未観測カテゴリーを含む文書データの自動分類手法に関する研究

荒川　貴紀[†]　　三川　健太[††a)]（正員）
後藤　正幸[††]（正員）

A Study on Document Classification Method with Containig Unknown Categories

Takanori ARAKAWA[†], *Nonmember*, Kenta MIKAWA[††a)], and Masayuki GOTO[††], *Members*

[†] 早稲田大学大学院創造理工学研究科，東京都
Graduate School of Creative Science and Engineering, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan
[††] 早稲田大学理工学術院，東京都
School of Creative Science and Engineering, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan
a) E-mail: kmikawa@aoni.waseda.jp

あらまし　本研究では，学習データ中に全く現れなかった未知のカテゴリー（未観測カテゴリー）の文書が出現するような状況での文書分類問題を対象とし，確率モデルに基づいた新しい分類手法を提案する．

キーワード　文書分類，未観測カテゴリー，混合 Polya 分布，EM アルゴリズム

1. まえがき

近年，蓄積された大量の文書を効率的に分類，整理するための技術の一つとして，機械学習に基づく自動文書分類が注目されている．一般の文書分類問題では，分類対象文書は学習用文書が帰属するカテゴリーのいずれかに帰属することを前提としている [1]．しかし，分類対象文書中には学習用文書中に全く現れていない未知のカテゴリー（以下，未観測カテゴリーと呼ぶ）に属する文書が存在している可能性もある(注1)．このような場合，既存のカテゴリーのいずれにも分類すべきでないような文書であったとしても，通常の分類手法では既存のカテゴリーのいずれかに分類されてしまうため，望ましい分類結果が得られないことになる．そのため，与えられた文書が既存カテゴリーのいずれかに帰属するのか，あるいはそれ以外の未観測カテゴリーに帰属するのかを正しく判別できる自動分類手法が望まれる．

データが未観測カテゴリーに帰属するかどうかの判別に関連する研究として，異常なデータを検出して弾くための異常値検出 [2] や自動分類におけるリジェクトルール [3] に関する研究が挙げられる．これらの分野の手法はあらかじめ何らかの尺度に対してしきい値を設定し，データがそのしきい値を超えたか否かによって異常/正常（リジェクト/アクセプト）を判定し後続の処理を行う．そのため，しきい値によって検出されるデータの数を調整できる一方で，最適なしきい値を設定するのが困難な場合もある．

そこで本研究では，データが未観測カテゴリーに帰属するかどうかの判別をしきい値に基づいて行うのではなく，データが未観測カテゴリーに帰属する確率をモデル化し，それをもとに分類誤り率最小化の観点から一意に判別を行える方法を提案する．提案手法は，確率論的手法である混合 Polya 分布による文書分類手法 [4] を拡張し，既存カテゴリーへの帰属確率に加えて未観測カテゴリーへの帰属確率も考慮したモデル化を行う．また，半教師あり学習 [6] の枠組みを援用することにより，分類対象文書のテキスト情報を有効に活用して未観測カテゴリーの性質を推定する．新聞記事データを用いた分類実験により，提案手法の有効性を検証する．

2. 混合 Polya 分布による文書モデル

単語頻度ベクトルにより表現された文書 $\boldsymbol{x} = (x_1, x_2, \cdots, x_V)$ に対する確率モデルとして，混合 Polya 分布によるモデル化が提案されている [7]．ただし，文書の特徴量 x_i は，単語 w_i の出現回数を表す．

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．

$$P_{PM}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{x}; \boldsymbol{\alpha}_m), \qquad (1)$$

ただし，$\sum_{m=1}^{M} \lambda_m = 1$，$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_M)$，$\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \cdots, \alpha_{mV})$，$\alpha_m = \sum_{v=1}^{V} \alpha_{mv}$，$x = \sum_{v=1}^{V} x_v$ であり，m 番目の Polya 分布は次式で与える．

$$P_{Polya}(\boldsymbol{x}; \boldsymbol{\alpha}_m) = \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + x)} \prod_{v=1}^{V} \frac{\Gamma(x_v + \alpha_{mv})}{\Gamma(\alpha_{mv})}.$$

貞光ら [7] では，M 個の潜在トピックが観測できない場合の学習として，EM アルゴリズムによる推定法を示している．また，正田ら [4] はカテゴリー既知の文書集合を用いて混合 Polya 分布を教師あり学習することで，文書分類へ適用する方法を提案している．

(注1)：例えばニュース記事をトピックに従って自動分類する際，学習用文書中には “経済”，“スポーツ”，“芸能”，“社会” の 4 カテゴリーしか存在していなかったとしても，分類対象文書中にはそれらとは全く異なるカテゴリー，例えば “科学” に関する記事が存在している可能性がある．

3. 提案手法

3.1 問題の定式化

カテゴリーが既知である N_L 件の文書からなる学習用文書集合を $\mathcal{D}_L = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_{N_L}, y_{N_L})\}$，カテゴリーが未知である N_T 件の文書からなる分類対象文書の集合を $\mathcal{D}_T = \{\boldsymbol{x}_{N_L+1}, \boldsymbol{x}_{N_L+2}, \cdots, \boldsymbol{x}_{N_L+N_T}\}$ とする．$\boldsymbol{x}_n$ は第 n 文書の単語頻度ベクトル，y_n は第 n 文書が帰属するカテゴリーである．カテゴリーの集合を $\mathcal{C} = \{c_1, c_2, \cdots, c_K, c_{K+1}\}$ とする．ここで c_{K+1} は既存カテゴリーに帰属しない文書が帰属する “未観測” という名のカテゴリーである．すなわち，第 n 文書が学習用文書であるならば $y_n \in \mathcal{C} \setminus \{c_{K+1}\}$ であり，第 n 文書が分類対象文書であるならば $y_n \in \mathcal{C}$ である．

学習用文書集合 $\mathcal{D}_L$ と分類対象文書集合 $\mathcal{D}_T$ を合わせた全文書中に含まれる異なり単語の集合を $\mathcal{W} = \{w_1, w_2, \cdots, w_V\}$ とする．w_v は第 v 単語，V は異なり単語数を表す．第 n 文書の単語頻度ベクトルを $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{nV})$ とする．ただし，x_{nv} は第 n 文書中の単語 w_v の出現回数である．本提案の目的は，学習用文書集合 $\mathcal{D}_L$ と分類対象文書集合 $\mathcal{D}_T$ が与えられたもとで，各分類対象文書に対し，それぞれが帰属するカテゴリー $y_{N_L+1}, y_{N_L+2}, \cdots, y_{N_L+N_T}$ を推定することである．分類先の候補は，K 個の既存カテゴリーに “未観測” という一つのカテゴリーを加えた $K+1$ 個のカテゴリーである．

3.2 確率モデルの設定

本研究では，正田ら [4] の混合 Polya 分布による従来研究のモデルをベースに，未観測カテゴリーの存在を前提としたモデルへと拡張する．従来は，K 個のカテゴリーのいずれかに帰属する文書集合に対して混合数 K の混合 Polya 分布を設定しているが，以下では，未観測カテゴリーに対応するために既存のカテゴリー数 K よりも 1 以上大きな値 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 個の既存カテゴリー $c_1, c_2, \cdots, c_K$ の文書をモデル化し，残りの $M-K$ 個の Polya 分布が未観測カテゴリー c_{K+1} の文書をモデル化する(注2)．

3.3 モデルの半教師あり学習

学習用文書を教師ありデータ，分類対象文書を教師なしデータとして半教師あり学習の枠組みを用いてモデルを学習する．学習用文書集合 $\mathcal{D}_L$ と分類対象文書集合 $\mathcal{D}_T$ が独立であるとすれば，$\mathcal{D}_L$ と $\mathcal{D}_T$ に対する混合 Polya 分布の対数ゆう度は以下の式で表される．

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

ただし，δ_{nk} は第 n 文書が帰属するカテゴリー y_n 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．上式を最大化するパラメータ $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ は，教師なし学習と教師あり学習を組み合わせた EM アルゴリズムによって推定することができる．$\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ の更新式を以下に示す．

$$\lambda_m = \frac{1}{N_L + N_T} \left(\sum_{n=1}^{N_L} \delta_{nm} + \sum_{n=1}^{N_T} P_{nm} \right), \qquad (3)$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_{n=1}^{N_L} \delta_{nm} \beta_{nmv} + \sum_{n=1}^{N_T} P_{nm} \beta_{nmv}}{\sum_{n=1}^{N_L} \delta_{nm} \gamma_{nm} + \sum_{n=1}^{N_T} P_{nm} \gamma_{nm}}.$$

ただし，

$$\beta_{nmv} = \frac{x_{nv}}{x_{nv} - 1 + \bar{\alpha}_{mv}}, \quad \gamma_{nm} = \frac{x_n}{x_n - 1 + \bar{\alpha}_m},$$

とし，$\bar{\boldsymbol{\lambda}}$, $\bar{\boldsymbol{\alpha}}$ は更新前のパラメータ値とする．これらは文献 [5] におけるパラメータの導出アルゴリズムを混合 Polya 分布で実行したものとなっている．

3.4 分類アルゴリズム

EM アルゴリズムによるモデルの学習が完了した段階で，分類対象文書ごとに各 Polya 分布への帰属確率を計算し，これをもとにして各カテゴリーへの帰属確率を計算する．第 n 文書の m 番目の Polya 分布への帰属確率 $P(z_n = m \mid \boldsymbol{x}_n)$ は以下の式により求められる．

$$\begin{aligned}&P(z_n = m \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \\ &= \frac{P(z_n = m) P(\boldsymbol{x}_n \mid z = m; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})}{\sum_{m=1}^{M} P(z = m) P(\boldsymbol{x}_n \mid z = m; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})} \qquad (4)\end{aligned}$$

各文書の各カテゴリーへの帰属確率を求めるには，1 番目から K 番目までの Polya 分布への帰属確率をそのままカテゴリー 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カテゴリー c_{K+1} への帰属確率とする．

(注2)：未観測カテゴリーに属する文書がみな似たような性質でまとまって存在しているような場合には，$M - K = 1$ で十分であると考えられるが，複数の未観測カテゴリーが存在すると想定される場合には，$M - K \geq 2$ である必要があると考えられる．

すなわち，

$$P(y_n = c_k \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) = \begin{cases} P(z_n = k \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}), & (1 \le k \le K) \\ \sum_{m=K+1}^{M} P(z_n = k \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}), & (k = K+1) \end{cases}$$

である．各分類対象文書を $(K+1)$ 個のカテゴリーの中で帰属確率が最大のカテゴリーへ分類する．すなわち，文書 $\boldsymbol{x}_n$ が帰属するカテゴリー y_n を

$$\hat{y}_n = \arg\max_{y_n \in \mathcal{C}} P(y_n \mid \boldsymbol{x}_n), \tag{5}$$

として推定する．

4. 実　　験

4.1 実験条件

毎日新聞（2005 年版）の記事データ集の中から社説，国際，経済，家庭，科学，芸能，スポーツ，社会の 8 カテゴリーについて，各カテゴリーごとに 200 件ずつ合計 1600 件の記事をランダムに取得した．8 カテゴリーのうち未観測カテゴリーが一つの場合（すなわち $K = 7$），二つの場合（$K = 6$），三つの場合（$K = 5$）についてそれぞれ実験を行った．

未観測カテゴリーについては 200 件を全て分類対象文書とし，未観測カテゴリー以外の K 個のカテゴリーについてはランダムに選択した 100 件を学習用文書，残りの 100 件を分類対象文書とした．未観測カテゴリーとして扱うカテゴリーを変えて全ての組合せで実験を行い，平均の分類精度を求めた(注3)．各文書の単語頻度ベクトルの作成に際しては，形態素解析により文書を単語単位に分割し，動詞，名詞，形容詞以外の単語と全体での出現回数が 10 以下の低頻度語は不要語として除外した．結果，用いた単語数 V は 4015 となった．

4.2 比較手法

提案手法の有効性を検証するために，以下の二つの手法と分類精度の比較を行う．

- 比較手法 1

リジェクトルールに基づき未観測カテゴリーを判別する．分類器は学習用文書を用いて教師付き学習をした，混合数 K の混合 Polya 分布を用いる．既存カテゴリーに対する確信度 $\max_{y \in \mathcal{C}} P(y \mid \boldsymbol{x})$ のしきい値に基づき，未観測カテゴリーかどうかを判別する処理を加える．分類器が出力した確信度があるしきい値以上の文書についてはそのまま既存のカテゴリーのいずれかに分類し，確信度がしきい値より低い文書については K 個のカテゴリーのいずれにも属さないとみなし，未観測カテゴリー c_{K+1} に分類する(注4)．

- 比較手法 2

余弦類似度に基づき未観測カテゴリーを判別する．K 個の既存カテゴリーについて学習用文書を用いて各カテゴリーの重心を求め，重心との余弦類似度が最も大きなカテゴリーに文書を分類する．既存カテゴリーとの類似度の最大値があるしきい値以上の文書については K 個の既存カテゴリーのいずれかに分類し，しきい値以下の文書については未観測カテゴリー c_{K+1} に分類する．

4.3 実験結果と考察

実験データに対し各手法を適用した結果について，未観測カテゴリーが一つの場合（図 1），二つの場合の結果（図 2），三つの場合の結果（図 3）を以下に示す．

二つの比較手法についてはしきい値の値を変え，提案手法については混合数 M を変えて，既存カテゴリーに対する分類精度，未観測カテゴリーに対する分類精度，全体に対する分類精度をそれぞれ算出した．M の値は $K+1$ から 20 の範囲で値を変えて実験を行った．

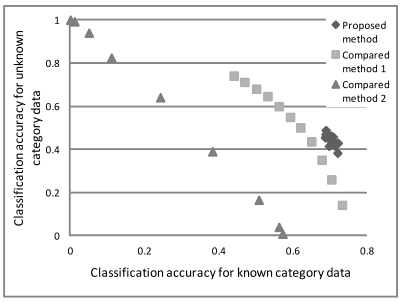

図 1　未観測カテゴリーが一つの場合の各手法の分類精度
Fig. 1　Classification accuracy ($K = 7$).

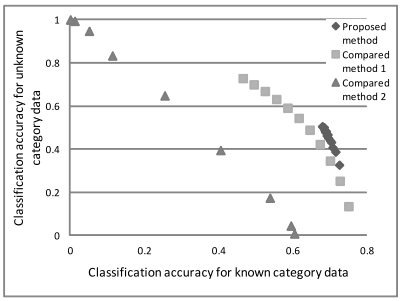

図 2　未観測カテゴリーが二つの場合の各手法の分類精度
Fig. 2　Classification accuracy ($K = 6$).

(注3)：組合せ数は，未観測カテゴリーが一つの場合は 8 通り，二つの場合は 28 通り，三つの場合は 56 通りである．また，分類精度は，（正しく分類できた文書の数）/（分類対象となった文書の数）で与えられる．

(注4)：確信度は 0 以上 1 以下の値をとるが，混合 Polya 分類器を用いると多くのデータに対して 1 に非常に近い値となるため，確信度のしきい値も 0.999 など，1 に近い値とする必要がある．

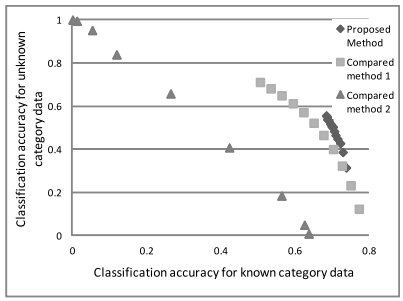

図 3 未観測カテゴリーが三つの場合の各手法の分類精度
Fig. 3 Classification accuracy ($K = 5$).

表 1 提案手法の分類精度（未観測カテゴリー数 1）
Table 1 Classification accuracy of proposed method ($K = 7$).

M	Classification accuracy for known category data	Classification accuracy for unknown category data	Classification accuracy for all data
8	0.7209	0.3833	0.6459
9	0.7220	0.4296	0.6570
10	0.7137	0.4167	0.6477
11	0.7061	0.4423	0.6475
12	0.7090	0.4577	0.6532
13	0.6979	0.4158	0.6352
14	0.7031	0.4635	0.6499
15	0.6993	0.4492	0.6438
16	0.6963	0.4652	0.6450
17	0.6960	0.4506	0.6414
18	0.6873	0.4548	0.6356
19	0.6886	0.4700	0.6400
20	0.6894	0.4890	0.6449

図 1〜図 3 のいずれにおいても，既存カテゴリー（未観測カテゴリー）に対する分類精度が提案手法と比較手法とで同程度になるしきい値で比較すると，未観測カテゴリー（既存カテゴリー）に対する分類精度は提案手法が比較手法を上回ることが分かる．以上より提案手法の有効性が確認できた．提案手法は既存カテゴリー情報をもつ学習用文書と，分類対象文書に含まれるテキスト情報を併せて用いて未観測カテゴリーの性質を推定しようとする方法であるため，より効果的に未観測カテゴリーを判別できたと考えられる．

また，未観測カテゴリーが一つの場合についての各手法の結果を表 1〜表 3 に示す．表 2，表 3 より，二つの比較手法はしきい値の設定に関して，未観測カテゴリーに対する分類精度と既存カテゴリーに対する分類精度がトレードオフの関係にあることが分かる．また，提案手法における混合数 M は分類器の構築の際に適当に設定する必要があるが，表 1 より，混合数 M を変化させても分類精度に著しい変化は見られなかった．よって，実用上は既存カテゴリー数 K より大きめの値を設定しておけばおおむね問題ないといえる．

5. む す び

本研究では，学習用文書と分類対象文書があらかじめ一括で与えられる状況に対して，未観測カテゴリーの存在を前提とした文書分類手法を提案した．新聞記事データを用いた文書分類実験により，分類対象文書を効果的に用いて未観測カテゴリーの性質を推定する提案手法の有効性を確認した．

提案手法では，K 個の既存カテゴリーをそれぞれ一つの Polya 分布でモデル化したが，各カテゴリーが潜在トピックをもつモデルの検討は今後の課題である．

表 2 比較手法 1 の分類精度（未観測カテゴリー数 1）
Table 2 Classification accuracy of compared method 1 ($K = 7$).

Threshold	Classification accuracy for known category data	Classification accuracy for unknown category data	Classification accuracy for all data
0.9	0.7333	0.1423	0.6020
0.99	0.7045	0.2602	0.6057
0.999	0.6784	0.3510	0.6056
0.9999	0.6504	0.4365	0.6029
0.99999	0.6209	0.5010	0.5943
0.999999	0.5932	0.5496	0.5835
0.9999999	0.5626	0.5994	0.5707
0.99999999	0.5321	0.6460	0.5574
0.999999999	0.5020	0.6804	0.5417
0.9999999999	0.4699	0.7119	0.5237
0.99999999999	0.4414	0.7410	0.5080

表 3 比較手法 2 の分類精度（未観測カテゴリー数 1）
Table 3 Classification accuracy of compared method 2 ($K = 7$).

Threshold	Classification accuracy for known category data	Classification accuracy for unknown category data	Classification accuracy for all data
0.1	0.5728	0.0088	0.4475
0.2	0.5630	0.0404	0.4469
0.3	0.5088	0.1654	0.4325
0.4	0.3838	0.3902	0.3852
0.5	0.2432	0.6406	0.3315
0.6	0.1113	0.8246	0.2698
0.7	0.0506	0.9404	0.2483
0.8	0.0118	0.9927	0.2298
0.9	0.0006	1.0000	0.2227

文 献

[1] 石田栄美，“テキスト自動分類の概要,” 情報の科学と技術，vol.56, no.10, pp.469–474, 2006.
[2] 竹内純一，山西健司，“データマイニングにおける統計的外れ値検出,” 応用数理，vol.11, no.2, pp.163–167, 2001.
[3] C. Chow, “On optimum recognition error and reject tradeoff,” IEEE Trans. Inf. Theory, vol.16, no.1, pp.41–46, 1970.
[4] 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,” 情処学論，vol.48, no.SIG11(TOD34), pp.14–26, 2007.
[5] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” Mach. Learn., vol.39, no.2-3, pp.103–134, 2000.
[6] M. Seeger, “Learning with labeled and unlabeled data,” Technical Report, University of Edingurgh, 2001.
[7] 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,” 信学論（D-II），vol.J88-D-II, no.9, pp.1771–1779, Sept. 2005.

（平成 25 年 4 月 9 日受付）