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

情報数理応用研究　　5211C002-0　荒川貴紀
指導教員　後藤正幸

A New Classification Method of Document Data with Unknown Category

ARAKAWA Takanori

1　はじめに

近年，情報技術の発達により，大量の文書がテキストデータとして電子的に蓄積されるようになった．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ある．機械学習に基づく自動文書分類とは，属するカテゴリが既知である文書（以下，学習用文書と呼ぶ）を用いて分類器を学習し，新たに入力されるカテゴリが未知の文書（以下，分類対象文書と呼ぶ）を自動的に分類する技術である．

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のいずれかに帰属することを前提としている [1]．しかしながら，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（以下，未観測カテゴリと呼ぶ）に属する文書が存在している可能性もある．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，例えば “科学”に関する記事が存在している可能性がある．このような場合，既存のカテゴリのいずれにも分類すべきでないような文書であったしても，通常の分類手法では必ず既存のカテゴリ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] に関する研究があげられる．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閾値を設定し，データがその閾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，閾値によって検出されるデータの数を調整できる一方で，最適な閾値を設定するのが困難な場合もある．

閾値に基づくこれらの従来研究とは異なり，本研究では確率モデルに基づくアプローチをとる．すなわち，データが未観測カテゴリに帰属するかどうかの判別を閾値に基づいて行うのではなく，データが未観測カテゴリに帰属する確率をモデル化し，それを基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，既存カテゴリへの帰属確率に加えて未観測カテゴリ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 [5] 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　準備

本研究では未観測カテゴリを含む文書分類問題に対し，混合 Polya 分布に基づく半教師あり学習の枠組みを援用した文書分類手法を提案する．本研究の立場を明確にするため，以下ではまず通常の文書分類問題を定義する．加えて，本研究で用いる混合 Polya 分布のパラメータ推定について，教師なし学習，教師あり学習の場合をそれぞれ説明する．これらをふまえ，本研究が対象とする問題への展開を述べる．

2.1　通常の文書分類問題

カテゴリの集合を $\mathcal{C} = \{c_1, c_2, \cdots, c_K\}$ とし，カテゴリが既知である N 件の文書からなる学習用文書の集合を $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$ とする．$\boldsymbol{x}_n$ は第 n 文書の特徴ベクトル，y_n は第 n 文書が帰属するカテゴリである（$y_n \in \mathcal{C}$）．学習用文書集合 $\mathcal{D}$ 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．ただし，w_v は第 v 単語，V は異なり単語数を表す．文書の特徴量としては単語の出現頻度が用いられることが多い．このとき，第 n 文書は単語頻度ベクトル $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{nV})$ として定義される．ただし，x_{nv} は第 n 文書の単語 w_v の出現回数である．

文書分類問題は，学習用文書集合 $\mathcal{D}$ が与えられた下で，新たに与えられるカテゴリ未知の入力文書 $\boldsymbol{x}$ が帰属するカテゴリ y を $\mathcal{C}$ の中から推定する問題である．入力文書の属するカテゴリの推定値を $\hat{y}$ とする．いま，入力文書 $\boldsymbol{x}$ が帰属するカテゴリ y の確率分布 $P(y \mid \boldsymbol{x})$ が既知であれば，分類精度を最大にする最適な推定量 $\hat{y}$ は，ベイズの定理より

$$\begin{aligned} \hat{y} &= \arg\max_{y \in \mathcal{C}} P(y \mid \boldsymbol{x}) \\ &= \arg\max_{y \in \mathcal{C}} \log P(y) \log P(\boldsymbol{x} \mid y), \end{aligned} \tag{1}$$

のように表せる．

2.2　混合 Polya 分布

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（多項分布のパラメータが混合ディリクレ分布に従う場合の合成分布）によるモデル化が提案されている [6]．混合 Polya 分布は，データがある潜在クラスに属しており，属する潜在クラスによって異なる Polya 分布から生成されるという仮定に基づく確率分布であり，多様な文書の話題をモデル化できる．

いま，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．

$$P_{PM}(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{x};\boldsymbol{\alpha}_m)$$
$$= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + x)} \prod_{v=1}^{V} \frac{\Gamma(x_v + \alpha_{mv})}{\Gamma(\alpha_{mv})}. \quad (2)$$

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．

2.3 EM アルゴリズムによる混合 Polya 分布の教師なし学習 [6]

以下では，教師なしの文書集合 $\mathcal{D}_U = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ から混合 Polya 分布のパラメータを最尤推定する方法について述べる．パラメータ $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ の下で $\mathcal{D}_U$ に対する対数尤度 $\log L(\mathcal{D};\boldsymbol{\lambda},\boldsymbol{\alpha})$ は

$$\log L(\mathcal{D};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \sum_{n=1}^{N} \log P_{PM}(\boldsymbol{x}_n;\boldsymbol{\lambda},\boldsymbol{\alpha}), \quad (3)$$

と表せる．この対数尤度を最大化するようなパラメータ $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ を求めたいが，解析的に推定量を導くことはできないため，EM アルゴリズムによってパラメータを推定する．EM アルゴリズムでは，まずパラメータの初期値を任意に与え，対数尤度が増加するように更新するというプロセスを繰り返すことにより，求めたい推定値の近似解を得ることができる．その際のパラメータ $\boldsymbol{\lambda}$ と $\boldsymbol{\alpha}$ の更新式は以下のように表すことができる．

$$\lambda_m = \frac{1}{N} \sum_{n=1}^{N} P_{nm}, \quad (4)$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_{n=1}^{N} P_{nm}\{x_{nv}/(x_{nv} - 1 + \bar{\alpha}_{mv})\}}{\sum_{n=1}^{N} P_{nm}\{x_n/(x_n - 1 + \bar{\alpha}_m)\}}. \quad (5)$$

ただし，$x_n = \sum_{v=1}^{V} x_{nv}$ であり，$\bar{\boldsymbol{\lambda}}$，$\bar{\boldsymbol{\alpha}}$ は更新前のパラメータ値である．P_{nm} は第 n 文書を生成した Polya 分布を表す潜在変数 z_n の事後確率であり，以下のように定義される．

$$P_{nm} = P(z_n = m \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$$
$$= \frac{\bar{\lambda}_m P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_m)}{\sum_{m=1}^{M} \bar{\lambda}_m P_{Polya}(\boldsymbol{x}_n; \bar{\boldsymbol{\alpha}}_m)}. \quad (6)$$

2.4 混合 Polya 分布の教師あり学習による文書分類への適用

カテゴリ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，文書分類への適用が可能である [4]．教師ありの文書集合 $\mathcal{D}_S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$ が与えられたもとでの混合 Polya 分布の対数尤度は以下の式で表される．

$$\log L(\mathcal{D};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \delta_{nk} \lambda_k P_{Polya}(\boldsymbol{x}_n;\boldsymbol{\alpha}_k). \quad (7)$$

ただし，δ_{nk} は第 n 文書が帰属するカテゴリ 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．

対数尤度を最大化するようなパラメータ $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ の推定値は以下のようにして求められる [2]．$\boldsymbol{\lambda}$ については

$$\lambda_k = \frac{1}{N} \sum_{n=1}^{N} \delta_{nk}, \quad (8)$$

と解析的に解が求まり，$\boldsymbol{\alpha}$ については

$$\alpha_{kv} = \bar{\alpha}_{kv} \frac{\sum_{n=1}^{N} \delta_{nk}\{x_{nv}/(x_{nv} - 1 + \bar{\alpha}_{kv})\}}{\sum_{n=1}^{N} \delta_{nk}\{x_n/(x_n - 1 + \bar{\alpha}_k)\}}, \quad (9)$$

という更新式を反復することで近似解を得られる．

新たな入力文書 $\boldsymbol{x}$ が帰属するカテゴリ y は，$\boldsymbol{x}$ がカテゴリ c_k に帰属する確率 $P(y = c_k \mid \boldsymbol{x})$ を最大にする c_k として

$$\hat{y} = \arg\max_{y \in \mathcal{C}} P(y \mid \boldsymbol{x})$$
$$= \arg\max_{y \in \mathcal{C}} \lambda_k P_{Polya}(\boldsymbol{x};\boldsymbol{\alpha}_k), \quad (10)$$

と推定する．

2.5 関連研究と本研究への展開

通常の文書分類問題では分類対象文書は K 個 の既存カテゴリのいずれかに帰属するのに対し，本研究では分類時に未観測カテゴリが存在する状況を前提とする．未観測カテゴリの判別に関連する研究分野として，正常なデータとは性質が大きく異なるデータを異常データとして検出する異常値検出 [2] や分類器が曖昧な出力をしたデータに関しては分類を行わずに結果を棄却するためのリジェクトルール [3] に関する研究があげられる．これらの分野の手法の多くは，あらかじめ何らかの尺度に対して閾値を設定し，データがその閾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）を判定する．異常値として検出されるデータの数やリジェクトルールにより棄却されるデータの数は，事前に設定した閾値に大きく依存する．そのため，閾値によって出力の傾向を調整できる一方で，最適な閾値を設定することが困難な場合もある．

閾値に基づくこれらの研究とは異なり，本研究では確率モデルに基づくアプローチをとる．すなわちデータが未観測カテゴリに帰属するかどうかの判別を閾値に基づいて行うのではなく，データが未観測カテゴリに帰属する確率をモデル化し，それを基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示す．

3 提案手法

3.1 問題の定式化

以下では，未観測カテゴリを含む文書データの分類問題について述べる．また，後述するカテゴリ集合や異なり単語集合については，本来なら 2.1 節とは異なる記号を用いて定義するべきであるが，表記が煩雑になるため同一の種類の変数に関しては同一の記号によって再定義する．

カテゴリが既知である 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である．カテゴリの集合を $\mathcal{C} = \{c_1, c_2, \cdots, c_K, c_{K+1}\}$ 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 $y_{N_L+1}, y_{N_L+2}, \cdots, y_{N_L+N_T}$ を推定することである．分類先の候補は，K 個の既存カテゴリに “未観測”という 1 つのカテゴリを加えた $K+1$ 個のカテゴリである．

3.2 確率モデルの設定

本研究では，2.4 に示した混合 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，未観測カテゴリに対応するために既存のカテゴリ数 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 $c_1, c_2, \cdots, c_K$ 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 c_{K+1} の文書をモデル化する．未観測カテゴリ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，複数の潜在カテゴリ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．

3.3 モデルの学習

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 分布の対数尤度は以下の式で表される．

$$\begin{aligned}
&\log L(\mathcal{D}_L, \mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&= \log L(\mathcal{D}_L; \boldsymbol{\lambda}, \boldsymbol{\alpha}) + \log L(\mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&= \sum_{n=1}^{N_L} \log \sum_{k=1}^{K} \delta_{nk} \lambda_k P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_k) \\
&+ \sum_{n=N_L+1}^{N_L+N_T} \log \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{x}_n; \boldsymbol{\alpha}_m). \qquad (11)
\end{aligned}$$

ただし，δ_{nk} は第 n 文書が帰属するカテゴリ 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，2.3 節に示した教師なし学習法と 2.4 節に示した教師あり学習法を組み合わせた 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．

$$\lambda_m = \frac{1}{N_L + N_T} \left(\sum_{n=1}^{N_L} \delta_{nm} + \sum_{n=1}^{N_T} P_{nm} \right), \qquad (12)$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_{n=1}^{N_L} \delta_{nm} \beta_{nmv} + \sum_{n=1}^{N_T} P_{nm} \beta_{nmv}}{\sum_{n=1}^{N_L} \delta_{nm} \gamma_{nm} + \sum_{n=1}^{N_T} P_{nm} \gamma_{nm}}. \qquad (13)$$

ただし，

$$\beta_{nmv} = \frac{x_{nv}}{x_{nv} - 1 + \bar{\alpha}_{mv}}, \qquad (14)$$

$$\gamma_{nm} = \frac{x_n}{x_n - 1 + \bar{\alpha}_m}. \qquad (15)$$

とし，$\bar{\boldsymbol{\lambda}}$，$\bar{\boldsymbol{\alpha}}$ は更新前のパラメータ値とする．

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．

$$\begin{aligned}
&P(z_n = m \mid \boldsymbol{x}_n; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \\
&= \frac{P(z_n = m) P(\boldsymbol{x}_n \mid z = m; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})}{\sum_{m=1}^{M} P(z = m) P(\boldsymbol{x}_n \mid z = m; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})} \qquad (16)
\end{aligned}$$

各文書の各カテゴリへの帰属確率を求めるには，1 番目の Polya 分布から K 番目の Polya 分布への帰属確率をそのままカテゴリ c_1 から c_K への帰属確率とする．$K+1$ 番目の Polya 分布から M 番目の Polya 分布までの帰属確率の和を未観測カテゴリ c_{K+1} への帰属確率とする．すなわち，

$$\begin{aligned}
&P(y_n = c_k \mid \boldsymbol{x}_n) \\
&= \begin{cases} P(z_n = k \mid \boldsymbol{x}_n) & (1 \leq k \leq K) \\ \displaystyle\sum_{m=K+1}^{M} P(z_n = k \mid \boldsymbol{x}_n) & (k = K+1), \end{cases} \qquad (17)
\end{aligned}$$

である．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へ分類する．すなわち，文書 $\boldsymbol{x}_n$ が帰属するカテゴリ y_n を

$$\hat{y}_n = \arg\max_{y_n \in \mathcal{C}} P(y_n \mid \boldsymbol{x}_n), \qquad (18)$$

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経済，科学，芸能，スポーツ，社会の 5 つのカテゴリについて，各カテゴリごとに 200 件ずつ合計 1000 件の記事をランダムに取得した．5 つのカテゴリのうちの 1 つを未観測カテゴリとして扱い (すなわち K=4)，未観測カテゴリについては 200 件を全て分類対象文書とし，未観測カテゴリ以外の 4 カテゴリ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．未観測カテゴリとして扱うカテゴリを変えて 5 通りの実験を行い，平均の分類精度を求めた．分類精度は以下の式により定義される．

$$\text{分類精度} = \frac{\text{カテゴリを正しく推定できた文書の数}}{\text{分類対象となった文書の数}}. \quad (19)$$

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 5 以下の低頻度語は不要語として除外した．用いた単語数 V は 2961 となった．

4.2 比較手法

比較手法として，リジェクトルールに基づき未観測カテゴリを判別する手法を用いる．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 分布とする．既存カテゴリに対する確信度 $\max_{y\in\mathcal{C}} P(y \mid \boldsymbol{x})$ の閾値に基づき，未観測カテゴリ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閾値以上の文書についてはそのまま既存のカテゴリのいずれかに分類し，確信度が閾値より低い文書については K 個のカテゴリのいずれにも帰属しないとみなし，未観測カテゴリ c_{K+1} に分類する．一般に，確信度の閾値が高いと既存カテゴリの文書を誤って未観測カテゴリに分類することが多くなり，逆に閾値が低いと未観測カテゴリの文書を誤って既存カテゴリのいずれかに分類することが多くなる．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，確信度の閾値として $0.9, 0.99, 0.999, \cdots, 0.999999999$ の 9 通りの値で実験を行った．

4.3 実験結果と考察

実験データに対し比較手法を適用した結果を表 1 に，提案手法を適用した結果を表 2 に示す．比較手法については確信度の閾値を変え，提案手法については混合数 M を変えて，既存カテゴリに対する分類精度，未観測カテゴリに対する分類精度，全体に対する分類精度をそれぞれ算出した．$K = 4$ であるため，M の値は 5 以上に設定する．表 1 より比較手法の分類精度（全体）は閾値の値が 0.99999 のときに最大値をとるが，表 2 より提案手法は M=5 から M=10 のいずれの場合においてもそれより高い値をとることがわかる．また，比較手法の閾値に関して，既存カテゴリに対する分類精度と未観測カテゴリに対する分類精度はトレードオフの関係にある．そのため，既存カテゴリに対する分類精度と未観測カテゴリに対する分類精度をそれぞれ別々に考えると比較手法が提案手法を上回る場合もあるが，既存カテゴリ（未観測カテゴリ）に対する分類精度が提案手法と同程度になる閾値で比較すると，未観測カテゴリ（既存カテゴリ）に対する分類精度は提案手法が比較手法を大きく上回る．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を判別できたと考えられる．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 2 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．

表 1 比較手法の分類精度

閾値	既存カテゴリ	未観測カテゴリ	全体
0.9	**0.801**	0.175	0.592
0.99	0.766	0.321	0.618
0.999	0.739	0.445	0.641
0.9999	0.704	0.532	0.647
0.99999	0.681	0.597	**0.653**
0.999999	0.651	0.649	0.650
0.9999999	0.625	0.703	0.651
0.99999999	0.590	0.738	0.639
0.999999999	0.550	**0.764**	0.621

表 2 提案手法の分類精度

混合数 M	既存カテゴリ	未観測カテゴリ	全体
5	**0.776**	0.735	0.762
6	0.767	0.786	**0.773**
7	0.755	0.775	0.761
8	0.753	0.791	0.766
9	0.742	**0.795**	0.759
10	0.738	0.770	0.748

5 まとめと今後の課題

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の性質を推定する提案手法の有効性を確認した．

また，本研究の提案手法は K 個の既存カテゴリをそれぞれ 1 つの Polya 分布でモデル化したが，一般には同じカテゴリに属する文書であってもトピックによって出現単語の傾向は異なると考えられるので，各カテゴリを複数の Polya 分布でモデル化する方法の検討が今後の課題である．

参考文献

[1] 石田栄美, “テキスト自動分類の概要,” 情報の科学と技術, 56 巻 10 号, 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,” 日本応用数理学会, Vol.11, No.2, 71–75, 2001.

[3] C. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Transactions on Information Theory*, 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,” 情報処理学会論文誌, Vol.48, pp. 14–26, 2007.

[5] Seeger, M. “Learning with labeled and unlabeled data,” Technical report, University of Edingurgh, 2001.

[6] 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.