統計的決定理論に基づく電報分類方法に関する一考察

前田康成[†,☆] 小原永[†]

従来，電報の分類作業は人手で行われており，その自動化が望まれている．電報分類問題については文書分類問題と同様の定式化を行うことができる．しかし，従来の文書分類方法には，学習用データが有限の場合にはその分類精度に何ら理論的な保証がない．そこで，本研究では統計的決定理論に基づいて，学習用データが有限の場合に電報を間違った分野に分類してしまう確率である誤り率をベイズ基準のもとで最小にすることが保証された電報分類方法を提案する．さらに，実際の分類実験を通して情報検索や文書分類で利用されているベクトル空間法を電報分類に適用した場合の基本的なアルゴリズムと提案方法との比較を行い，提案方法の方がベクトル空間法を電報分類に適用した場合のアルゴリズムよりも分類精度が高いことを示す．また，同じく文書分類等で利用されている Naive-Bayes 法を電報分類に適用した場合のアルゴリズムと提案方法との比較を，そのアルゴリズムの導出過程の違いから考察する．

A Note on Telegram Categorization Algorithm Based upon Statistical Decision Theory

Yasunari Maeda[†,☆] and Hisashi Ohara[†]

Usually telegram categorization is done by human. And an automatic method for telegram categorization is needed. Telegram categorization problem can be resolved by using the same mathematical model of text categorization. But algorithms in previous research on text categorization have no theoretical guarantee. So in this research we propose a new algorithm for telegram categorization, which is based upon statistical decision theory and minimizes an error rate with respect to Bayes criterion. And we show an effectiveness of our proposed algorithm by some simulations.

1. はじめに

従来，電報の分類作業は，実際に電報文を打つオペレータによって行われている．しかし，この作業にかかる時間によってオペレータの作業効率が低下するという問題点が指摘されている．そこで，本研究では電報の自動分類方法を研究対象とする．

電報の自動分類方法は従来からルールベースによる方法[10]が研究されているが，ルールの作成が難しい等の問題点をかかえていた．本研究では，電報の生起するモデルに，従来から文書分類の分野で多くの研究者によって採用されている確率モデル[3),4),11)]を採用し，電報分類問題を文書分類問題と同様に定式化する．

しかし，文書分類の従来研究では，分類精度を理論的に保証することができないか，あるいは学習用の電報数が無限の場合には最適性が保証できても有限の場合には分類精度を理論的に保証することができない分類方法が提案されている．そこで，本研究では統計的決定理論[1),9)]に基づき，ベイズ基準のもとで学習用の電報数が有限の場合でも電報を誤った分野に分類してしまう確率である誤り率を最小にすることが保証された分類方法を提案する．また，従来のベクトル空間法[8),12)]を電報分類問題に適用した基本的な分類方法に対して提案方法の方が誤り率が低いことを，一例にすぎないが分類実験を通して検証し，さらに従来の最尤推定法を採用した Naive-Bayes 法[4),6)]を電報分類問題に適用した分類方法と提案方法の関係をアルゴリズムの導出過程の違いから考察する．

まず，2 章で電報分類問題の概要について述べ，3 章で文書分類の従来研究について述べる．4 章で統計的決定理論に基づく電報分類問題について述べ，定式化を行う．5 章で実際にベイズ基準のもとで誤り率を

† NTT サイバースペース研究所メディア処理プロジェクト
Media Processing Project, NTT Cyber Space Laboratories
☆ 現在，NTT 東日本研究開発センタ
Presently with Research and Development Center, NTT EAST

最小にする電報分類方法を提案する．6 章で一例にすぎないが実際の電報データを用いた電報分類実験を行い，提案方法が実問題において基本的なベクトル空間法を電報分類問題に適用した方法よりもより多くの電報を正しい分野に分類できることを示す．さらに，最尤推定法を採用した Naive-Bayes 法を電報分類問題に適用した従来方法と本研究における提案方法の関係について考察する．最後に，7 章でまとめを行う．

2. 電報分類問題の概要

電報分類問題について述べる前に，まず，いくつかの定義を行う．c_i は電報が内容に応じて分類される，「結婚」,「クリスマス」等の分野を示し，$c_i \in C$ で，C は電報の分野の集合，$C = \{c_1, c_2, \ldots, c_{|C|}\}$ である．なお，$|\cdot|$ は集合の要素数を示す．前提として，1 つの電報は必ず 1 つの分野のみに分類されるものとする．key_i は電報の電報文中に出現するキーワードを示し，$key_i \in KEY$ で，KEY はキーワードの集合，$KEY = \{key_1, key_2, \ldots, key_{|KEY|}\}$ である．

電報分類問題とは，すでに各分野 c_i に分類されている学習用の電報 doc_{l_i} を用いて学習し，新規に分類したい未知の電報 doc_u をいずれかの分野 c_i に分類する問題である．

学習用の電報 doc_{l_i} は，doc_{l_i} が分類されている分野 w_i，$w_i \in C$ と doc_{l_i} の電報文に含まれるキーワードの系列 z^{N_i} の 2 項組み (w_i, z^{N_i}) で表現される．なお，N_i は doc_{l_i} に含まれるキーワードの延べ数を示し，z^{N_i} は $z_{i,1}z_{i,2}\cdots z_{i,N_i}$ と同一で，$z_{i,j}$，$z_{i,j} \in KEY$ は doc_{l_i} に含まれるキーワードの系列中の j 番目に並んでいるキーワードを示す．学習用の電報 doc_{l_i} は G 個与えられ，G 個の学習用の電報全体は，長さ G の電報の系列 $doc_{l_i}^G$ として表現される．また，$doc_{l_i}^G$ は $doc_{l_1}doc_{l_2}\cdots doc_{l_G}$ や $(w_i, z^{N_i})^G$ や $(w, z^N)^G$ のように表現されることもある．

新規に分類したい未知の電報 doc_u は，doc_u が分類されるべき真の分野 c^*，$c^* \in C$ と doc_u の電報文に含まれるキーワードの系列 y^M の 2 項組み (c^*, y^M) で表現される．なお，M は未知の電報 doc_u に含まれるキーワードの延べ数を示し，y^M は $y_1y_2\cdots y_M$ と同一である．また，実際に観測されるのはキーワードの系列 y^M のみで，真の分野 c^* は未知である．

すなわち，電報分類問題とは，新規に分類したい未知の電報 doc_u のキーワード系列 y^M を観測したもとで，doc_u が分類されるべき真の分野 c^* を推定する問題として解釈できる．さらに，電報分類問題とは数学的には，従来から研究されている文書分類問題と同様の問題である．

電報分類問題に対しては，従来からキーワードの有無を基本とするルールベースによる研究[10]等が行われているが，ルールの作成が困難である等の問題点をかかえている．そこで，本研究では，従来から文書分類問題に対して多くの研究者によって採用されている確率モデルによる定式化を行う．しかし，従来の文書分類方法は分類精度を理論的に保証することができないか，あるいは学習用の電報数が無限の場合には最適性が保証できても有限の場合には分類精度を理論的に保証することができない．そこで，本研究では統計的決定理論に基づき，学習用の電報数が有限の場合に未知の電報を間違った分野に分類してしまう確率である誤り率をベイズ基準のもとで最小化するという意味で，理論的に最適な電報分類方法を提案する．

3. 従来の文書分類方法

3.1 ベクトル空間法に基づく電報分類方法

情報検索の分野で広く用いられているベクトル空間法を文書分類に適用した文書分類方法がいくつか提案されており[4),8),11),12)]，文書分類に適した様々な提案が盛り込まれたアルゴリズムが提案されている．しかし，本論文ではベクトル空間法の基本的な性質に着目したいので，ここでは，我々が考えた最も単純な TF・IDF 法を加味した基本的なベクトル空間法に基づく文書分類方法を電報分類問題に適用した場合を紹介する．なお，以下では単にベクトル空間法に基づく電報分類方法と呼ぶ．

ベクトル空間法に基づく電報分類方法では，次式によって分類する分野が決定される．なお，$d_{Vec}(y^M)$ は未知の電報 doc_u のキーワード系列 y^M を引数にとり，ベクトル空間法に基づいて未知の電報を分類する分野を決定する決定関数である．

$$\begin{aligned} d_{Vec}(y^M) &= \arg\max_{x \in C} \cos(V(x), V(doc_u)) \\ &= \arg\max_{x \in C} \frac{V(x) \cdot V(doc_u)}{||V(x)||||V(doc_u)||}, \end{aligned} \quad (1)$$

ただし，

$$\begin{aligned} V(x) = \Bigg(& F((x, key_1)|(w, z^N)^G) \log \frac{|C|}{A(key_1)}, \\ & F((x, key_2)|(w, z^N)^G) \log \frac{|C|}{A(key_2)}, \\ & \ldots, F((x, key_{|KEY|})|(w, z^N)^G) \\ & \log \frac{|C|}{A(key_{|KEY|})} \Bigg), \end{aligned} \quad (2)$$

$$V(doc_u) = \Bigg(F(key_1|y^M)\log\frac{|C|}{A(key_1)}, F(key_2|y^M)\log\frac{|C|}{A(key_2)}, \ldots, F(key_{|KEY|}|y^M)\log\frac{|C|}{A(key_{|KEY|})}\Bigg), \tag{3}$$

$V(x)$ は分野 x，$x \in C$ の特徴ベクトルで，$F((x, key_i)|(w, z^N)^G)$ は学習用の電報全体中で分野 x に分類されている電報でキーワード key_i が生起した回数，$A(key_i)$ は $F((x, key_i)|(w, z^N)^G) > 0$ が成立している分野の数，$V(doc_u)$ は新規に分類したい未知の電報 doc_u の特徴ベクトルで，$F(key_i|y^M)$ は未知の電報 doc_u のキーワード系列 y^M 中でキーワード key_i が生起した回数，cos はベクトル間の余弦の値を求める関数，$V(x) \cdot V(doc_u)$ はベクトル $V(x)$，$V(doc_u)$ 間の内積，$||V(x)||$ はベクトル $V(x)$ のノルムを示す．

上記の決定関数から理解できるように，ベクトル空間法に基づく電報分類方法ではキーワードの生起の仕方の相関が最大となる分野に未知の電報 doc_u を分類していると解釈できる．定性的には相関が大きければ関係深い分野に分類できそうであるが，相関が最大となる分野に分類しても，誤り率を理論的に何らかの基準に基づいて最小化することは保証できない．

3.2 最尤推定法を採用した Naive-Bayes 法に基づく電報分類方法

文書分類の分野において，確率モデルを採用した様々な分類方法が提案されている[3),4),11)]が，本研究ではその中でも最も多くのその他の分野においても適用されている方法の 1 つである最尤推定法を採用した Naive-Bayes 法[5),6)]を（以下では，単に Naive-Bayes 法と呼ぶ）を取り上げる．Naive-Bayes 法に基づく文書分類方法を電報分類に適用した場合を紹介する前に，いくつかの定義を行う．

$p(c_i|\theta)$ は分野 c_i が生起する確率分布，$p(key_j|c_i, \theta)$ は分野 c_i が生起した条件のもとでキーワード key_j が生起する確率分布を示し，$p(c_i|\theta)$ も $p(key_j|c_i, \theta)$ もともに連続パラメータ $\theta, \theta \in \Theta$ によって支配されていて既知である．$\theta^*, \theta^* \in \Theta$ は真のパラメータで未知である．

上記の確率モデルのもとでは，電報の生成とは，まず確率分布 $p(c_i|\theta)$ に従って分野 c_i が生起し，次にいくつかのキーワードが独立に確率分布 $p(key_j|c_i, \theta)$ に従って生起することに相当する．パラメータ θ のもとでの学習用の電報 doc_{l_i} の生起確率 $p(doc_{l_i}|\theta)$ は次式で示される．

$$\begin{aligned} p(doc_{l_i}|\theta) &= p(w_i|\theta)p(z^{N_i}|w_i, \theta) \\ &= p(w_i|\theta)\prod_{j=1}^{N_i} p(z_{i,j}|w_i, \theta). \end{aligned} \tag{4}$$

現実の世界では各キーワードは $p(key_k|key_j, c_i, \theta)$ という確率分布で示されるような，1 つ前に生起したキーワードによって次に生起するキーワードの確率分布が左右されるマルコフ性等を有するかもしれないが，上式では各キーワードが他のキーワードとは独立に生起すると仮定されている．現実世界のこのようなモデル化の仕方が Naive-Bayes 法の考え方である．本論文でも全体を通して，このモデル化を採用している．

同様に未知の電報 doc_u の生起確率 $p(doc_u|\theta)$ は次式で示される．

$$\begin{aligned} p(doc_u|\theta) &= p(c^*|\theta)p(y^M|c^*, \theta) \\ &= p(c^*|\theta)\prod_{i=1}^{M} p(y_i|c^*, \theta). \end{aligned} \tag{5}$$

次に，Naive-Bayes 法に基づく電報分類方法では未知の電報を分類する分野が次式によって決定される．なお，$d_{NB}(y^M)$ は未知の電報 doc_u のキーワード系列 y^M を引数にとり，Naive-Bayes 法に基づいて未知の電報を分類する分野を決定する決定関数である．

$$d_{NB}(y^M) = \arg\max_{x \in C} \hat{p}(x)\prod_{i=1}^{M} \hat{p}(y_i|x), \tag{6}$$

ただし，$\hat{p}(x)$ および $\hat{p}(y_i|x)$ はそれぞれ $p(x|\theta^*)$ および $p(y_i|x, \theta^*)$ に対する最尤推定法による推定値を示す．

上記の Naive-Bayes 法における推定値が真のパラメータと一致する場合には，上式によって分類を間違えてしまう確率である誤り率を最小にするという意味での最適性が理論的に保証されている[2)]．しかし，最尤推定法を採用しているので，この最適性は学習用の電報数が無限の場合のみ保証され，学習用の電報数が有限の場合には最適性は保証されない．

そこで，本研究では統計的決定理論に基づき，学習用の電報数が有限の場合に未知の電報を間違った分野に分類してしまう確率である誤り率をベイズ基準のもとで最小化するという意味で，最適な電報分類方法を以下で提案する．

4. 統計的決定理論に基づく電報分類問題

本研究では，統計的決定理論[1),9)]に基づいて，電報分類問題を考え直す．

4.1 事前確率密度関数の定義

ここでは，本研究で新たに電報分類問題の定式化に導入する事前確率密度関数の定義を行う．その他の記号等に関しては，2 章および 3 章における定義と同様である．

$p(\theta)$ はパラメータ θ の事前確率密度関数を示し既知である．

4.2 統計的決定理論に基づく電報分類問題の概要

本研究では，学習用の電報数が有限の場合に未知の電報を間違った分野に分類してしまう確率である誤り率を統計的決定理論に基づいて最小化するという意味で最適な電報分類方法を提案する．

真のパラメータを $\theta^*, \theta^* \in \Theta$ とすると，まず，真のパラメータ θ^* によって支配される $p(c_i|\theta^*)$ と $p(key_j|c_i,\theta^*)$ に基づいて学習用の電報 $(w_i, z^{N_i})^G$ と未知の電報の分類されるべき真の分野 c^* と未知の電報のキーワード系列 y^M の 2 項組み (c^*, y^M) が生起する．しかし，実際には真のパラメータ θ^* と未知の電報の分類されるべき真の分野 c^* は未知で，学習用の電報 $(w_i, z^{N_i})^G$ と未知の電報のキーワード系列 y^M のみが観測される．よって，電報分類問題は学習用の電報 $(w_i, z^{N_i})^G$ と未知の電報のキーワード系列 y^M を観測したもとで，未知の電報 doc_u の分類されるべき真の分野 c^* を推定する問題として解釈できる．

本研究では，未知の電報 doc_u の分類されるべき真の分野 c^* を推定する際に，間違った分野を推定結果として出力してしまう確率である誤り率を損失関数として導入し，統計的決定理論に基づいて，誤り率の最小化を図る．最適性の基準にはミニマックス基準，ベイズ基準等様々な基準が存在するが，本研究ではベイズ基準を採用する．

4.3 統計的決定理論に基づく電報分類問題の定式化

以下で，統計的決定理論に基づく電報分類問題の定式化を行う．

4.3.1 損 失 関 数

式 (7) で示される損失関数は，パラメータ θ によって支配される $p(c_i|\theta)$ と $p(key_j|c_i,\theta)$ に基づいて，未知の電報 doc_u が分類されるべき真の分野 x（x は前述の c^* のことである）と未知の電報 doc_u のキーワード系列 y^M の 2 項組み (x, y^M) が生起する場合に，決定関数 $d(y^M)$ を用いて間違った分野 c_i を分野 x の推定結果として出力する確率である誤り率を表す．定性的には，パラメータ θ によって支配される状況で，未知の電報 doc_u のキーワード系列 y^M を観測した場合に間違った分野 c_i に未知の電報 doc_u を分類する確率を意味する．

$$
\begin{aligned}
L(d,\theta) &= \sum_{x\in C}\sum_{y^M\in KEY^M} p(x,y^M|\theta) I(d(y^M)) \\
&= \sum_{x\in C}\sum_{y^M\in KEY^M} p(x|\theta)\prod_{i=1}^{M} p(y_i|x,\theta) I(d(y^M)),
\end{aligned}
\tag{7}
$$

ただし，$d(y^M)$ は未知の電報 doc_u のキーワード系列 y^M を引数にとり未知の電報 doc_u の分類されるべき分野 x の推定結果を出力する決定関数（左辺の d も同じ決定関数を示す），$I(d(y^M))$ は $d(y^M)$ が分野 x と等しい正しい分野を出力すれば 0，分野 x と異なる間違った分野を出力すれば 1 を返すインディケータで次式で示される，

$$
I(d(y^M)) = \begin{cases} 1\,, & d(y^M) \neq x; \\ 0\,, & d(y^M) = x. \end{cases}
\tag{8}
$$

4.3.2 リスク関数

式 (9) で示されるリスク関数は，パラメータ θ によって支配される $p(c_i|\theta)$ と $p(key_j|c_i,\theta)$ に基づいて学習用の電報 $(w_i, z^{N_i})^G$ が生起し，決定関数 $d(y^M)$ を用いる場合の損失関数の期待値，言い換えると誤り率の学習用の電報に関する期待値を表す．

$$
\begin{aligned}
R(d,\theta) &= \sum_{(w_i,z^{N_i})^G\in(C,KEY^{N_i})^G} p((w_i,z^{N_i})^G|\theta) L(d,\theta) \\
&= \sum_{(w_i,z^{N_i})^G\in(C,KEY^{N_i})^G} \prod_{i=1}^{G}\left(p(w_i|\theta)\prod_{j=1}^{N_i} p(z_{i,j}|w_i,\theta)\right) L(d,\theta).
\end{aligned}
\tag{9}
$$

4.3.3 ベイズリスク

式 (10) で示されるベイズリスクは，パラメータ θ の事前確率密度関数 $p(\theta)$ に関するリスク関数の期待値を表す．

$$
BR(p(\theta)) = \int_{\Theta} p(\theta) R(d,\theta) d\theta.
\tag{10}
$$

ベイズリスクを最小にする決定をベイズ決定 $BD(p(\theta))$ と呼び，ベイズ決定はベイズ基準のもとで最適な決定を示す．式 (10) によるベイズリスクを最小にする式 (11) によるベイズ決定は，誤り率をベイズ基準のもとで最小にするという意味で最適な電報分類方法である．

$$
BD(p(\theta)) = \arg\min_{d} BR(p(\theta)).
\tag{11}
$$

5 章において，実際に式 (11) を満足する最適な電報

分類方法を提案する．

5. 提案方法

学習用の電報数が有限の場合に未知の電報 doc_u を間違った分野に分類する確率である誤り率をベイズ基準のもとで最小にするという意味で最適な電報分類方法の実際のアルゴリズムを提案するにあたって，まず，式 (10) によるベイズリスクを書き下してみる．

$$BR(p(\theta)) = \sum_{(w,z^N)^G} br\_1 \times br\_2 \times \cdots \times br\_g \times \cdots \times br\_G \times br\_u, \tag{12}$$

ただし，

$$br\_1 = \int_\Theta p(\theta)p(w_1|\theta)d\theta \int_\Theta p(\theta|w_1)p(z_{1,1}|w_1,\theta)d\theta \int_\Theta p(\theta|w_1,z_{1,1})p(z_{1,2}|w_1,\theta)d\theta \cdots \int_\Theta p(\theta|w_1,z^{N_1-1})p(z_{1,N_i}|w_1,\theta)d\theta, \tag{13}$$

$$br\_2 = \int_\Theta p(\theta|(w_1,z^{N_1}))p(w_2|\theta)d\theta \int_\Theta p(\theta|(w_1,z^{N_1}),w_2)p(z_{2,1}|w_2,\theta)d\theta \cdots \int_\Theta p(\theta|(w_1,z^{N_1}),w_2,z^{N_2-1}) p(z_{2,N_2}|w_2,\theta)d\theta, \tag{14}$$

$$br\_g = \int_\Theta p(\theta|(w,z^N)^{g-1})p(w_g|\theta)d\theta \prod_{i=1}^{N_g}\int_\Theta p(\theta|(w,z^N)^{g-1},w_g,z^{i-1}) p(z_{g,i}|w_g,\theta)d\theta, \tag{15}$$

$$br\_G = \int_\Theta p(\theta|(w,z^N)^{G-1})p(w_G|\theta)d\theta \prod_{i=1}^{N_G}\int_\Theta p(\theta|(w,z^N)^{G-1},w_G,z^{i-1}) p(z_{G,i}|w_G,\theta)d\theta, \tag{16}$$

$$br\_u = \sum_x \int_\Theta p(\theta|(w,z^N)^G)p(x|\theta)d\theta \sum_{y^M}\prod_{i=1}^{M}\int_\Theta p(\theta|(w,z^N)^G,x,y^{i-1}) p(y_i|x,\theta)d\theta I(d(y^M)), \tag{17}$$

$(w,z^N)^{g-1}$ は長さ g—1 の学習用の電報の系列 $(w_1,z^{N_1})(w_2,z^{N_2})\cdots(w_{g-1},z^{N_{g-1}})$ を示し，z^{i-1} および y^{i-1} はそれぞれの電報の 1 番目から i—1 番目のキーワードが並んだ長さ i—1 のキーワード系列を示し，$p(\theta|(w,z^N)^{g-1})$，$p(\theta|(w,z^N)^G)$，$p(\theta|(w,z^N)^{g-1},w_g,z^{i-1})$，$p(\theta|(w,z^N)^G,x,y^{i-1})$ はパラメータ θ の事後確率密度関数を示す．さらに，式 (17) の br_u を変形してみると次式のようになる．

$$br\_u = \sum_{y^M}\sum_x \int_\Theta p(\theta|(w,z^N)^G)p(x|\theta)d\theta \prod_{i=1}^{M}\int_\Theta p(\theta|(w,z^N)^G,x,y^{i-1}) p(y_i|x,\theta)d\theta I(d(y^M)). \tag{18}$$

学習用の電報 $(w,z^N)^G$ と未知の電報のキーワード系列 y^M を受け取ったもとで，実際にどの分野を未知の電報の分類先として決定するかによってベイズリスクの値を変化させるのは，

$$\sum_x \int_\Theta p(\theta|(w,z^N)^G)p(x|\theta)d\theta \prod_{i=1}^{M}\int_\Theta p(\theta|(w,z^N)^G,x,y^{i-1})p(y_i|x,\theta)d\theta I(d(y^M))$$

の部分である．よって，最適な決定 $d_{Bayes}(y^M)$ は次式によって実現できる．

$$d_{Bayes}(y^M) = \arg\max_{x\in C}\int_\Theta p(\theta|(w,z^N)^G)p(x|\theta)d\theta \prod_{i=1}^{M}\int_\Theta p(\theta|(w,z^N)^G,x,z^{i-1}) p(y_i|x,\theta)d\theta. \tag{19}$$

以上のように，学習用の電報数が有限の場合に未知の電報を間違った分野に分類してしまう確率である誤り率をベイズ基準のもとで最小にするという意味で最適な電報分類方法を導出することができた．

なお，パラメータ θ の事前確率密度関数 $p(\theta)$ としてベータ分布を採用することにより，式 (19) 中の積分の計算は式 (20) および式 (21) のように容易になる．

$$\int_\Theta p(\theta|(w,z^N)^G)p(x|\theta)d\theta = \frac{F(x|w^G)+\beta(x)}{\sum_{x\in C}(F(x|w^G)+\beta(x))}, \tag{20}$$

ただし，$F(x|w^G)$ は学習用の電報全体の中で分野 x の電報が生起した回数，$\beta(x)$ はベータ分布の $p(x|\theta)$ に対するパラメータを示す．

$$\int_\Theta p(\theta|(w,z^N)^G,x,z^{i-1})p(y_i|x,\theta)d\theta = \frac{F((x,y_i)|(w,z^N)^G)+F(y_i|y^{i-1})+\beta(y_i|x)}{\sum_{y_i\in KEY}(F(x,y_i)|(w,z^N)^G)+F(y_i|y^{i-1})+\beta(y_i|x))}, \tag{21}$$

表 1 提案方法（bayes）とベクトル空間法（vec）の比較
Table 1 Comparison of proposed algorithm (bayes) with previous algorithm (vec).

		学習用の電報数			
		5000	10000	100000	150000
成功率	bayes	80.70	83.77	84.09	84.43
（%）	vec	79.25	79.13	79.27	79.54

ただし，$F((x,y_i)|(w,z^N)^G)$ は学習用の電報全体の中で分野 x の電報中でキーワード y_i が生起した回数，$F(y_i|y^{i-1})$ はキーワード系列 y^{i-1} 中でキーワード y_i が生起した回数，$\beta(y_i|x)$ はベータ分布の $p(y_i|x,\theta)$ に対するパラメータを示す．

6. 従来方法と提案方法の比較

6.1 ベクトル空間方法に基づく電報分類方法との比較

ここでは，一例にすぎないが実際の電報データを用いた分類実験を行うことによって，我々が説明用に先に考えた基本的なベクトル空間法に基づく電報分類方法と提案方法との比較を行う．

この実験では人手によって「結婚」,「クリスマス」等の 20 の分野に分類されている電報データを用いた（$|C| = 20$）．学習用の電報を全体で 150,000 通，学習用の電報とは別に試験用の未知の電報を 40,981 通用意した．キーワードは学習用の電報の本文を形態素解析に掛け，各分野ごとに名詞，動詞，独立詞の 3 品詞を合わせて，頻度上位 50 個ずつ採用し，分野間の重複を除くと，キーワード数は 438 個となった（$|KEY| = 438$）．採用されたキーワードには,「おめでとう」,「祈」,「赤ちゃん」,「合格」等が含まれている．**表 1** および **図 1** は分類実験結果である．学習用の電報数は学習用に利用した電報数を示し，電報数 G が 5,000，10,000，100,000，150,000 と徐々に増加している．成功率は未知の電報 40,981 通に対して何通の分類に成功したかを相対頻度による百分率で示している．なお，ここでの正解は未知の電報に対して事前に人手で付与された分野である．また，Vec は式 (1) のベクトル空間方法による分類結果を示し，Bayes は式 (19) の提案方法による分類結果を示す．

表 1 および図 1 による実験結果を統計的に評価するため，分類に成功するか失敗するかが 2 項分布に従っていると仮定して，母不良率に関する検定[6]を行った．その結果，99%の信頼率のもとで，提案方法の分類精度が基本的なベクトル空間法に基づく電報分類方法の分類精度を上回るという検定結果が得られた．

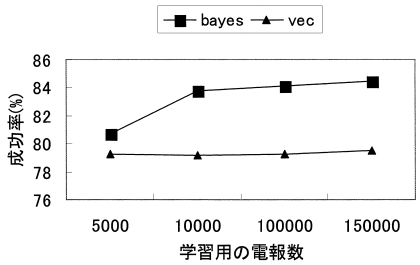

図 1 提案方法（bayes）とベクトル空間法（vec）の比較
Fig. 1 Comparison of proposed algorithm (bayes) with previous algorithm (vec).

6.2 Naive-Bayes 法に基づく電報分類方法との比較

ここでは，Naive-Bayes 法に基づく電報分類方法と提案方法との関係について考察する．

電報分類問題においては，分野の生起確率等を支配する真のパラメータ θ^* は未知であるが，仮に既知であった場合を考えてみる．θ^* が既知の場合には，式 (9) のリスク関数を最小にする電報分類方法が最適な電報分類方法 $d^*(y^M)$ であり，次式で示される．

$$d^*(y^M) = \arg\max_{x\in C} p(x|\theta^*) \prod_{i=1}^{M} p(y_i|x,\theta^*). \tag{22}$$

Naive-Bayes 法に基づく電報分類方法の式 (6) と式 (22) を比較してみると，Naive-Bayes 法に基づく電報分類方法においてパラメータの推定値が真のパラメータと一致した場合には，Naive-Bayes 法において誤り率を最小にするという意味で最適性が保証されることが分かる[2]．これは最尤推定法を採用している場合だと，学習用の電報数が無限の場合に相当する．しかし，学習用の電報数が有限の場合には定性的には電報数が増加するにともないパラメータの推定精度が向上し，分類精度も高くなることは想像できるが，理論的な保証はない．

一方，提案方法は統計的決定理論に基づいて，学習用の電報数が有限の場合にベイズ基準のもとで誤り率を最小にするような決定関数および推定値を導出しているので，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性が保証されている．

7. ま と め

本研究では，電報を既存の分類大系に従って分類する電報分類問題を統計的決定理論に基づいて考え直した．確率モデルを導入して定式化した電報分類問題は，従来から研究されている文書分類問題と同一の問題で

あった．しかし，従来の文書分類方法を適用した電報分類方法は，分類精度を理論的に保証することができないか，あるいは学習用の電報数が無限の場合には最適性が保証できても有限の場合には分類精度を理論的に保証することができない．

そこで，本研究では統計的決定理論に基づき，学習用の電報数が有限の場合に新規の未知の電報を間違った分野に分類してしまう確率である誤り率をベイズ基準のもとで最小にするという意味で最適な電報分類方法を提案した．

我々が説明用に考えた基本的なベクトル空間法に基づく分類方法と提案方法の比較として，一例ではあるが，実際の電報を用いた分類実験を通して基本的なベクトル空間法に基づく電報分類方法よりも提案方法の分類精度の方が統計的に見て高いことを確認した．また，Naive-Bayes 法に基づく電報分類方法と提案方法との違いを，アルゴリズムの導出過程の違いから考察した．

なお，ベクトル空間法と提案方法との比較においては，情報検索の分野で研究されてきたベクトル空間法の本質的な部分のみを基本的なベクトル空間法として定義して比較実験を行った．一方，実際の文書分類の分野における従来研究においては，文書分類に適した様々な工夫が盛り込まれている．よって，本論文における比較実験では，単にキーワードの生起の仕方の相関を見るだけの単純なベクトル空間法と提案方法の比較を行ったにすぎない．そのため，従来研究における様々な工夫が盛り込まれたベクトル空間法に基づく分類方法は分類精度に関しての理論的な保証はないが，実際の分類精度では提案方法よりも優れている場合もある．

今後の課題としては，本論文の提案方法に上記従来研究の様々な工夫を盛り込むことによる高精度化も期待できる．また，現在は単に学習用の電報から学習した 1 つの確率モデルを利用しているが，電報というものの性質を考えてみると，電報の内容には季節的な変動がある．そこで，季節によって異なった確率モデルを利用することによって，さらなる高精度化が期待される．また，電報には送り先等の情報が付与されているので，送り先が結婚式場であれば「結婚」に分類するというようなルールと組み合わせることによる高精度化も考えられる．

参 考 文 献

1) Berger, J.: *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York (1980).
2) Domingos, P. and Pazzani, M.: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*, Vol.29, pp.103–130 (1997).
3) Iwayama, M. and Tokunaga, T.: 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).
4) Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *International Conference on Machine Learning'97*, pp.143–151 (1997).
5) Fujii, A., Inui, K., Tokunaga, T. and Tanaka, H.: Selective Sampling for Example-based Word sense Disambiguation, *Computational Linguistics*, Vol.24, No.4, pp.573–597 (1998).
6) Manning, C. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, pp.237–239, MIT Press, London (1999).
7) 永田 靖：入門統計解析法，pp.223–225，日科技連，東京 (1992).
8) Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing & Management*, Vol.24, No.5, pp.513–523 (1988).
9) 繁桝算男：ベイズ統計入門，東京大学出版会 (1985).
10) 砂場倫太郎，堀井統之，今村賢治，大山芳史：メッセージ種別判定方式の検討，電子情報通信学会春季大会論文集，No.6, p.114 (1993).
11) 徳永健伸：情報検索と言語処理，東京大学出版会，東京 (1999).
12) Witten, I., Moffat, A. and Bell, T.: *Managing Gigabytes*, Van Norstrand Reinhold, New York (1994).

(平成 12 年 3 月 29 日受付)
(平成 14 年 9 月 5 日採録)

前田 康成（正会員）

平成 7 年早稲田大学理工学部工業経営学科卒業．平成 9 年同大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．NTT サイバースペース研究所を経て，現在，NTT 東日本研究開発センタ勤務．機械学習，統計的決定理論，自然言語処理の研究に従事．

小原　永（正会員）

昭和 52 年慶應義塾大学工学部電気工学科卒業．昭和 54 年同大学大学院工学研究科電気工学専攻修士課程修了．同年日本電信電話公社情報通信研究所入社．現在，日本電信電話（株）サイバースペース研究所勤務，プロジェクトマネージャ．主に推敲支援，韻律生成技術の研究に従事．電子情報通信学会会員．