潜在トピックを考慮した未観測なカテゴリを含む文書集合の自動分類手法の提案

1X10C121-7　山本祐生
指導教員　後藤正幸

1　はじめに

近年，蓄積された膨大な電子文書に対する自動分類手法が注目されている．文書分類では，あらかじめカテゴリが既知な文書集合 (以下，学習用文書) を用いて，分類の対象であるカテゴリが未知な新規文書 (以下，分類対象文書) を，テキストの内容に基づいて分類する．ただし，一般的な分類対象文書の分類先は，学習用文書内に出現している既存のカテゴリのいずれかであることを前提としているため，分類対象文書に学習用文書内に出現していない未観測なカテゴリ (以下，未観測カテゴリ) に所属する文書が存在している場合，うまく分類が行えないといった問題が生じる．そのため，新たに与えられた入力文書が既存カテゴリのいずれかに所属する文書であるのか，もしくは未観測カテゴリに所属する文書であるのかを自動的に判別する分類手法が望まれる．

そこで荒川ら [1] は，文書が各カテゴリごとに異なる単一な Polya 分布に従って出現するという仮定に基づく混合 Polya 分布を用い文書の生成確率モデルを表現することで，未観測カテゴリを含む文書集合の分類手法を提案している．しかし，新聞記事等にみられる文書のカテゴリは，1 つのカテゴリ内に複数の潜在トピックが存在していると考えられる．例えば,「スポーツ」という大域的なカテゴリに対し,「サッカー」や「野球」等の観測されない潜在的なトピックが考えられる．荒川らの手法は各カテゴリごとに単一な Polya 分布を仮定していることから，これらの潜在トピックを考慮していない．そこで本研究では荒川らの手法にこれらの潜在トピックを新たに導入することで，予測精度の向上を目指す．また，新聞記事データを用いて分類精度を検証する実験を行い，本手法の有効性を示す．

2　準備

2.1　記号の定義

文書の特徴量は，形態素解析によって分割された各単語の出現頻度によって構成されている．すなわち，全文書中に含まれる単語の集合を $\mathcal{W} = \{w_1, w_2, \cdots, w_V\}$ とすれば，第 n 文書 $\boldsymbol{x}_n$ は単語頻度ベクトル $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{nV})$ という特徴量によって表される．ただし，w_v は第 v 単語，x_{nv} は第 n 文書中の単語 w_v の出現回数である．

2.2　混合 Polya 分布

文書が潜在カテゴリごとに異なる Polya 分布から生成されるという仮定に基づいた，混合 Polya 分布によるモデル化が提案されている [2]．混合数を M，混合比を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$，$m$ 番目の Polya 分布のパラメータを $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \cdots, \alpha_{mV})$ とすると，文書 $\boldsymbol{x}_n$ の生成する確率モデルは混合 Polya 分布 $P_{PM}(\boldsymbol{x}_n; \boldsymbol{\lambda}, \boldsymbol{\alpha})$ によって表される．

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

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

3　従来手法

3.1　概要

荒川らの手法 [1] では，各カテゴリに単一な Polya 分布を仮定したうえで，未観測カテゴリに対応するために $M(M \geq K+1)$ を混合数とする混合 Polya 分布を設定する．M 個の Polya 分布のうち，K 個の Polya 分布がそれぞれ K 個の既存カテゴリ $c_1, c_2, \cdots, c_K$ の文書をモデル化し，残りの $M-K$ 個の Polya 分布が「未観測」というカテゴリ c_{K+1} の文書をモデル化する．

3.2　モデルの学習

N_L 件の文書からなる文書の帰属先カテゴリが既知な学習用文書の集合 $\mathcal{D}_L = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{N_L})$ と，N_T 件の文書からなる帰属先カテゴリが未知な分類対象文書の集合 $\mathcal{D}_T = (\boldsymbol{x}_{N_L+1}, \boldsymbol{x}_{N_L+2}, \cdots, \boldsymbol{x}_{N_L+N_T})$ が独立であるとすれば，対数尤度は以下の式 (2) で表される．ただし，δ_{nk} は学習用文書 $\boldsymbol{x}_n$ の帰属するカテゴリが c_k と一致するときに 1，それ以外で 0 をとるインジケータ関数である．

$$\begin{aligned}\log L(\mathcal{D}_L, \mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \sum_{n=1}^{N_L} \log \sum_{k=1}^{K} \delta_{nk} \lambda_k P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_k) \\ &\quad + \sum_{n=N_L+1}^{N_L+N_T} \log \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_m) \end{aligned} \quad (2)$$

上式を極大化するパラメータ λ_m,α_{mv} は EM アルゴリズムによって推定を行う．EM アルゴリズムとは潜在変数の期待値を求める E-Step と，対数尤度を最大にするパラメータを求める M-Step を交互に繰り返すことでパラメータの最尤推定値を得る学習法である．

3.3　分類

EM アルゴリズムによるモデルの学習が完了した段階で，各分類対象文書ごとに Polya 分布への帰属確率を計算し，これを用いて各カテゴリへの帰属確率を計算する．第 n 文書 $\boldsymbol{x}_n$ の m 番目の Polya 分布への帰属確率 $P(z_n = m \mid \boldsymbol{x}_n)$ は式 (3) により求められる．

$$P(z_n = m \mid \boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}, \bar{\lambda}_m) = \frac{P(z_n = m) P(x_n \mid z = m, \bar{\alpha}_{mv}, \bar{\lambda}_m)}{\sum_{m=1}^{M} P(z_n = m) P(x_n \mid z = m, \bar{\alpha}_{mv}, \bar{\lambda}_m)} \quad (3)$$

ただし，$\bar{\alpha}_{mv}, \bar{\lambda}_m$ は EM アルゴリズムにおける更新前のパラメータである．1 番目から K 番目の Polya 分布への帰属確率をカテゴリ c_1 から c_K への帰属確率とし，$K+1$ 番目から M 番目の Polya 分布までの帰属確率の和を「未観測」というカテゴリ c_{K+1} への帰属確率とする．すなわち，各カテゴリへの帰属確率は式 (4) によって与えられる．

$$P(c_k \mid \boldsymbol{x}_n) = \begin{cases} P(z_n = k \mid \boldsymbol{x}_n) & (1 \leq k \leq K) \\ \displaystyle\sum_{m=K+1}^{M} P(z_n = k \mid \boldsymbol{x}_n) & (k = K+1) \end{cases} \quad (4)$$

各分類対象文書を，$K+1$ 個のカテゴリの中で帰属確率が最大のカテゴリ $\hat{c}_k$ へ分類する．

4　提案手法

4.1　概要

新聞記事等にみられる文書のカテゴリは，カテゴリ「スポーツ」における「サッカー」や「野球」のように，大域的な情報をもつカテゴリが上位に，それを細分化する複数の観測されないトピックが下位に存在するといった，階層的構造

をもつと考えられる．同一のカテゴリに所属する文書集合においても，どのトピックを持つかによって，文書中の単語出現頻度が異なると仮定する．荒川らの手法では 1 つの既存カテゴリに対し，単一な Polya 分布を仮定しているため，潜在トピックごとの単語出現頻度の差異が考慮されていない．そこで本研究では，各カテゴリ c_k に対し新たに S 個の潜在トピック $t_{k1}, t_{k2}, \cdots, t_{kS}$ を導入し，各潜在トピックに対し Polya 分布 $P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_{ms})$ を仮定する．そして，新たな混合比 $\boldsymbol{\pi}_k = (\pi_{k1}, \pi_{k2}, \cdots, \pi_{kS})$ により S 個の Polya 分布を混合し，1 つのカテゴリ c_k の確率モデルを表す．それにより，複数の潜在トピックで構成される 1 つのカテゴリをより幅広い確率領域で表現できる．ただし，$\sum_{s=1}^{S} \pi_{ks} = 1$ とする．

4.2 モデルの学習

提案手法の対数尤度は以下のように定義される．

$$\begin{aligned}\log L(\mathcal{D}_L, \mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\pi}, \boldsymbol{\alpha}) \\ = \sum_{n=1}^{N_L} \log \sum_{k=1}^{K} \delta_{nk} \lambda_k \sum_{s=1}^{S} \pi_{ks} P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_{ks}) \\ + \sum_{n=N_L+1}^{N_L+N_T} \log \sum_{m=1}^{M} \lambda_m \sum_{s=1}^{S} \pi_{ms} P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_{ms}) \quad (5)\end{aligned}$$

上式を極大化するパラメータ $\lambda_m, \pi_{ms}, \alpha_{msv}$ の最適値は EM アルゴリズムによって推定される．

E-Step)

m 番目のカテゴリにおける s 個目の Polya 分布を表す潜在変数 z_{nm} の事後確率 P_{nms} は次式によって定義される．

$$\begin{aligned}P_{nms} &= P(z_{nm} = s \mid \boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms}, \bar{\pi}_{ms}) \\ &= \frac{\bar{\pi}_{ms} P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms})}{\sum_{s=1}^{S} \bar{\pi}_{ms} P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms})} \quad (6)\end{aligned}$$

P_{nms} が与えられたもとで，$\boldsymbol{x}_n$ を生成した m 個目の混合 Polya 分布を表す潜在変数 z_n の事後確率 P_{nm} は以下のように定義される．

$$\begin{aligned}P_{nm} &= P(z_n = m \mid \boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms}, \bar{\lambda}_m, \bar{\boldsymbol{\pi}}_m) \\ &= \frac{\bar{\lambda}_m \sum_{s=1}^{S} \bar{\pi}_{ms} P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms})}{\sum_{m=1}^{M} \bar{\lambda}_m \sum_{s=1}^{S} \bar{\pi}_{ms} P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_{ms})} \quad (7)\end{aligned}$$

M-Step)

各パラメータ $\lambda_m, \pi_{ms}, \alpha_{msv}$ の更新式は以下のように与えられる．

$$\lambda_m = \frac{1}{N_L + N_T} \left(\sum_{n=1}^{N_L} \delta_{nm} + \sum_{n=N_L+1}^{N_L+N_T} P_{nm} \right) \quad (8)$$

$$\pi_{ms} = \frac{M}{N_L + N_T} \left(\sum_{n=1}^{N_L} \delta_{nm} P_{nms} + \sum_{n=N_L+1}^{N_L+N_T} P_{nm} P_{nms} \right) \quad (9)$$

$$\alpha_{msv} = \bar{\alpha}_{msv} \times \frac{\sum_{n=1}^{N_L} \delta_{nm} P_{nms} \beta_{nmsv} + \sum_{n=N_L+1}^{N_L+N_T} P_{nm} P_{nms} \beta_{nmsv}}{\sum_{n=1}^{N_L} \delta_{nm} P_{nms} \gamma_{nm} + \sum_{n=N_L+1}^{N_L+N_T} P_{nm} P_{nms} \gamma_{nm}} \quad (10)$$

ただし，

$$\beta_{nmsv} = \frac{x_{nv}}{x_{nv} - 1 + \bar{\alpha}_{msv}} \quad (11)$$

$$\gamma_{nms} = \frac{x_n}{x_n - 1 + \bar{\alpha}_{ms}} \quad (12)$$

とし，$\bar{\lambda}_m, \bar{\pi}_{ms}, \bar{\alpha}_{msv}$ は更新前のパラメータである．

5 実験

5.1 実験概要

毎日新聞 (2005 年版) の新聞記事データ 1600 件を用いて実験を行う．社説，国際，経済，家庭，科学，芸能，スポーツ，社会の計 8 カテゴリを用い，記事 200 件ずつをランダムに抽出した．そのうち 1 つのカテゴリを未観測カテゴリと規定し，残りのカテゴリを既存カテゴリとした．既存カテゴリは 100 件を学習用文書，残りの 100 件の分類対象文書とし，未観測カテゴリについては 200 件すべてを分類対象文書とした．また，提案手法の潜在トピック数 S は 1〜4 として実験を行う．$S = 1$ の場合が従来手法である．モデルの混合数 M を 8〜25 の範囲で変化させ，既存カテゴリ，未観測カテゴリの両カテゴリに対する分類精度を評価する．ただし，分類精度は次式によって定義する．

$$\text{全体の分類精度} = \frac{\text{正しく分類できたテストデータ数}}{\text{全テストデータ数}} \quad (13)$$

5.2 実験結果

表 1: 実験結果

全体の分類精度				
混合数	S=1(従来)	S=2	S=3	S=4
8	0.6459	**0.6609**	0.6597	0.6580
9	0.6570	**0.6648**	0.6594	0.6641
10	0.6477	0.6517	**0.6611**	0.6509
11	0.6475	0.6615	**0.6621**	0.6582
12	0.6532	0.6558	**0.6559**	0.6527
13	0.6352	0.6500	**0.6528**	0.6501
14	0.6499	**0.6551**	0.6461	0.6493
15	0.6438	**0.6538**	**0.6538**	0.6522
16	0.6450	0.6497	0.6517	**0.6522**
17	0.6414	0.6560	0.6493	**0.6570**
18	0.6356	0.6533	**0.6551**	0.6490
19	0.6400	**0.6527**	0.6446	0.6513
20	0.6449	**0.6514**	0.6505	0.6481
21	0.6402	0.6432	**0.6577**	0.6480
22	0.6441	0.6551	**0.6568**	0.6482
23	0.6458	0.6464	**0.6546**	0.6504
24	0.6443	**0.6512**	0.6503	0.6469
25	0.6468	0.6501	**0.6512**	0.6507

実験結果より，とりわけ潜在トピック数 2，及び 3 と設定した場合，提案手法が従来手法と比較して優れた性能を示していることがわかる．

5.3 考察

追加実験より，各カテゴリに単一な Polya 分布を仮定した従来手法と比較し，提案手法では潜在トピックを考慮することにより，未観測カテゴリの分類精度が向上していることが確認できた．これにより，全体として優れた分類精度が得られたことが考えられる．また，提案手法では潜在トピック数を 4 個以上に設定した場合に，従来手法の分類精度と同程度の結果が示された．これは，潜在トピック数を適切な個数よりも多く設定したために，パラメータ数が多すぎることによる過学習に陥り，それが誤分類の原因となってしまったためだと考えられる．

6 まとめと今後の課題

各カテゴリに Polya 分布を仮定した従来手法に対し，提案手法では新たに各カテゴリごとに潜在トピックを導入したモデルの構築を行い，実験により，潜在トピックを考慮した提案手法が従来手法に比べ，分類精度の観点で優れていることを示した．また，今後の課題として潜在トピック数を自動的に判断するモデルの提案などが挙げられる．

参考文献

[1] 荒川貴紀，三川健太，後藤正幸，“未観測カテゴリ—を含む文書データの自動分類手法に関する研究,”電子情報通信学会論文誌 (D)，vol.J96-D, No.8, pp.1956-1959, 2013.

[2] 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.