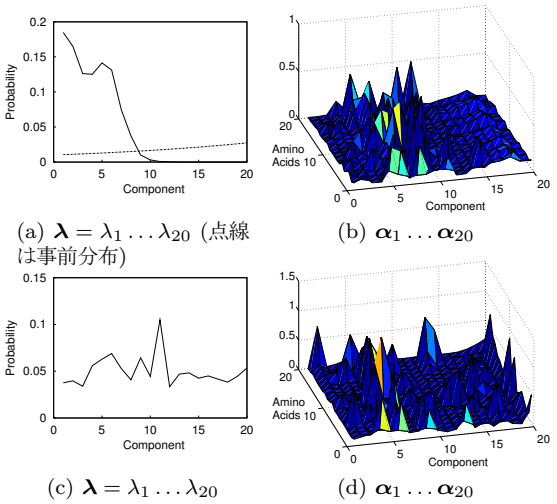
$$
\begin{aligned}
\gamma_1^{new} &= \gamma_1 + \left[\gamma_1 x + \gamma_1 - \gamma_2 y\right]/z\,, \qquad (33) \\
\gamma_2^{new} &= \gamma_2 + \left[\gamma_2 x + \gamma_2 \Psi'(\gamma_1)(\gamma_1 - \gamma_2 y)\right]/z\,. \qquad (34)
\end{aligned}
$$

となる. ただし,

$$
\begin{cases}
x = \log \gamma_2 - \Psi(\gamma_1) + \sum_k \big(\Psi(q_k) - \log p_k\big) & (35) \\
y = \sum_k p_k/q_k & (36) \\
z = \gamma_1 \Psi'(\gamma_1) - 1 & (37)
\end{cases}
$$

である.

ただし, この方法でディリクレ分布の精度パラメータである σ をベイズ推定したところ, 最尤推定である [4] の方法に比べて小さな値に揃いすぎるという現象がみられ, テストデータに対する精度が悪化した. このため, 実際には使用せず, [4] と同様に最尤推定を用いた. 完全な変分ベイズ法を用いて μ および σ を推定したとき, σ に今回のような単純な Γ 事前分布のほかにどんな事前分布を考えるべきかは今後の課題である.



(a) $\boldsymbol{\lambda} = \lambda_1 \ldots \lambda_{20}$ (点線は事前分布)　(b) $\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_{20}$

(c) $\boldsymbol{\lambda} = \lambda_1 \ldots \lambda_{20}$　(d) $\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_{20}$

図 3: アミノ酸配列データのパラメータ推定値. (a)(b): Infinite model, (c)(d): 通常の混合モデル.

3.2 実装と初期化

ディリクレ過程混合モデルの推定法は, 本論文のように変分ベイズ法により決定的に最適化を行う場合は, 事前分布の更新を除いて DM と同様, 本質的にソフト K 平均法と同一である. したがって, 特にパラメータの自由度のきわめて高い自然言語データにおいては, 注意深い初期化が重要となる. 実装に際しては, 予備実験の結果から $\hat{\sigma}$ を $\epsilon \times$ (語彙数) で初期化し ($\epsilon = 0.01$), $\hat{\mu}$ を全体のユニグラム分布を中心として, 精度を語彙数 ×10 としたディリクレ分布によりきわめてわずかにずれた値に初期化した.[5]

α の初期化は混合数の決定に大きな影響を及ぼすため, α の事前分布 $\mathrm{Ga}(s_1, s_2)$ の選択には注意が必要である. [13] では $\mathrm{Ga}(1,1)$ を用いているが, この分布は α について一様ではなく, 尤度に大きな差のない自然言語の場合には常にきわめて小さい混合数が選ばれてしまう. ここでは DP の閾値 K に対し, $\mathrm{Ga}(1, 1/K)$ と初期化して, α に関してほぼ一様な事前分布を用いた.

4 実験

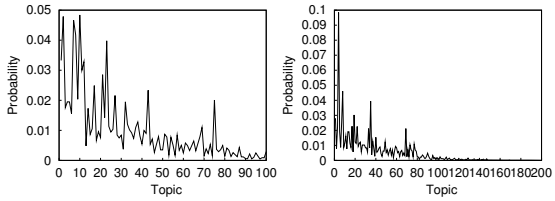
最初にバイオ分野で提案された DM [2] の動機となったアミノ酸配列データ, および小規模コーパスである Cranfield コーパスと文書分類のための中規模コーパスである Reuters-21578 を用いて実験を行った.

4.1 アミノ酸配列データ

DM の最初の目的となったタンパク質のアミノ酸配列データとして, Blocks データベース [17] 中の “k80c” データを用いた.[6] このデータは各配列が, 20 個のアミノ酸の特徴量を表す数値からなる 70,989 配列のデータである. このうちランダムに選択した 1,000 配

[5] 実際にはさらに, ϵ を全体に加えてスムージングを行った.

[6] `http://www.cse.ucsc.edu/compbio/dirichlets/` より入手可能.



(a) $K=100$ (b) $K=200$

図 4: Cranfield のトピック事前分布.

Model	Infinite DM	DM
$K=100$	**444.00**	449.15
$K=200$	**455.78**	465.22

表 1: Cranfield のテストセットパープレキシティ.

列をテストデータ, 残りの 69,989 配列を訓練データとした. データの最大次元数は 20 であるため, DP の切り捨てレベルは $K=20$ と設定した.

図 3(a) に各コンポーネントの事前確率 $\boldsymbol{\lambda}$ を, 図 3(b) に各コンポーネントの持つディリクレ事前分布パラメータ $\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_{20}$ を示す. 図 3(a) では, 点線が事前分布を表す. 図 3(c) および図 3(d) の通常の混合モデルに比べて, 必要な混合数 (最大 11～12) を推定できていることがわかる.

ただし, この時, テストデータのパープレキシティは 5.39 (DP) および 5.35 (通常) であり, 未知データに対する予測精度の面ではほぼ同程度となっていた.

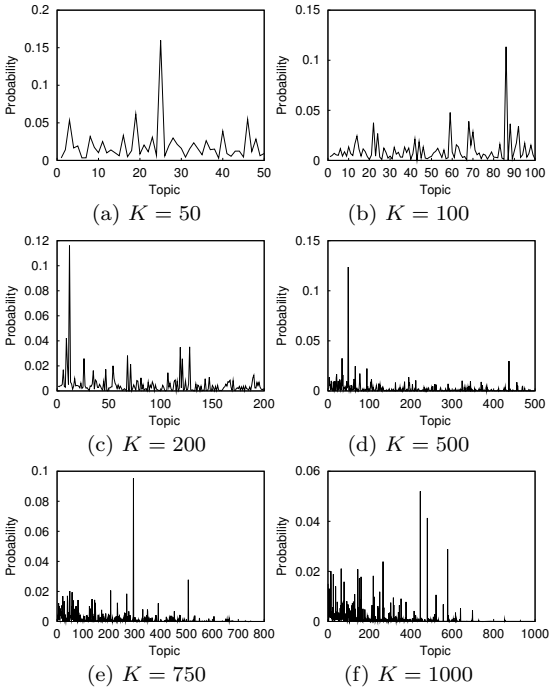
4.2 Cranfield コーパス

Cranfield コーパス [18] は, PLSI, LDA [10] などの文書モデルの評価に使われている小規模なコーパスであり, 1,400 文書, 218,865 単語からなる. このうちランダムに選択した 1,000 文書を訓練データ, 残り 400 文書をテストデータとして実験を行った. 語彙は頻度 2 以上の 5,177 語である. DP の切り捨てレベルを $K=100$ および $K=200$ とした時のトピック事前分布 $\boldsymbol{\lambda}$ を図 4 に示す. 表 1 に, テストデータに対するパープレキシティを 5 回の平均を取って示す.

4.3 Reuters-21578 データセット

中～大規模コーパスでの実験として, 文書分類のためのデータセット Reuters-21578 [19] を使用した. このデータは [20] でも使われている. 'BRIEF' タグの付いた短い記事を除き, ランダムに選択した 500 文書をテストデータ, 残り 18,543 文書, 2,487,020 語を訓練データとした. 語彙は頻度 5 以上の 14,874 語である. 図 5 に, $K=50 \sim 1000$ のそれぞれに設定した時のトピック事前分布 $\boldsymbol{\lambda}$ の推定値を示す. 閾値 K が小さい時は通常の混合モデルと変わらないが, K を充分に大きくすると, このデータに必要なトピック数は最大 500～600 程度であることが見てとれる.

付録 B に, $K=1000$ でのトピック毎の単語出現確率を $p(w|k)/p(w)$ の順にソートした特徴語をトピック 1～5 およびトピック 101～105 について示した.



(a) $K=50$ (b) $K=100$

(c) $K=200$ (d) $K=500$

(e) $K=750$ (f) $K=1000$

図 5: Reuters-21578 の DP によるトピック事前分布.

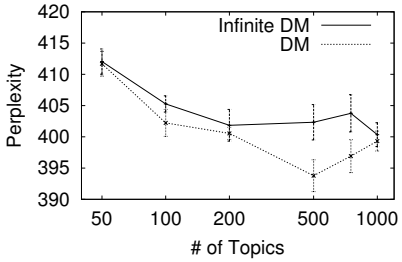


図 6: Reuters-21578 でのテストセット・パープレキシティ.

トピック番号に何ら意味を持たない通常の混合モデルと異なり, 若いトピック番号にこのコーパスにおける中心的な話題が集まり, 古いトピック番号にはより特化したトピックが割り当てられていることがわかる.

図 6 に, テストデータに対するパープレキシティを 5 回の平均を取って示す. 通常の DM では混合数 500 をピークとして過適応が起こるが, 無限 DM ではパープレキシティ自体は通常の DM にわずかに及ばないものの (差は最大でも約 8.5), 混合数を増やしても精度が一定していることがわかる.

5 考察および結論

本論文では, 可算無限個の混合数に事前分布を与えることにより混合数をデータから同時に推定するディリクレ過程混合モデルを DM について適用し, 実テキストデータに対する結果を示した. 通常の混合モデルに比べて最高性能ではわずかに及ばないものの, 混合数の大きい場合にも過適応することなく, 安定した性能を見せることがわかった.

しかしながら, この方法によるメリットは性能そのものよりも, 推定されたトピック事前分布を観察することで, データを表現するのに必要なトピック数が見積もれること, 及び, トピックの間に相対的な依存性があり, その番号にトピック全体での位置が反映されることにあると思われる. ある未知文書のトピック分布を求めたとき, それがトピック事前分布の裾に偏っていれば, その文書は「外れ値」であることの判断基準となる. 通常の混合モデルを用いては, このような判断は行えない.

本論文では DM に対する完全な変分ベイズ法 (貞光 2006; 未発表) は 1 つ欠陥が見つかり使用できなかったため, Reversing EM 法を用いたが, その点が修正されれば 3 章で述べた Newton 法は全体の変分ベイズ法の枠組みの中で自然に導くことができる. 無限個のコンポーネント分布を扱う事前分布には, 今回用いたディリクレ過程以外にもさまざまな可能性があり, 単語単体全域をモデル化する DM をそのような方向に発展させることは将来の課題であると考えられる.

謝辞: 本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[2] K. Sjölander, K. Karplus, M.P. Brown, R. Hughey, R. Krogh, I.S. Mian, and D. Haussler. Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology. *Computing Applications in the Biosciences*, 12(4):327–245, 1996.

[3] 山本 幹雄, 貞光 九月, 三品 拓也. 混合ディリクレ分布を用いた文脈のモデル化と言語モデルへの応用. **情報処理学会研究報告** *2003-SLP-48*, pages 29–34, 2003.

[4] 貞光 九月, 待鳥 裕介, 山本 幹雄. 混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法. **情報処理学会研究報告** *2004-SLP-53*, pages 1–6, 2004.

[5] Hagai Attias. A Variational Bayesian Framework for Graphical Models. In *NIPS 1999*, 1999.

[6] Naonori Ueda and Zoubin Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15:1223–1241, 2002.

[7] Max Welling and Kenichi Kurihara. Bayesian K-Means as a “Maximization-Expectation” Algorithm. In *SIAM Conference on Data Mining SDM2006*, 2006.

[8] 大羽成征. 変分法的ベイズ推定による混合主成分分析, 2001. 奈良先端科学技術大学院大学 情報科学研究科 修士論文, NAIST-IS-MT9951021.

[9] Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. Technical Report 653, Department of Statistics, University of California at Berkeley, 2004.

[12] C. E. Rasmussen. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems 12*, pages 554–560, 2000.

[13] David Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2005.

[14] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[15] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[16] Jayaram Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

[17] Steven Henikoff and Jorja G. Henikoff. Automated assembly of protein blocks for database searching. *Nucletic Acids Research*, 19(23):6565–6572, 1991.

[18] C. J. van Rijsbergen and W. Bruce Croft. Document clustering: An evaluation of some experiments with the cranfield 1400 collection. *Information Processing and Management*, 11(5–7):171–182, 1975. http://ir.dcs.gla.ac.uk/resources/test_collections/cran/.

[19] David D. Lewis. Reuters-21578 text categorization test collection distribution 1.0, 1997. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[20] Kai Yu, Shipeng Yu, and Volker Tresp. Dirichlet Enhanced Latent Semantic Analysis. In *AI & Statistics (AISTATS-2005)*, 2005.

付録 A: $\langle \log \sigma \rangle$ の導出

指数分布族

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^T T(\mathbf{x}) - A(\boldsymbol{\theta})\right] \tag{38}$$

において,

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left\langle T(\mathbf{x}) \right\rangle_{p(\mathbf{x}|\boldsymbol{\theta})} \tag{39}$$

であることは, 簡単な計算によってわかる.

さて, ガンマ分布 $\mathrm{Ga}(s|a,b)$ は指数分布族で,

$$\mathrm{Ga}(s|a,b) = \frac{b^a}{\Gamma(a)} s^{a-1} \exp(-bs) \tag{40}$$

$$= \exp\bigl\{a \log b - \log \Gamma(a) + (a-1) \log s - bs\bigr\} \tag{41}$$

$$= \exp\bigl\{[a-1 \;\; b][\log s \;\; -s]^T + a \log b - \log \Gamma(a)\bigr\} \tag{42}$$

ゆえ, $A(\boldsymbol{\theta}) = \log \Gamma(a) - a \log b$ であるから,

$$\left\langle \log s \right\rangle = \frac{\partial}{\partial (a-1)} A(\boldsymbol{\theta}) \tag{43}$$

$$= \Psi(a) - \log b. \quad \square \tag{44}$$

付録 B: Reuters-21578, $K = 1000$ におけるトピック特徴語

トピック $1 \sim 5$

#	特徴語 Top 10
1	fed, repurchase, arrange, indirect, agreements, customer, temporary, entered, reserve, reserves
2	vehicles, fry, ford, consolidated, geneva, fn, atlas, fold, gm, europe
3	nasdaq, nasd, unusual, exception, activity, explain, stock, delist, exchange, securities
4	economy, growth, economists, economic, forecast, inflation, exports, rise, gross, gdp
5	lending, liquidity, banks, borrow, bankers, bonds, maturity, funds, rates, loans

トピック $101 \sim 105$

#	特徴語 Top 10
101	ounces, mine, feasibility, receipt, ounce, gold, costs, geodome, earnings, reflected
102	nasdaq, composite, showboat, opening, casino, trading, amex, vestor, stock, gaming
103	uranium, oxide, sanctions, african, apartheid, aid, veto, ore, passed, hurt
104	agricultural, agriculture, crops, irrigation, grain, india, harvest, yielding, bales, usda
105	vastagh, potash, nevin, machold, hardie, somc, nissen, marris, wyttenbach, labrecque (空トピック)