混合 Polya 分布に基づくサブトピックを考慮した文書分類手法に関する一考察

1G06H016-9　牛尼　夏海
指導教員　後藤　正幸

1　はじめに

近年，WWW や電子図書館において，膨大な文書が電子的なテキストデータとして蓄積されるようになった．このようなテキストデータの量は増加の一途を辿っており，自動文書分類といったテキストデータの自動処理技術の重要性が増している．このため，様々なテキストデータ分析のための文書モデルや多くの自然言語処理技術が研究されている．中でも，確率的言語モデルの有用性が示され，PLSA，LDA など数多くの提案がなされている．その 1 つにモデルに混合 Polya 分布 [1] を用いた分類手法がある [2]．この手法は，単語の出現確率を“ 点 ”推定をするのではなく，分布で推定して混合をとり，Polya 分布を構成する．Polya 分布によって出現確率の揺らぎを表現しつつ，さらに，複数の内容や話題を含む Polya 分布の混合をとることにより，複数の内容や話題を含む文書のモデルを表現するものである．文書は，その形が様々であるが，このモデルは広い確率領域をモデル化しているので，文書の多様性に対し汎用性が高く，近年注目を集めている．

本研究では，文書分類問題を対象とし，この混合 Polya 分布を用いた分類方法を考える．与えられた文書データを，スポーツ，政治，社会などのカテゴリに分類することを考えた場合，カテゴリはさらに複数のサブカテゴリが混在していることが考えられる．そこで，このサブカテゴリを混合 Polya 分布で推定することを考え，新たな文書に対してサブカテゴリへの帰属度を用いて適切なカテゴリに分類する手法を提案し，その有効性を示す．

2　準備

2.1　文書分類問題

カテゴリ集合を $\mathcal{C}$，カテゴリを $c_k \in \mathcal{C}$，カテゴリ数を C とする．ここで，カテゴリが未知の入力文書 $\boldsymbol{y}$ に対して，推定されたカテゴリを $\hat{c}$ とする．入力文書 $\boldsymbol{y}$ が，カテゴリ c_k に属する確率 (以下，帰属度とする)$P(c_k|\boldsymbol{y})$ が既知であれば，分類精度を最大にする最適な推定量 $\hat{c}$ は，以下の式のように表せる．

$$\begin{aligned}\hat{c} &= \arg\max_{c_k \in \mathcal{C}} P(c_k|\boldsymbol{y}) \\ &= \arg\max_{c_k \in \mathcal{C}} \{\log P(\boldsymbol{y}|c_k) + \log P(c_k)\}. \end{aligned} \quad (1)$$

これより，$P(\boldsymbol{y}|c_k)$ が計算できれば，式 (1) で最適な分類が可能である．すなわち，問題は，いかに確率モデル $P(\boldsymbol{y}|c_k)$ を精度よく構成するかに帰着する．

2.2　混合 Polya 分布

$\mathcal{D}$ を文書集合，$\boldsymbol{y}_i (i = 1, 2, \cdots, D)$ を文書，$y_{iv} (v = 1, 2, \cdots, V)$ を $\boldsymbol{y}_i$ における単語の出現頻度とする．M 個のディリクレ分布を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$ で重み付けした混合ディリクレ分布 $P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}^M)$ を考える．ただし，$\sum_{m=1}^{M} \lambda_m = 1$ とする．また，$\boldsymbol{\alpha}_m$ を第 m コンポーネントのディリクレ分布のパラメータとし，$\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \cdots, \alpha_{mV})$ とする $(\alpha_{mv} > 0)$．さらに，$\boldsymbol{\alpha}^M = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_m, \cdots, \boldsymbol{\alpha}_M)$ である．また，$P_{Mul}(\boldsymbol{y}_i|\boldsymbol{p})$ を $\boldsymbol{p} = (p_1, p_2, \cdots, p_V)$ がパラメータである多項分布とすると，混合 Polya 分布 $P_{PM}(\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha})$ は次式のようになる [1].

$$\begin{aligned} P_{PM}(\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \int P_{Mul}(\boldsymbol{y}_i|\boldsymbol{p}) P_{PM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}^M) d\boldsymbol{p} \\ &= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y_i)} \prod_{v=1}^{V} \frac{\Gamma(y_{iv} + \alpha_{mv})}{\Gamma(\alpha_{mv})}. \end{aligned} \quad (2)$$

ただし，$y_i = \sum_{v=1}^{V} y_{iv}$,$\alpha_m = \sum_{v=1}^{V} \alpha_{mv}$ とする．また，

$$P_{Polya}(\boldsymbol{y}_i; \boldsymbol{\alpha}_m) = \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y_i)} \prod_{v=1}^{V} \frac{\Gamma(y_{iv} + \alpha_{mv})}{\Gamma(\alpha_{mv})}, \quad (3)$$

とおく．ここで $P(\cdot|*)$ は $*$ の条件付確率を表すのに対し $P(\cdot; *)$ は $*$ のパラメータのもとでの確率を表すことにする．

2.3　EM アルゴリズムを用いたパラメータの推定

式 (2) で定義した混合 Polya 分布に，どのトピックに属するかを表す観測できない隠れ変数 $\boldsymbol{Z} = (z_1, z_2, \cdots, z_D)$ を導入したとき，$\boldsymbol{D} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_D)$ が与えられたもとでの，混合ディリクレ分布の対数尤度関数 $\mathcal{L}(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\lambda}, \boldsymbol{\alpha})$ は，

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{i=1}^{D} \log P_{PM}(\boldsymbol{y}_i, z_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}), \quad (4)$$

のように与えられる．この対数尤度を最大にするようなパラメータが求めるパラメータである．いま，対数尤度関数の最大値を直接計算する事が出来ないので，EM アルゴリズムを用いてパラメータの推定を行う．すると，以下のようなパラメータ $\boldsymbol{\lambda}$ と $\boldsymbol{\alpha}$ の更新式を得る．

$$\lambda_m \propto \sum_{i=1}^{D} P_{im}, \quad (5)$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_{i=1}^{D} P_{im} \{\Psi(y_{iv} + \bar{\alpha}_{mv}) - \Psi(\bar{\alpha}_{mv})\}}{\sum_{i=1}^{D} P_{im} \{\Psi(y_i + \bar{\alpha}_m) - \Psi(\bar{\alpha}_m)\}}. \quad (6)$$

ただし，$\bar{\lambda}$，$\bar{\alpha}$ は現在の値，$\Psi(x)$ は digamma 関数と呼ばれる対数ガンマ関数の一次導関数，P_{im} は各文書のトピックを表す隠れ変数 z_i の事後確率であり次式のように定義される．

$$P_{im} = P(z_i = m|\boldsymbol{y}_i; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) = \frac{\bar{\lambda}_m P_{Polya}(\boldsymbol{y}_i; \bar{\boldsymbol{\alpha}}_m)}{\sum_{m=1}^{M} \bar{\boldsymbol{\lambda}}_m P_{Polya}(\boldsymbol{y}_i; \bar{\boldsymbol{\alpha}}_m)}. \quad (7)$$

ただし，P_{Polya} は Polya 分布である．さらに，leaving-one-out 法を用いてより高速な更新式が得られる．

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_{i=1}^{D} P_{im} \{y_{iv}/(y_{iv} - 1 + \bar{\alpha}_{mv})\}}{\sum_{i=1}^{D} P_{im} \{y_i/(y_i - 1 + \bar{\alpha}_m)\}}. \quad (8)$$

3　従来手法

正田ら [2] は，カテゴリ c が与えられたもとでの文書データの分布が，単一の Polya 分布に従うことを仮定し，全カテゴリに対する文書データ全体で混合 Polya 分布となるモデルを用い，未知の文書を分類する手法を提案した．

3.1　混合 Polya 分類器

式 (1) に対し，混合 Polya 分類器では $P(c_k|\boldsymbol{y})$ 入力文書 $\boldsymbol{y}$ に対して次の式のように表せる．

$$P(c_k|\boldsymbol{y};\hat{\boldsymbol{\alpha}}) \propto P(c_k)P_{Polya}(\boldsymbol{y}_i;\hat{\boldsymbol{\alpha}}), \tag{9}$$

よって，この式 (9) に対し，入力文書 $\boldsymbol{y}$ のカテゴリを以下のように決定する．

$$\hat{c} = \arg\max_{c_k \in \mathcal{C}} P(c_k|\boldsymbol{y};\hat{\boldsymbol{\alpha}}). \tag{10}$$

3.2　学習

式 (9) のため $\boldsymbol{\alpha}$，及び $P(c_k)$ のため $\boldsymbol{\lambda}$ を事前に学習する必要がある．ここで，文書集合 $\mathcal{D}$ をカテゴリが既知の文書集合，$\lambda = 1/C$ とし，式 (8) を用いて，α_{mv} を推定する．この際, 混合数 $M = C$ とする．

4　提案手法

4.1　背景

新聞など多くの文書では，ある程度大きなカテゴリがあり，カテゴリ内はそこからさらに細分化されサブカテゴリに分かれてゆく階層的構造をしている．例えば，スポーツ，政治などのカテゴリに対し，さらにスポーツの中には野球，サッカーなどがあるといった具合である．従来研究においては，1 つのカテゴリにつき，1 つの Polya 分布を対応させ，カテゴリ数だけ混合するモデルを考えていた．しかし，1 つのカテゴリには複数のサブカテゴリが混在しているので，仮に文書がカテゴリ c_k に属していたとしても，その文書の単語頻度分布が，必ずしも c_k の単語頻度分布に近いとは限らない．これはサブカテゴリごとに単語頻度分布がバラバラに存在すると考えられる．よって，仮にあるカテゴリに属している文書であっても違うカテゴリに誤分類される要因となる．本研究では，この問題を解決するために，混合 Polya 分布をカテゴリの数だけ用意するモデルを考え，サブカテゴリへの帰属度を用い分類を行う，

4.2　サブカテゴリを考慮した混合 Polya 分類器

式 (1) に対し，サブカテゴリを考慮した混合 Polya 分類器は入力文書 $\boldsymbol{y}$ に対し，推定されたカテゴリを $\hat{c}$ とすると，以下のように表せる．ここで，カテゴリ c_k に対し，それを構成するサブカテゴリ集合を $\mathcal{T}_k = (t_{k1}, t_{k2}, \cdots, t_{ks}, \cdots, t_{kS})$ とする．

$$P(t_{ks}|\boldsymbol{y};\hat{\boldsymbol{\alpha}}) \propto P(t_{ks})P_{Polya}(\boldsymbol{y}|\hat{\boldsymbol{\alpha}}), \tag{11}$$

$$\hat{c} = \arg\max_{c_k \in \mathcal{C}} \{\max_{s} P(t_{ks}|\boldsymbol{y};\hat{\boldsymbol{\alpha}})\}. \tag{12}$$

4.3　学習

文書集合より 1 カテゴリ (カテゴリ c_k とする) の文書集合を取り出し，これを文書 $\mathcal{D}_k$ とし，これを従来手法における $\mathcal{D}$ として，式 (5)〜(8) を用いて，パラメータを推定する．カテゴリ集合をサブカテゴリ集合に置き換え，サブカテゴリ集合を $\mathcal{T}_k$ とする．ここで，S はカテゴリ c_k のサブカテゴリ数とする．また，推定されたパラメータを $\hat{\boldsymbol{\alpha}}_k$,$\hat{\lambda}_k$ とする．これを全てのカテゴリに対して行う．

5　評価実験

提案手法の有効性を検討するために新聞記事を用いて評価実験を行う．なお評価基準として以下の式に示す分類精度を用いた．

$$\text{分類精度} = \frac{\text{正しく分類された個数}}{\text{総分類個数}}. \tag{13}$$

5.1　実験条件

実験データは，毎日新聞 2005 年のデータ 4 カテゴリ（経済，芸能，スポーツ，社会）を使用する．データから各カテゴリ 1000 文書ずつの 4000 文書を 1 セットとし，これを 10 セット作り，9 セットを学習データ，1 セットをテストデータとする．この条件で 10 パターン実験する 10 分割ローテーション法実験をし，この 10 回の分類精度の平均で比較する．

5.2　実験結果及び考察

実験結果を以下の図 1 に示す．

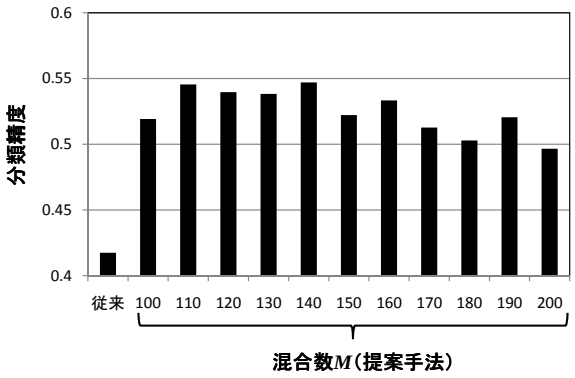

図 1．各手法による分類精度

実験を行った結果，提案手法は従来手法に分類精度で上回った．提案手法においては，混合数が増加すると 140 以降 1 つの山ごとに分類精度が下がる傾向がある．これは，サブトピックを増やしていくと，トピック同士の特徴が近くなり，誤分類を起こしやすくなったのではないかと考えられる．

6　まとめと今後の課題

本研究では，混合 Polya 分布を用いた文書分類において，カテゴリごとに 1 つの混合 Polya 分布を与え，サブカテゴリへの帰属度で分類する分類法を提案し，その有効性を示した．

今後の課題は，サブカテゴリの混合数について，カテゴリごとに混合数を変えることや，その混合数を自動的に推定することなど，検討が必要である．

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- , No.9, pp. 1771–1779, 2005.

[2] 正田備也, 高須淳宏, 安達淳,“混合ディリクレ分布を用いた文書分類の精度について,”情報処理学会論文誌, Vol.48, No.SIG 11(TOD 34), pp.14–26, 2007.