電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.131 No.8 pp.1459-1466　　DOI: 10.1541/ieejeiss.131.1459

論　文

学習データが少量しかない場合の文書分類に関する一考察

非会員　前田　康成*　　非会員　吉田　秀樹*
非会員　鈴木　正清*　　非会員　松嶋　敏泰**

A Note on Document Classification with Small Training Data

Yasunari Maeda*, Non-member, Hideki Yoshida*, Non-member, Masakiyo Suzuki*, Non-member,
Toshiyasu Matsushima**, Non-member

（2010 年 3 月 19 日受付，2010 年 8 月 27 日再受付）

Document classification is one of important topics in the field of NLP (Natural Language Processing). In the previous research a document classification method has been proposed which minimizes an error rate with reference to a Bayes criterion. But when the number of documents in training data is small, the accuracy of the previous method is low. So in this research we use estimating data in order to estimate prior distributions. When the training data is small the accuracy using estimating data is higher than the accuracy of the previous method. But when the training data is big the accuracy using estimating data is lower than the accuracy of the previous method. So in this research we also propose another technique whose accuracy is higher than the accuracy of the previous method when the training data is small, and is almost the same as the accuracy of the previous method when the training data is big.

キーワード：文書分類，学習データ，事前分布，事後分布
Keywords：document classification, training data, prior distribution, posterior distribution

1. まえがき

文書分類は，自然言語処理において重要な問題の 1 つである。文書分類については，従来から数多くの研究が行われている。大別すると，距離に基づく文書分類方法と，確率モデルに基づく文書分類方法とに分けられる。

距離に基づく文書分類方法には，テンプレートマッチング法[1]を用いた文書分類方法や，k 最近傍法[1][2]を用いた文書分類方法などがある。さらに，テンプレートマッチング法の 1 種にベクトル空間法[3]~[6]がある。ベクトル空間法は，文書分類や自然言語処理に限らず多くの分野で，よく研究および実用化されている方法である。自然言語処理という 1 分野だけでも，情報検索，情報フィルタリング，機械翻訳など多くの実用例が報告されている。ベクトル空間法を用いた文書分類方法は実用面での評価が高い。それは，直感的に理解し易く，言語的な性質やさまざまな経験則での微調整を行い易く，かつ計算量が小さいという長所があるためである。しかし，その分類精度に理論的な保証はない。

他方，確率モデルに基づく文書分類方法[3][6]~[8]についても，いろいろな研究が行われている。その中で最もよく研究されているのは，文書のクラスと文書中にあるキーワードが多項分布に従って生起する確率モデルを採用して，学習データから求めた最尤推定値を利用して，新規文書（新規に分類したい文書）を分類する方法である。確率モデルを採用すると，分類精度を理論的に誤り率（分類を間違えてしまう確率）で保証できる。学習データによる最尤推定値を用いる方法では，多項分布の真のパラメータが既知の場合に誤り率を最小にできる文書分類方法に，多項分布の最尤推定値をあてはめて，新規文書を分類する。よって，無限の学習データを仮定すると最尤推定値が真のパラメータに一致するので，漸近的に誤り率を最小にできる。しかし，有限の学習データに対しては，理論的な精度保証ができない。

従来研究[9]でも，文書のクラスとキーワードが多項分布に従って生起する確率モデルを採用している。この従来研究では，統計的決定理論[10]に基づいて文書分類問題の定式化を行い，有限の学習データに対して誤り率をベイズ基準の

* 北見工業大学情報システム工学科
〒090-8507　北海道北見市公園町 165 番地
Dept. of Computer Science, Kitami Institute of Technology
165, Koen-cho, Kitami-shi, Hokkaido 090-8507, Japan
** 早稲田大学応用数理学科
〒169-8555　東京都新宿区大久保 3-4-1
Department of Applied Mathematics, Waseda University
3-4-1, Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

もとで最小にすることを保証する文書分類方法を提案している。この従来研究で提案された文書分類方法（以下，ベイズ最適な従来方法と呼ぶ）は，漸近的には，最尤推定値を用いる文書分類方法と同様の性質も保持する。さらに，実際の電報データを用いた比較実験によって，ベクトル空間法を用いた文書分類方法よりも高い分類精度が報告されている。しかし，その分類精度は学習データ量に依存する性質があり，学習データが少量の場合には分類精度はどうしても低くなってしまう（誤り率が高くなる）。なお，学習データ量に分類精度が依存するのはベクトル空間法を用いた文書分類方法でも同様であり，学習データが少量の場合でもベイズ最適な従来方法の分類精度がベクトル空間法を用いた文書分類方法よりも高いことが報告されている。文書分類問題では多項分布を支配する真のパラメータが未知なので，従来研究では多項分布を支配するパラメータに対する事前分布を導入している。しかし，事前に何も情報が無いことを示す無情報な事前分布を採用しているため，学習データが少量の場合には分類精度は低くなる。

そこで，本研究では，新規文書や学習データとは性質が異なるが，ある程度似ているような既存の文書データを，事前分布の推定用データとして利用する。そして，事前分布の推定用データを利用して推定した事前分布を，ベイズ最適な従来方法の事前分布として使用することにより，新規文書や学習データとは性質が異なるが，ある程度似ている既存データ利用の有効性を検証する。無情報な事前分布を使用することは，学習データが無いもとで全てのクラスとキーワードの生起の仕方が等しく尤もらしいと仮定することに相当する。これに対し，学習データとは別の既存データを事前分布の推定用データとして利用することは，学習データが無いもとでクラスとキーワードの生起の仕方の尤もらしさに既存データに従って重み付けをすることに相当する。よって，既存データの性質が学習データや新規文書に似ているほど，学習データが少量の場合の分類精度が高くなると期待される。

最初に本研究では，学習データが少量の場合に，本研究における1番目の検証例の分類精度が学習データのみ利用する場合の分類精度よりも高くなることを確認した。しかし，学習データが増加した場合には，1番目の検証例の分類精度は学習データのみ利用する場合の分類精度よりも低くなることも確認した。これは，事前分布の推定用データの影響が，学習データが少量の場合には分類精度を高くするように作用し，学習データが増加した場合には逆に分類精度を低くするように作用したことによる。この原因は過学習である。

次に本研究では，1番目の検証例の事前分布の設定に改良を加えた，2番目の検証例について検証する。2番目の検証例では，学習データが少量の場合には学習データのみ利用するときよりも高い分類精度を達成し，かつ学習データが増加した場合には学習データのみ利用するときと同程度の分類精度を達成することを確認した。

第2章で従来研究と本研究に共通する定義と従来研究（ベイズ最適な従来方法）について説明し，第3章で事前分布の推定用データを利用する1番目の検証例と2番目の検証例を説明し，第4章で本研究における2番目の検証例と学習データのみ利用する場合の従来方法との比較実験結果について紹介し，第5章でまとめと今後の課題について述べる。

2. 文書分類問題の定義と従来研究

〈2･1〉 文書分類問題の定義 ここでは，従来研究[9]と本研究に共通する文書分類問題の定義を行う。

c_i，$c_i \in C$ は文書のクラスを示し，C はクラスの集合，$C=\{c_1,c_2,\ldots,c_{|C|}\}$ である。$|C|$ はクラスの集合 C の要素数を示す。key_i，$key_i \in KEY$ は文書中にあるキーワードを示し，KEY はキーワードの集合，$KEY=\{key_1,key_2,\ldots,key_{|KEY|}\}$ である。$p(c_i|\theta)$ はクラス c_i が生起する確率を示し，$p(c_i|\theta)$ は連続パラメータ θ，$\theta \in \Theta$ によって支配されている。Θ はパラメータの集合を示す。真のパラメータ θ^*，$\theta^* \in \Theta$ は未知である。$p(key_j|c_i,\psi)$ はクラス c_i の文書中でキーワード key_j が生起する確率を示し，$p(key_j|c_i,\psi)$ は連続パラメータ ψ，$\psi \in \Psi$ によって支配されている。Ψ はパラメータの集合を示す。真のパラメータ ψ^*，$\psi^* \in \Psi$ は未知である。doc^L は，未知パラメータ θ^* と ψ^* について学習するために利用する L 個の文書からなる学習データで，次式で示される。

$$doc^L=(x,y^n)^L=(x_1,y^{n_1})(x_2,y^{n_2})\cdots(x_L,y^{n_L}) \quad \cdots\cdots(1)$$

ただし，x_i は学習データ中にある i 番目の文書のクラス，n_i は i 番目の文書中にあるキーワードの数，y^{n_i} は i 番目の文書中にあるキーワードの系列，

$$y^{n_i}=y_{i,1}y_{i,2}\cdots y_{i,n_i} \quad \cdots\cdots(2)$$

$y_{i,j}$ はキーワード系列 y^{n_i} 中にある j 番目のキーワードである。学習データは既知である。学習データ doc^L が生起する確率は次式で計算できる。

$$\begin{aligned} p(doc^L|\theta,\psi) &= p\left((x,y^n)^L|\theta,\psi\right) \\ &= p\left((x_1,y^{n_1})(x_2,y^{n_2})\cdots(x_L,y^{n_L})|\theta,\psi\right) \\ &= \prod_{i=1}^{L}\left(p(x_i|\theta)\prod_{j=1}^{n_i}p(y_{i,j}|x_i,\psi)\right) \end{aligned} \quad \cdots\cdots(3)$$

doc' は新規文書で，次式で示される。

$$doc'=(x',y'^{n'}) \quad \cdots\cdots(4)$$

ただし，x' は新規文書 doc' のクラス，n' は新規文書 doc' 中にあるキーワードの数，$y'^{n'}$ は新規文書 doc' 中にあるキーワードの系列である。クラス x' は未知で，キーワード系列 $y'^{n'}$ は既知である。新規文書 doc' が生起する確率は次式で計算できる。

$$p(doc'|\theta,\psi)=p(x',y'^{n'}|\theta,\psi)=p(x'|\theta)\prod_{i=1}^{n'}p(y'_i|x',\psi) \quad \cdots\cdots(5)$$

上記より，文書分類問題は，学習データ doc^L と新規文書

のキーワード系列 $y'^{n'}$ を受け取ったもとで，新規文書のクラス x' を推定する問題として定義できる。

〈2･2〉　従来研究　ここでは，従来研究[9]について説明する。

従来研究では，統計的決定理論に基づきベイズ基準のもとで誤り率を最小にする文書分類方法（ベイズ最適な従来方法）が提案されている。誤り率を最小化するために次式の損失関数が用いられている。

$$L\left(d(y'^{n'},doc^{L}),x'\right)=\begin{cases}1, & d(y'^{n'},doc^{L})\neq x'\\ 0, & d(y'^{n'},doc^{L})=x'\end{cases} \quad \cdots\cdots(6)$$

ただし，$d(y'^{n'},doc^{L})$ は，新規文書のキーワード系列と学習データを受け取って，新規文書のクラスの推定結果を返す決定関数である。(6)式は，決定関数が間違った推定結果を返す場合に1，正しい推定結果を返す場合に0になる0–1損失である。パラメータ θ と ψ という条件下における，損失の期待値であるリスクは次式で定義される。

$$R\left(d(y'^{n'},doc^{L}),\theta,\psi\right)=\sum_{(x,y^{n})^{L}\in(C,KEY^{n})^{L}}\sum_{(x',y'^{n'})\in(C,KEY^{n'})}p\left((x,y^{n})^{L}\mid\theta,\psi\right)p(x',y'^{n'}\mid\theta,\psi)L\left(d(y'^{n'},doc^{L}),x'\right) \quad \cdots\cdots(7)$$

(7)式のリスクは，パラメータ θ と ψ という条件下における誤り率に相当する。文書分類問題では，真のパラメータは未知なので，次式で定義されるベイズリスクを最小化する。

$$BR\left(p(\theta),p(\psi)\right)=\int_{\theta\in\Theta}\int_{\psi\in\Psi}p(\theta)p(\psi)R\left(d(y'^{n'},doc^{L}),\theta,\psi\right)d\psi d\theta \quad \cdots\cdots(8)$$

ただし，$p(\theta)$ と $p(\psi)$ はパラメータ θ と ψ の事前分布である。ベイズリスクは誤り率の期待値である。ベイズリスクを最小にする決定関数 $d_{B}(y'^{n'},doc^{L})$ が，ベイズ基準のもとで誤り率を最小にするという意味で最適な文書分類方法であり，次式で与えられる。

$$d_{B}(y'^{n'},doc^{L})=\arg\max_{\hat{x}'\in C}\int_{\theta\in\Theta}p(\theta\mid x^{L})p(\hat{x}'\mid\theta)d\theta\prod_{i=1}^{n'}\int_{\psi\in\Psi}p\left(\psi\mid(x,y^{n})^{L},y'^{i-1}\right)p(y'_{i}\mid\hat{x}',\psi)d\psi \quad \cdots\cdots(9)$$

ただし，

$$\int_{\theta\in\Theta}p(\theta\mid x^{L})p(\hat{x}'\mid\theta)d\theta=\frac{F(\hat{x}'\mid x^{L})+\beta(\hat{x}')}{\sum_{x''\in C}\left(F(x''\mid x^{L})+\beta(x'')\right)} \quad \cdots\cdots(10)$$

$$\int_{\psi\in\Psi}p\left(\psi\mid(x,y^{n})^{L},y'^{i-1}\right)p(y'_{i}\mid\hat{x}',\psi)d\psi=\frac{FF_{1}}{\sum_{y''\in KEY}FF_{2}} \quad \cdots\cdots(11)$$

$$FF_{1}=F\left((\hat{x}',y'_{i})\mid(x,y^{n})^{L}\right)+F(y'_{i}\mid y'^{i-1})+\gamma(y'_{i}\mid\hat{x}') \quad \cdots\cdots(12)$$

$$FF_{2}=F\left((\hat{x}',y'')\mid(x,y^{n})^{L}\right)+F(y''\mid y'^{i-1})+\gamma(y''\mid\hat{x}') \quad \cdots\cdots(13)$$

$\beta(\hat{x}')$ と $\gamma(y'_{i}\mid\hat{x}')$ は事前分布 $p(\theta)$ と $p(\psi)$ として使用するディリクレ分布のパラメータ，$F(\hat{x}'\mid x^{L})$ は学習データ中にあるクラス $\hat{x}'$ の文書数，$F\left((\hat{x}',y'_{i})\mid(x,y^{n})^{L}\right)$ は学習データ中にあるクラス $\hat{x}'$ の文書中にあるキーワード y'_{i} の数，$F(y'_{i}\mid y'^{i-1})$ は新規文書のキーワード系列 y'^{i-1} 中にあるキーワード y'_{i} の数である。なお，ディリクレ分布は多項分布に対する自然共役事前分布である。

実際に文書分類を行うためには，自然共役事前分布であるディリクレ分布のパラメータ $\beta(\hat{x}')$ と $\gamma(y'_{i}\mid\hat{x}')$ に何らかの数値を設定する必要がある。通常，学習データと新規文書以外に使用する文書データはない。パラメータ $\beta(\hat{x}')$ と $\gamma(y'_{i}\mid\hat{x}')$ の設定に関する考え方は数多くあるが，ジェフリーズの法則に従って無情報を示す0.5に設定されることが多い[11][12]。パラメータ $\beta(\hat{x}')$ と $\gamma(y'_{i}\mid\hat{x}')$ を，無情報を示す0.5に設定して実際に分類実験を行うと，学習データが少量の場合には分類精度は低くなってしまう。学習データのみ利用する場合の従来方法は，有限の学習データに対してベイズ基準のもとで誤り率を最小にできるが，学習データが少量の場合の分類精度は低い。

そこで，本研究では，新規文書や学習データとは性質が異なるが，ある程度似ているような既存の文書データを，事前分布の推定用データとして利用する。そして，事前分布の推定用データを利用して推定した事前分布を，ベイズ最適な従来方法の事前分布として使用することにより，新規文書や学習データとは性質が異なるが，ある程度似ている既存データ利用の有効性を検証する。

3.　事前分布の推定用データを利用する検証例

〈3･1〉　1番目の検証例　ここでは，本研究における1番目の検証例について説明する。

本研究では，事前分布 $p(\theta)$ と $p(\psi)$ を推定するための推定用データを利用する。通常，新規文書と学習データは，真のパラメータが同一の多項分布に従って生起するという点で，同一の性質を持つと仮定される。本研究では，新規文書と比較して，性質が学習データのように同一ではないが，ある程度似ているような既存の文書データを事前分布の推定用データとして利用する。事前分布の推定用データ doc^{G} は次式で示される。

$$doc^{G}=(v,w^{m})^{G}=(v_{1},w^{m_{1}})(v_{2},w^{m_{2}})\cdots(v_{G},w^{m_{G}}) \quad \cdots\cdots(14)$$

ただし，G は推定用データ中にある文書数，v_{i} は推定用データ中にある i 番目の文書のクラス，m_{i} は i 番目の文書中にあるキーワードの数，$w^{m_{i}}$ は i 番目の文書中にあるキーワードの系列，

$$w^{m_{i}}=w_{i,1}w_{i,2}\cdots w_{i,m_{i}} \quad \cdots\cdots(15)$$

$w_{i,j}$ はキーワード系列 $w^{m_{i}}$ 中にある j 番目のキーワードで

ある。推定用データ doc^G が生起する確率は次式で計算できる。

$$\begin{aligned} p(doc^G \mid \theta,\psi) &= p\left((v,w^m)^G \mid \theta,\psi\right) \\ &= p\left((v_1,w^{m_1})(v_2,w^{m_2})\cdots(v_G,w^{m_G}) \mid \theta,\psi\right) \\ &= \prod_{i=1}^{G}\left(p(v_i \mid \theta)\prod_{j=1}^{m_i} p(w_{i,j} \mid v_i,\psi)\right) \end{aligned} \quad \cdots\cdots(16)$$

1 番目の検証例においても，新規文書のクラスの推定には従来方法と同様に(９)式を用いる。ただし，(10)式，(12)式および(13)式中のパラメータ $\beta(\hat{x}')$ と $\gamma(y_i' \mid \hat{x}')$ はジェフリーズの法則に従って設定するのではなく，(17)式および(18)式によって設定する。

$$\beta(\hat{x}') = \frac{F(\hat{x}' \mid v^G) + \beta'(\hat{x}')}{\sum_{x'' \in C}\left(F(x'' \mid v^G) + \beta'(x'')\right)} \quad \cdots\cdots(17)$$

$$\gamma(y_i' \mid \hat{x}') = \frac{F\left((\hat{x}', y_i') \mid (v,w^m)^G\right) + \gamma'(y_i' \mid \hat{x}')}{\sum_{y'' \in KEY}\left(F\left((\hat{x}', y'') \mid (v,w^m)^G\right) + \gamma'(y'' \mid \hat{x}')\right)} \quad \cdots\cdots(18)$$

ただし，$\beta'(\hat{x}')$ と $\gamma'(y_i' \mid \hat{x}')$ は事前分布 $p(\theta)$ と $p(\psi)$ として使用するディリクレ分布のパラメータ，$F(\hat{x}' \mid v^G)$ は推定用データ中にあるクラス $\hat{x}'$ の文書数，$F\left((\hat{x}', y_i') \mid (v,w^m)^G\right)$ は推定用データ中にあるクラス $\hat{x}'$ の文書中にあるキーワード y_i' の数である。(17)式および(18)式を利用することは，事前分布の推定用データを利用して推定した未知パラメータの推定値（事前分布の推定用データに対応する未知パラメータの推定値）を新規文書および学習データに対応する未知パラメータの事前分布として利用することである。

事前分布の推定用データと新規文書および学習データでは，文書が生起する確率分布の未知パラメータが異なる。よって，学習データが少量の場合の分類精度は，事前分布の推定用データに対応する未知パラメータと新規文書および学習データに対応する未知パラメータがどれだけ似ているかに依存する。仮に，新規文書および学習データの未知パラメータと比較して，無情報の事前分布によるパラメータの推定値よりも似ていない（遠くはなれた）未知パラメータに対応するような既存文書データを事前分布の推定用データとして利用した場合には，無情報の事前分布による分類精度よりも低くなることが想定される。よって，事前分布の推定用データの選択が重要になるが，ある程度，新規文書および学習データに似たデータを利用すれば，無情報の事前分布を使用する場合よりも分類精度が高くなることが期待される。

学習データのみ利用する場合の従来方法と，本研究における 1 番目の検証例との比較実験の結果を Fig.1 から Fig.4 に示す。学習データのみ利用する場合の従来方法のパラメータ $\beta(\hat{x}')$ と $\gamma(y_i' \mid \hat{x}')$，および 1 番目の検証例のパラメータ $\beta'(\hat{x}')$ と $\gamma'(y_i' \mid \hat{x}')$ はジェフリーズの法則に従って無情報を示す 0.5 に設定した。実際の文書データとして新聞記事のデータを用いている。新規文書（10000 記事）と学習データには，

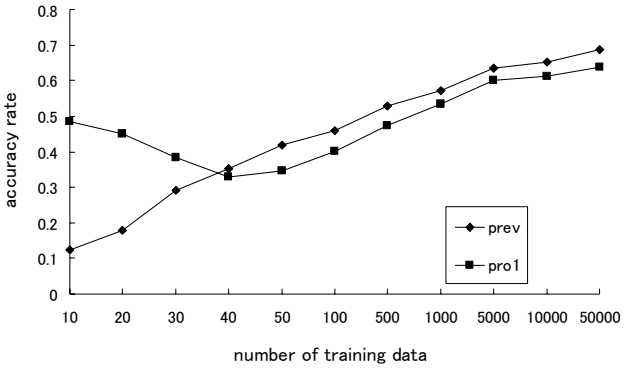

Fig. 1. Experiments results 1 on the first verification.

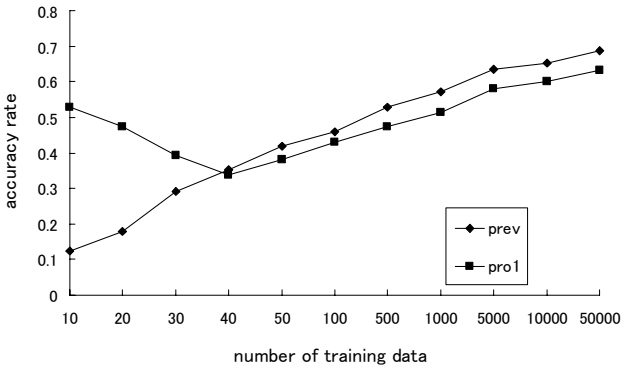

Fig. 2. Experiments results 2 on the first verification.

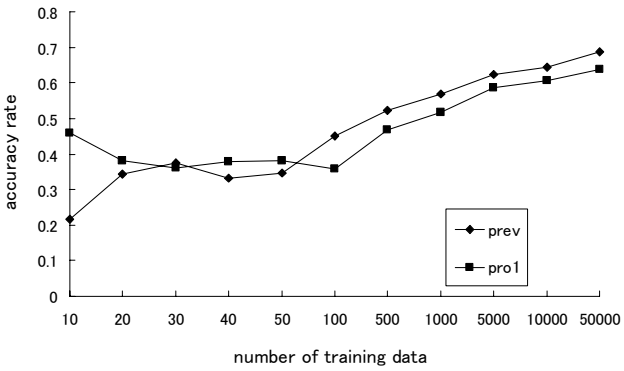

Fig. 3. Experiments results 3 on the first verification.

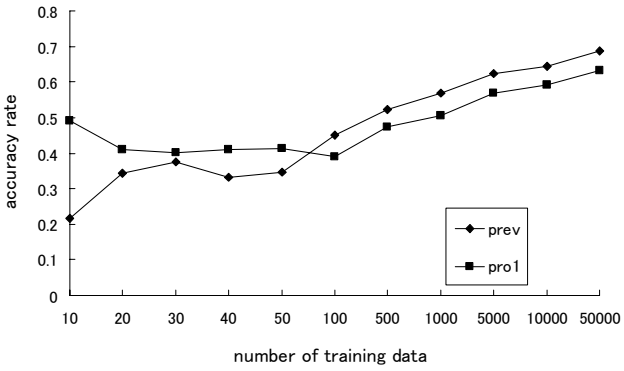

Fig. 4. Experiments results 4 on the first verification.

Fig.1 と Fig.2 の比較実験では 2007 年版の毎日新聞記事[13]，Fig.3 と Fig.4 の比較実験では 2008 年版の毎日新聞記事[14]を

用いた。事前分布の推定用データ（50000 記事）には Fig.1 と Fig.3 の比較実験では 1994 年版の毎日新聞記事[15]，Fig.2 と Fig.4 の比較実験では 2001 年版の毎日新聞記事[16]を用いた。今回用いた新聞記事データ集では記事ごとに見出しからキーワードが抽出されており，比較実験ではこれらの見出しキーワードをキーワードとして利用した。また，データ集の記事は国際，経済，科学等の 16 個のクラスに分類されており，比較実験ではデータ集の分類体系を利用した。各 Fig 中の prev が学習データのみ利用する場合の従来方法の正解率（accuracy rate），pro1 が 1 番目の検証例の正解率を示す。正解率は新規文書数に対する，クラスの推定に成功した文書数の割合で算出している。

なお，本研究では学習データや新規文書と性質は異なるがある程度似ているデータを事前分布の推定用データとして用いているが，事前分布の推定用データに関する厳密な定義は行っていない。事前分布の推定用データの選択は分類対象ごとに検討が必要になると考えられる。今回の比較実験では，記事の年代が異なることによって使用されるキーワード集合が変化しているが，新聞記事という性質上ある程度の数のキーワードは年代が異なっても同じものが生起すると考えて，年代の古い記事を事前分布の推定用データとして用いることにした。

各データ集のキーワード集合の大きさは，1994 年版が 171729，2001 年版が 222744，2007 年版が 61606，2008 年版が 59046 である。Fig.1 の比較実験での共通のキーワード数は 24564，Fig.2 の比較実験での共通のキーワード数は 27947，Fig.3 の比較実験での共通のキーワード数は 24147，Fig.4 の比較実験での共通のキーワード数は 27336 である。発行年が遠い 1994 年版よりも，発行年が近い 2001 年版との共通のキーワード数が大きいことから，年代の近い新聞記事ほど性質が似ていることが確認できる。2007 年版と 2008 年版が本社版のデータなのに対して，1994 年版と 2001 年版は本社版の記事と地方版の記事が混在したデータ集になっているため，本社版の 2007 年版と 2008 年版のキーワード集合は 1994 年版と 2001 年版よりも小さい。これは，地方版には小さな街の地名など地方版特有のキーワードが数多く存在するためだと考えられる。

学習データ（training data）が少量の場合には，1 番目の検証例の正解率が，学習データのみ利用する場合の従来方法の正解率よりも高い。また，Fig.1 と Fig.2 および Fig.3 と Fig.4 の学習データ数が 10 のときの正解率を比較すると，事前分布の推定用データの年代が学習データおよび新規文書に近い方が正解率が高くなっている。年代が近い新聞記事の方が共通のキーワード数が大きいことと合わせて解釈すると，年代が近い新聞記事の方が性質が似ており，性質が似ているほど，学習データが少量時の分類精度が高くなることが確認できる。

しかし，学習データが増加すると，1 番目の検証例の正解率は，従来方法の正解率よりも多少低くなっている。学習データが少量の場合の分類精度を向上させるという当初の目的は達成できたが，学習データ増加時の分類精度が従来方法よりも低くなってしまうのは大きな短所である。

この短所の原因は過学習だと考えられる。事前分布を導入する場合，学習データの増加にともなって事前分布の影響が消えるのが普通である。しかし，今回のデータでは各記事から観測されるキーワードは見出しのキーワードのため，1 記事あたり約 8 キーワードとあまり多くないのに対して，キーワード集合は大きく，疎なデータである。よって，学習データとして用いている 50000 記事（学習データ最大時）では事前分布の推定用データによる過学習の影響を打ち消すには不十分だと考えられる。仮に 100 万記事，1000 万記事と学習データを増加することが可能であれば，学習データの増加にともなって過学習の影響を打ち消すことも可能だと考えられるが，それだけの学習データを確保することは難しい。1 番目の検証例の短所への対処については，次の 2 番目の検証例で説明する。

〈3･2〉　2 番目の検証例　ここでは，本研究における 2 番目の検証例について説明する。

学習データが少量の場合には，1 番目の検証例の分類精度は学習データのみ利用する場合の従来方法の分類精度よりも高い。しかし，学習データが増加した場合にも事前分布の推定用データの影響が残ってしまうために，1 番目の検証例の分類精度は学習データ増加時には従来方法よりも低くなってしまう。これは過学習が起きているためである。

事前分布の推定用データによる過学習が起きても，学習データを膨大に増加することが出来れば学習データの増加にともなって事前分布の推定用データによる影響を打ち消すことが出来る。しかし，データが疎な今回の新聞記事データの場合には必要な学習データ量が非常に膨大で非現実的である。そこで，1 番目の検証例に，学習データが少量の場合には主に推定用データを用いて新規文書のクラスを推定し，学習データ増加時には主に学習データを用いて新規文書のクラスを推定するような仕組みを追加して，2 番目の検証例とする。追加する仕組みは具体的にはパラメータ $\beta(\hat{x}')$ と $\gamma(y_i'|\hat{x}')$ の設定についてである。追加する仕組みは理論的に導出されたものではなく，経験則から考え出したものである。2 番目の検証例では，1 番目の検証例における(17)式および(18)式を，(19)式および(20)式に変更する。

$$\beta(\hat{x}')=\left(\frac{F(\hat{x}'|v^G)+\beta'(\hat{x}')}{\sum_{x''\in C}\left(F(x''|v^G)+\beta'(x'')\right)}\right)^{\log_{A_1}L+2}+\frac{\beta'(\hat{x}')}{(A_2)^{A_3^L}} \quad \cdots\cdots(19)$$

$$\gamma(y_i'|\hat{x}')=\left(\frac{FF_3}{\sum_{y''\in KEY}FF_4}\right)^{\log_{A_1}L+2}+\frac{\gamma'(y_i'|\hat{x}')}{(A_2)^{A_3^L}} \quad \cdots\cdots(20)$$

ただし，

$$FF_3=F\left((\hat{x}',y_i')|(v,w^m)^G\right)+\gamma'(y_i'|\hat{x}') \quad \cdots\cdots(21)$$

$$FF_4 = F\left((\hat{x}', y'') \mid (v, w^m)^G\right) + \gamma'(y'' \mid \hat{x}') \quad \cdots\cdots\cdots\cdots\cdots\cdots (22)$$

A_1，$1 < A_1 < \infty$，A_2，$1 < A_2 < \infty$，A_3，$0 < A_3 < 1$ は経験則で設定する定数である。(19)式および(20)式の右辺の第 1 項が事前分布の推定用データに対応する部分で学習データが少量時に強調したい項である。第 2 項が学習データ増加時に強調したい（無情報を示すように設定される）パラメータに対応する部分である。

学習データが少量の場合には，第 2 項の分母を大きな値にすることによって第 2 項の値をほぼ 0 にして，第 1 項を強調する。また，学習データの増加にともなって，第 1 項の指数部の値を大きくすることによって第 1 項の値を 0 に近づけ，かつ第 2 項の分母を 1 に近づけることによって第 2 項の分子のパラメータを強調する。第 1 項と第 2 項の形にもいろいろなものが考えられるが，試行錯誤の結果，(19)式および(20)式の形を採用した。A_1 によって事前分布の推定用データの影響の打ち消し方を調整する。A_1 を大きく設定するほど，事前分布の推定用データの影響が，学習データ量が増加した後にも残る設定になる。A_2 は学習データ少量時に第 2 項の分母を大きくするために，大きな値に設定するパラメータである。A_2 を小さく設定すると，学習データの増加に対してすぐに第 2 項を強調することになる。A_3 は学習データの増加にともなって A_2 の指数部を 0 に近づけるために $0 < A_3 < 1$ の範囲で設定する。A_3 を小さく設定すると，学習データの増加に対してすぐに第 2 項を強調することになる。(19)式および(20)式を用いることにより，学習データが少量の場合には主に推定用データを用いて新規文書のクラスを推定し，学習データ増加時には主に学習データを用いて新規文書のクラスを推定することができる。

4. 比較実験

学習データのみ利用する場合の従来方法と，本研究における 2 番目の検証例との分類精度の比較実験結果を Fig.5 から Fig.8 に示す。比較実験では，学習データのみ利用する場合の従来方法のパラメータ $\beta(\hat{x}')$ と $\gamma(y_i' \mid \hat{x}')$，および本研究における 2 番目の検証例のパラメータ $\beta'(\hat{x}')$ と $\gamma'(y_i' \mid \hat{x}')$ はジェフリーズの法則に従って無情報を示す 0.5 に設定した。本研究における 2 番目の検証例の A_1，A_2，A_3 は，試行錯誤の結果，それぞれ 10，1000000，0.995 に設定した。新規文書（10000 記事）と学習データには，Fig.5 と Fig.6 の比較実験では 2007 年版の毎日新聞記事[13]，Fig.7 と Fig.8 の比較実験では 2008 年版の毎日新聞記事[14]を用いた。事前分布の推定用データ（50000 記事）には Fig.5 と Fig.7 の比較実験では 1994 年版の毎日新聞記事[15]，Fig.6 と Fig.8 の比較実験では 2001 年版の毎日新聞記事[16]を用いた。各 Fig 中の prev が学習データのみ利用する場合の従来方法の正解率（accuracy rate），pro2 が 2 番目の検証例の正解率を示す。

Fig.5 から Fig.8 より，学習データが少量の場合には，本研究における 2 番目の検証例の正解率が，学習データのみ利用する場合の従来方法の正解率よりも高いことがわかる。

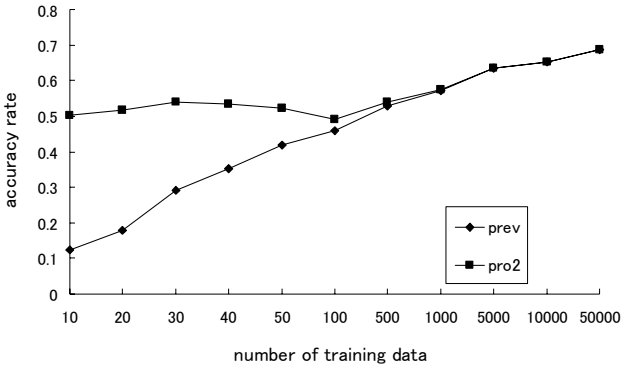

Fig. 5. Experiments results 5 on the second verification.

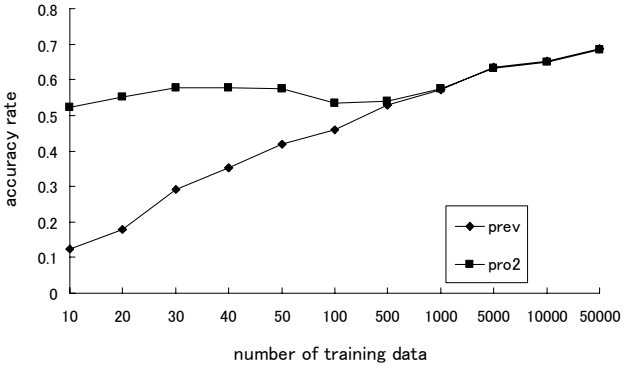

Fig. 6. Experiments results 6 on the second verification.

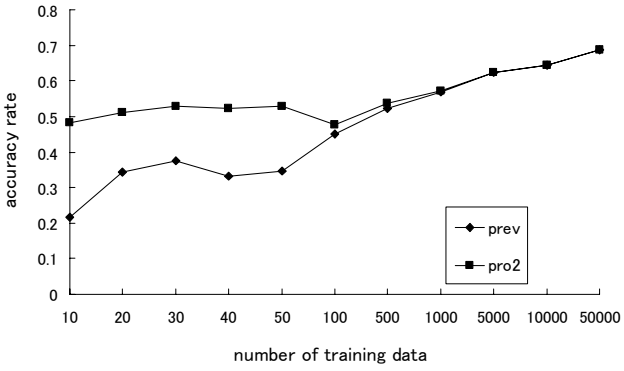

Fig. 7. Experiments results 7 on the second verification.

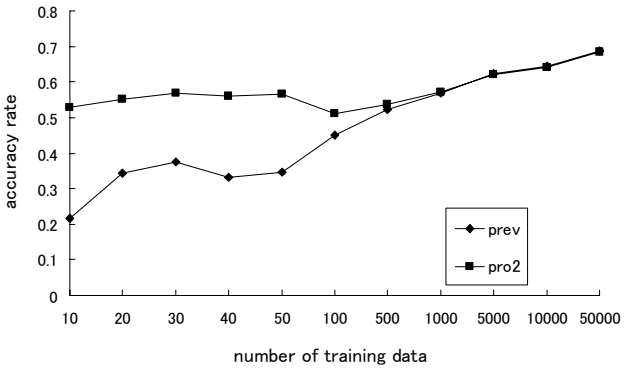

Fig. 8. Experiments results 8 on the second verification.

また，学習データ増加時には，両方の正解率が同程度であることもわかる。これは，A_1，A_2，A_3 を 10，1000000，0.995

に設定した結果，学習データが少量の場合には(19)式および(20)式の第 1 項（事前分布の推定用データ）が適切に強調され，学習データ増加時には(19)式および(20)式の第 2 項（無情報の事前分布のパラメータ）が適切に強調されたからである。(19)式および(20)式は経験則であり，A_1，A_2，A_3 の値の設定も 1 例に過ぎないが，比較実験結果より，学習データや新規文書と性質は異なるがある程度似ている既存文書データを適切に利用すれば，学習データ少量時の分類精度の向上に有効に作用することが確認できた。また，不適切に利用すると学習データ増加時に分類精度を低下させることもあるが，適切に利用すれば従来方法と同程度の分類精度を達成できることも確認できた。

5. ま と め

従来研究において，有限の学習データに対して，ベイズ基準のもとで誤り率を最小にするという点で最適な文書分類方法（ベイズ最適な従来方法）が提案されている。しかし，学習データが少量の場合には，従来方法の分類精度は低くなってしまう。

そこで，本研究では，新規文書や学習データとある程度似た性質の既存データを学習データが少量しかない場合の分類精度の向上に利用し，その有効性を検証した。利用の仕方はいろいろ考えられるが，本研究では事前分布の推定用データとして利用し，限られたデータによる数例の実験ではあるが，学習データが少量しかない場合に事前分布の推定用データを利用することによって，分類精度が向上することを 1 番目の検証例で確認した。また，学習データが少量の場合には主に事前分布の推定用データを用いて新規文書のクラスを推定し，学習データ増加時には主に学習データを用いて新規文書のクラスを推定するような仕組みを経験則から考え出し，限られたデータによる数例の実験ではあるが，学習データが少量の場合には分類精度が従来方法よりも高く，学習データ増加時には両方の分類精度が同程度になることを 2 番目の検証例で確認した。

本研究では，学習データが不足している場合の対応策として，新規文書や学習データとある程度似た性質の既存データの利用について，事前分布の推定用データとして利用することによって，その有効性を確認した。しかし，文書の生起を仮定している確率モデルについて考えてみると，今回利用した既存データは学習データや新規文書とは異なる未知パラメータが支配する確率分布から生起したデータである。よって，この既存データを含めた文書分類問題を本質的に考えるならば，情報通信における通信路のギルバートモデル[17]や，経営工学における需要予測などの季節変動モデル[18]のように，パラメータが時間経過にともなって変化するモデルを文書分類にも導入する必要があると考えられる。自然言語処理における形態素解析などでは単語長（単語を構成する文字列の長さ）を示す確率モデル[19]も検討されており，文書分類においても文書中のキーワード数を示す確率モデルを導入することも考えられる。これらのモデルの拡張については今後の課題としたい。

また，本研究では新規文書や学習データとある程度似た性質の既存データを利用しているが，データの類似性に関する尺度などは未定義である。本研究の比較実験では発行年ごとに新聞記事データを分けて扱ったが，本来はデータの類似性の尺度を定義したもとで，新聞記事データのクラスタリングを行うことによって，類似データの整備を行う必要がある。尺度としてはいろいろな候補が考えられるが，例えば 2 つの確率分布間で定義されるカルバックライブラー情報量[20]などが考えられる。クラスタリング手法についても既存の各種手法が候補になる。他方，分類誤り率を最小化あるいは近似的に最小化することを目的にデータをクラスタリングすることも考えられる。この場合には，文書分類と文書分類の前段のクラスタリング部分をまとめて 1 つの最適化問題として定式化することにより，文書分類向けのクラスタリング手法を導出することが可能だと考える。データの類似性およびクラスタリングに関する具体的な検討は今後の課題としたい。

文　献

(1) 麻生英樹・津田宏治・村田　昇：パターン認識と学習の統計学，岩波書店，東京 (2003)
(2) Y. Yang : “An Evaluation of Statistical Approaches to Text Categorization”, Journal of Information Retrieval, Vol.1, No.1, pp.67-88 (1999)
(3) T. Joachims : “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization”, International Conference on Machine Learning ’97, pp.143-151 (1997)
(4) G. Salton and C. Buckley : “Term-Weighting Approaches in Automatic Text Retrieval”, Information Processing & Management, Vol.24, No.5, pp.513-523 (1988)
(5) I. Witten, A. Moffat, and T. Bell : Managing Gigabytes, Academic Press, California (1999)
(6) 徳永健伸：情報検索と言語処理，東京大学出版会，東京 (1999)
(7) M. Iwayama and T. Tokunaga : “A Probabilistic Model for Text Categorization based on a Single Random Variable with Multiple Values”, Conference on Applied Natural Language Processing, pp.162-167 (1994)
(8) C. D. Manning and H. Schutze : Foundations of Statistical Natural Language Processing, MIT Press (1999)
(9) Y. Maeda and H. Ohara : “A Note on Telegram Categorization Algorithm Based upon Statistical Decision Theory”, IPSJ Journal, Vol.43, No.10, pp.3119-3126 (2002) (in Japanese)
前田康成・小原　永：「統計的決定理論に基づく電報分類方法に関する一考察」，情処学論，Vol.43, No.10, pp.3119-3126 (2002)
(10) J. Berger : Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York (1980)
(11) 鈴木　譲：ベイジアンネットワーク入門，培風館，東京 (2009)
(12) T. Matsushima and S. Hirasawa : “A Bayes coding algorithm for Markov models”, TECHNICAL REPORT OF IEICE, IT95-1, pp.1-6 (1995)
(13) 毎日新聞社 : CD 毎日新聞 2007 データ集本社版，日外アソシエーツ，東京 (2007)
(14) 毎日新聞社 : CD 毎日新聞 2008 データ集本社版，日外アソシエーツ，東京 (2008)
(15) 毎日新聞社 : CD 毎日新聞 94 データ集，日外アソシエーツ，東京 (1994)
(16) 毎日新聞社 : CD 毎日新聞 2001 データ集，日外アソシエーツ，東京 (2001)
(17) 今井秀樹：情報理論，昭晃堂，東京 (1984)
(18) 廣松　毅・浪花貞夫：経済時系列分析，朝倉書店，東京 (1990)

(19) M. Nagata : “A Japanese Morphological Analysis Method Using a Statistical Language Model and an N-best Search Algorithm”, IPSJ Journal, Vol.40, No.9, pp.3420-3431 (1999) (in Japanese)
永田昌明：「統計的言語モデルと N-best 探索を用いた日本語形態素解析法」，情処学論，Vol.40, No.9, pp.3420-3431 (1999)
(20) 平澤茂一：情報理論，培風館，東京 (1996)

前 田 康 成

（非会員） 1995 年早大・理工・工業経営卒。1997 年同大大学院・修士課程了。同年，日本電信電話（株）入社。現在，北見工大・准教授。博士（工学）。統計的決定理論の学習問題への応用に関する研究に従事。電子情報通信学会，情報処理学会等各会員。

吉 田 秀 樹

（非会員） 1991 年九大・工・電子卒。現在，鹿児島大・教授。博士（工学）。医用生体工学の研究に従事。BMFSA 等各会員。

鈴 木 正 清

（非会員） 1982 年北大・工・電子卒。1987 年同大大学院博士課程了。同年同大応用電気研究所助手。1993 年北見工大・助教授，1996 年北大・電子研助教授。センサアレー信号処理，鮭追跡システムの開発，国際会議運営支援システムの開発，電子波包絡の回路モデルの研究に従事。2001 年より北見工大教授。工博。

松 嶋 敏 泰

（非会員） 1978 年早大・理工・工業経営卒。1980 年同大大学院修士課程了。同年，日本電気（株）入社。1986 年早大・理工学研究科・博士後期課程入学。現在，早大・応用数理学科教授。知識情報処理及び情報理論とその応用に関する研究に従事。工学博士。IEEE 等各会員。