川崎医療福祉学会誌　Vol. 15　No. 2　2006　539–552

原　著

Bayesian Spam Filterを用いた要約の自動分類の試み

田　中　昌　昭[*1]

要　　約

保健医療分野では，病名や検査データのように構造化されたデータ以外に，インシデントレポート，放射線読影レポート，退院時サマリなど，構造を持たないテキストデータを扱うことが多い．これらのテキストデータを有効に活用するには何らかの方法でコード化して保存や検索を容易にする必要がある．しかしながら，日常の診療において日々大量に発生するテキストデータを人手によって分類し，コード化する作業は容易なことではない．そのような場合，機械学習の分野で培われてきたテキスト分類技術を利用して，分類作業の自動化を行うことに期待が集まる．

本研究では，迷惑メール（Spam）のフィルタとして考案されたBayesian Spam Filterを医学テキストの自動分類に適用し，その可用性を検討した．Bayesian Spam Filterは，Bayes理論に基づき，分類済みのコーパスからテキストを構成する単語と分類カテゴリの関連性の度合いを学習し，未知のテキストを分類する技術である．医学テキストとしてPubMedからキーワードを指定して収集したabstractを用いた．Bayesian Spam Filterには当初Grahamによって考案され，その後，Robinsonによって改良が加えられたモデルを利用した．予備実験として，これらのモデルが使うパラメタの最適値を求め，それらを用いて各モデルの分類性能を調べた．

その結果，最大でRecallが96.0%，Precisionが92.9%という分類性能を得た．これは，Bayesian Spam Filterの本来の目的であるSpamの分類成績には遠く及ばないものの，改善次第では十分に実用に耐えられる成績である．同時に，Robinsonの改良モデルにχ^2による特徴選択を適用することにより，分類性能が向上することを明らかにした．

はじめに

テキスト分類（Text Categorization：TC）は，自然言語で記述されたテキストを，その内容に基づいて，あらかじめ定義されたカテゴリに自動分類する問題である[1)]．近年，医療分野では，診療情報の電子化が進み，インシデントレポート[2)]，放射線読影レポート[3)]，退院時サマリ[4−6)]など，自然言語で記述されたテキストデータが蓄積されつつあるが，それらを有効に活用する上で，テキスト分類技術は重要な役割を果たすものと期待されている．

テキスト分類の手法には様々なものがあるが，なかでもBayesの定理に基づく確率的手法は単純な割には効果的な方法として多くの研究がなされている[7−9)]．最近では，迷惑メール（Spamメールとも呼ばれる）のフィルタに応用され，大きな成果をあげている[10−16)]．本研究では，Grahamによって提案され[13,14)]，その後，Robinsonによって改良されたSpamフィルタ（Bayesian Spam Filter：BSF）[15,16)]を用いて，医学文書の自動分類がどの程度まで可能かを検証するために，PubMed[17)]から収集した医学論文の抄録の自動分類を試みた．

Bayesian Spam Filter

BSFは，メールがSpamかどうかをそのメールに含まれる単語wのSpam確率（以降，これをToken Spam Probability：TSPと呼ぶ）に基づいて推定する．ここで，単語wのSpam確率$P(Spam \mid w)$とは，メール中に単語wが含まれることがわかっているという条件のもとで，そのメールがSpamである事後確率で，Bayesの定理によって次式によって計算することができる．

[*1] 川崎医療福祉大学　医療福祉マネジメント学部　医療情報学科
（連絡先）田中昌昭　〒701-0193　倉敷市松島288　川崎医療福祉大学
E-Mail: mtanaka@mw.kawasaki-m.ac.jp

$$P(Spam \mid w) = \frac{P(Spam) \cdot P(w \mid Spam)}{P(Spam) \cdot P(w \mid Spam) + P(Nonspam) \cdot P(w \mid Nonspam)} \quad (1)$$

ここで,$P(Spam)$,$P(w \mid Spam)$,$P(Nonspam)$,$P(w \mid Nonspam)$ はそれぞれ，ランダムに 1 通のメールを取り出したとき，それが Spam である確率，Spam メールからランダムに 1 通のメールを取り出したとき，それが単語w を含む確率，ランダムに 1 通のメールを取り出したとき，それが Spam でない（Nonspam）確率，Spam でないメールからランダムに 1 通のメールを取り出したとき，それが単語w を含む確率で，いずれも過去に受け取った Spam/Nonspam に分類済みのメールを訓練用コーパスとして用いて学習によって見積もることができる事前確率である．このように，BSF は教師付きの機械学習である．次に，メールd の Spam 確率 $P(Spam \mid d)$ を求めるには，メールを単語の集合 $d = \{w_1, \cdots, w_n\}$とみなし，構成単語のTSPの結合確率を計算する．結合確率の計算方法は，GrahamとRobinsonでは異なり，その詳細は次章で説明する．

Graham モデル

Graham はメールd の Spam 確率$P(Spam \mid d)$ を求めるために，メール中に出現する単語は互いに独立であるという仮定（独立性の仮定：Independence assumption）のもとに次のような結合確率[18] を計算した[13,14]．

$$P(Spam \mid d) = P(Spam \mid w_1, \cdots, w_n) \quad (2)$$

$$= \frac{{}_k P(Spam \mid w_k)}{{}_k P(Spam \mid w_k) + {}_k P(Nonspam \mid w_k)} \quad (3)$$

また，事前確率は次式によって見積もった．

$$P(Spam) = DF(Spam)/\sharp Tr$$
$$P(Nonspam) = DF(Nonspam)/\sharp Tr$$
$$P(w \mid Spam) = DF(w, Spam)/DF(Spam)$$
$$P(w \mid Nonspam) = DF(w, Nonspam)/DF(Nonspam)$$

ここで，DF は文書頻度（Document Frequency）を意味し，$DF(Spam)$，$DF(Nonspam)$，$DF(w, Spam)$，$DF(w, Nonspam)$ は，それぞれ，訓練用メール中の Spam 数，Spam でないメール数，単語w を含む Spam 数，単語w を含む Spam でないメール数である．また，$\sharp Tr$ は訓練用メールの総数である．これらの式を（1）式に代入すると

$$p(w) = \frac{b(w)}{b(w) + g(w)} \quad (4)$$

となる．ここで，Graham にしたがって$p(w) = P(Spam \mid w)$，$b(w) = DF(w, Spam)$，$g(w) = DF(w, Nonspam)$ と略記した．また，この表記を用いると，（3）式は

$$P(Spam \mid d) = \frac{\prod_k p(w_k)}{\prod_k p(w_k) + \prod_k (1 - p(w_k))} \quad (5)$$

と表すことができる．なお，Graham は，（4）式において TSP を計算する際に，FP（False positive：この場合，Spam でないメールを Spam として誤判定すること）を小さくするようにバイアスをかけるため，分母の$g(w)$ を$2g(w)$ とし計算しているが，本研究では（4）式をそのまま使った．

（5）式によってメールの Spam 確率を計算する際，判定対象のメール中に訓練用コーパスに出現しない未知の単語w が含まれている場合に問題が発生する（ゼロ頻度問題[19]）．まず，$b(w) = g(w) = 0$ の場合，（4）式は計算できない．また，$b(w) = 0$, $g(w) \neq 0$ の場合, $p(w) = 0$ となり，(5) 式からメールの Spam 確率が 0 になる．これは，判定対象のメールが Spam である確率が 0，即ち，Spam でないという意味になる．逆に，$b(w) \neq 0, g(w) = 0$ の場合，$p(w) = 1$ となり，(5) 式からメールの Spam 確率は 1 になる．これは, 判定対象のメールが Spam である確率が 1，即ち，Spam であるという意味になる．しかし，これは訓練用コーパスの不完全さに由来するもので，そのまま受け入れるわけにはいかない．訓練用コーパスに出現する単語の頻度 $n(w) = b(w) + g(w)$ が少ない場合についても同様のこと（訓練用コーパスの不完全さに由来）が言えるので，事前に定めたある特定の頻度（これを *DF threshold* と呼ぶことにする）よりも出現頻度$n(w)$ が小さい単語については，これも事前に定めたある TSP の値（これを *Assumed probability* と呼ぶことにする）を未知の単語に付与した．

さらに，Graham は（5）式を計算する際，すべての単語w_k の$p(w_k)$ を使うのではなく，文書の性

格を最も特徴付ける単語をいくつか選択して計算した．これは特徴選択（Feature Selection）と呼ばれる手法で，一般にテキスト分類問題では特徴選択によって分類性能が向上することが報告されている[9,20,21]．特徴選択の方法にも様々なやり方があるが，Grahamは（4）式によって計算されるTSPが中立値である0.5から最も離れているものを事前に定めた単語数（これを*Features*と呼ぶことにする）だけ選択して計算を行った．

これ以外に，これは多分に計算上の便宜的なものと考えられるが，Grahamは（4）式で得られたTSPに上下限を設定し，それを超えた（あるいは下回った）ものを次式によって補正した．

$$p(w) = \begin{cases} 1 - TSPCO & p(w) > 1 - TSPCO \\ TSPCO & p(w) < TSPCO \end{cases}$$

ここで，*TSPCO*(Token Spam Probability Cut-Off)は補正パラメタである．このような作為的な補正が必要となる理由は，結合確率を求める（5）式の分母には，$p(w)$や$1 - p(w)$が積の形で現われるため，$p(w)$が十分0（または1）に近い場合，誤差が拡大して計算結果が不安定になる可能性があるためと思われる．以上述べたモデルのパラメタを表1にまとめた．Grahamは*Assumed probability*=0.4，*DF threshold*=5，*Features*=15，*TSPCO*=0.01を使ったが，これらの値は対象とするコーパスに依存すると考えられるので，実際の計算では，予備実験を行って最適値を求めた．詳細については後述する．

表1　Graham モデルのパラメタ

パラメタ	説明
Assumed probability	訓練用コーパスの不完全さを補うために仮定したToken Spam Probability
DF threshold	*Assumed probability*付与の目安となる文書頻度
Features	選択する単語の数
TSPCO	Token Spam Probabilityの上下限

表中に示すパラメタの名称はいずれも著者が独自につけたもの．

こうして得られたメールのSpam確率$P(Spam \mid d)$がある閾値を超えた場合，そのメールをSpamと判定した．

Robinson–Fisher モデル

RobinsonはGrahamが考案したモデルに2つの改良を加えた[15,16]．まず，第一に，低出現頻度の単語に関する改良である．GrahamはTSPを算出する（4）式において，根拠があまり明確でない作為的な方法でゼロ頻度問題を回避した．これに対してRobinsonは，（4）式の代わりに

$$f(w) = \frac{robs \cdot robx + n(w) \cdot p(w)}{robs + n(w)} \tag{6}$$

を用いた．ここで，$robx$, $robs$はパラメタで，それぞれ，ゼロ頻度語（$n(w) = 0$）に付与するTSP及び$robx$の重みである．また，$n(w)$は訓練用コーパス中の単語wを含むメールの総数で，$p(w)$は（4）式で与えられるGrahamの提案したTSPである．（6）式は，形式的には$robx$と$p(w)$の重みつき平均で，それぞれの重みが$robs$及び$n(w)$になっている．（6）式は$n(w) \to 0$で$robx$となり，$n(w) \to \infty$で$p(w)$になる．つまり，（6）式はGrahamの提案したTSPを連続的に低頻度語やゼロ頻度語まで拡張したものとなっている．

Robinsonが行った2番目の改良は結合確率の計算方法である．Robinsonは結合確率を計算するのにFisherの方法[22]を用いた．つまり，「$f(w)$は正確で，かつ，メールは一様な分布$f(w)$にしたがって互いに独立に出現するランダムな単語の集まりである」という帰無仮説のもとでは，$-2\ln\prod_{k=1}^{n} f(w_k)$は自由度$2n$の$\chi^2$分布に従うとし，単語のTSPから結合確率

$$S = Chidist(-2\ln\prod_{k=1}^{n} f(w_k), 2n)$$
$$H = Chidist(-2\ln\prod_{k=1}^{n} (1 - f(w_k)), 2n)$$

を計算し，これらを使ってメールのSpam確度

$$I = \frac{1 + S - H}{2} \tag{7}$$

を求めた．ここで，$Chidist(\cdots)$はχ^2の分布関数で，nは文書中に含まれる単語の数である．Robinsonは，このアプローチが他の多くのアプローチと決定的に違うところは独立性の仮定を前提として

いない点であると主張している．実際，Grahamは（3）式を得るのに独立性の仮定を用いており，またBayes理論に基づく他の多くの手法は独立性の仮定を用いて計算を簡略化している（Naive Bayes）[7]．そして独立性の仮定は，メール中に単語が互いに独立に出現する，という極めて非現実的な仮定であるので，そのような仮定を前提として作ったモデルは計算方法としては不適切であると述べている．一方，上述したFisherの結合確率の計算方法では，単語の独立性は棄却すべき帰無仮説の中に取り込まれており，したがって結合確率の計算自体は単語の独立性とは無関係に正しいと主張している．この問題については特徴選択との関連で実験結果を踏まえながら後で考察を行う．表2にRobinson-Fisherモデルで使うパラメタをまとめた．

表2 Robinson-Fisherモデルのパラメタ

パラメタ	説明
robx	訓練用コーパスの不完全さを補うために仮定したToken Spam Probability
robs	*robx*に与える重み

コーパス

本研究では，メールではなくPubMedから収集した英文abstractに対してBSFによるテキスト分類を試みた．コーパスは，PubMedでキーワードに“liver cancer”を入力して得たabstract（便宜上，これをSpamと呼ぶ）と，キーワードを何も指定しないで得たabstract（同様に，これをNonspamと呼ぶ）の2種類を作成し，各々訓練用コーパスと検証用コーパスに2分した．なお，SpamとNonspamのいずれにも含まれるabstractは除去した．得られたコーパスの要約を表3に示す．コーパスは全部で8,970件の英文abstractからなり，1件当たり平均で116語を含んでいた．また，コーパス中に出現する異なり語の数は，訓練用では48,637語，検証用では48,257語，そして全体では71,779語であった．

評価指標

モデルの分類性能を評価する指標として，以下の4種類の指標を使った．

1．最小Bayesエラー（BE）[23]

ここで，TP（True Positive），FN（False Negative），TN（True Negative），FP（False Positive）は，それぞれモデルがSpamを正しくSpamと判定した数，Spamを間違ってNonspamと判定した数，Nonspamを正しくNonspamと判定した数，そしてNonspamを誤ってSpamと判定した数である．これらの値はSpam確率の閾値を0から1に変化させるとそれに応じて変わるので，それに伴って上式の値もまた変化するが，その最小値が最小Bayesエラーである．

2．Precision（π）[7,24]

Precisionはモデルが Spamと判定したabstractのうち，実際にSpamである割合を示す指標で，次式によって計算される．

$$\pi = \frac{TP}{TP + FP}$$

3．Recall（ρ）[7,24]

Recallは実際のSpamのうち，モデルがSpamと判定したabstractの割合を示す指標で，次式によって計算される

$$\rho = \frac{TP}{TP + FN}$$

4．F–measure（F）[7,24]

F–measureは，Precisionと Recallの調和平均で，次式によって計算される．

$$F = \frac{2\pi\rho}{\pi + \rho}$$

予備実験

表1，2に示すGrahamモデルとRobinson–Fisherモデルのパラメタの最適値を得るために予備実験を行った．訓練用データ数$\sharp Tr$を400から2,400まで400ずつ増やしながら（それぞれSpam，Nonspamは同数とした），検証用データの半分（2245件; Spam 1122件，Nonspam 1123件）を使って分類実験を行い，最小Bayesエラーが最も小さくなるパラメタの値を求めて最適値とした．

1．Grahamモデルのパラメタの最適値

1.1．*Assumed probability*の最適値

図1は，*Assumed probability*を0.1から0.8まで0.1ずつ変化させたときの最小Bayesエラーをプロットしたものである．なお，このとき，他のパラメタについては表4に示すGrahamが用いた値を使った．

図が示すように，すべての$\sharp Tr$に対して*Assumed probability*の最適値は0.2であった．これは，はじめて見る語，あるいは滅多に現れない語のスパム確度（この場合“liver cancer”確度）は0.2，つまり低いということを示している．

表 3　本研究で使用した英文 abstract コーパス

コーパス	カテゴリ	Abstract数	平均語数	異なり語数
訓練用	Spam	2,241	135	48,637
	Nonspam	2,240	98	
検証用	Spam	2,244	133	48,257
	Nonspam	2,245	99	
全体		8,970	116	71,779

PubMedからキーワード “liver cancer” を指定して収集した英文 abstract を Spam，キーワードに何も指定しないで収集したものを Nonspam と記してある．異なり語数はコーパス中に出現する異なる単語の数である．

表 4　パラメタの最適値

モデル	パラメタ	既定値	$\sharp Tr$					
			400	800	1,200	1,600	2,000	2,400
Graham	*Assumed probability*	0.4	0.2	0.2	0.2	0.2	0.2	0.2
	DF threshold	5	9	9	6	4	8	9
	Features	15	16	16	16	16	16	16
	TSPCO	0.0100	0.0050	0.0001	0.0050	0.0100	0.0001	0.0005
Robinson-Fisher	*robx*	0.5	0.6	0.7	0.7	0.7	0.7	0.7
	robs	1.0	0.4	2.6	2.0	2.6	2.6	2.6

調べた範囲内での各モデルのパラメタの最適値．最小 Bayes エラーを極小にする値をもって最適値とした．Graham モデルの場合は Graham が使用した値を初期値として *Assumed probability, DF threshold, Features, TSPCO* の順に逐次最適値を決定した．Robinson-Fisher モデルは事前に決めた$(robx, robs)$ の組合せを総当りで調べて最適値を決定した．

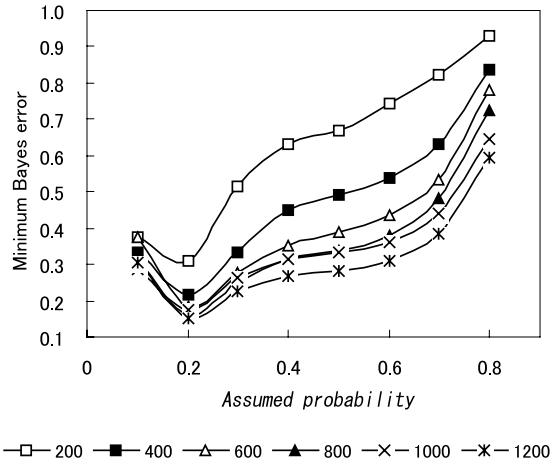

図 1　*Assumed probability* 依存性
Graham モデルにおいて，最小 Bayes エラー（縦軸）を指標とする分類性能（小さいほど性能は高い）の *Assumed probability* 依存性を訓練データ数$\sharp Tr$ 別にプロットしたもの．他のパラメタについては Graham の既定値（表 4 参照）を用いた．

1.2．*DF threshold* の最適値

図 2 は，*DF threshold* を 1 から20まで 1 ずつ変化させたときの最小 Bayes エラーをプロットしたものである．なお，このとき *Assumed probability* は前節で得た最適値を用い，その他のパラメタについては表 4 に示す Graham が用いた値を使った．

図から，訓練用データ数$\sharp Tr$ の増加に伴って，性能が *DF threshold* に依存しない範囲が広がっていく様子がわかる．$\sharp Tr \geq 400$ では，$5 \leq DF\ threshold \leq 15$ の範囲で最小 Bayes エラーはほぼ一定の値に収まっ

ている．もともと *DF threshold* は，訓練用データの不完全さを補うパラメタであった．つまり，訓練用データにおける出現頻度が *DF threshold* 未満の語に対しては，その出現頻度に基づいて計算した語のスパム確度は信頼できないため *Assumed probability* を付与するというものであった．したがって，$\sharp Tr$ の増加に伴って *DF threshold* 依存性が減少するのは理にかなっている．

1.3. *Features* の最適値

図3は，*Features* を1から512まで2倍ずつ変化させながら最小 Bayes エラーを計算してプロットしたものである（横軸は対数目盛にしてある）．なお，このとき *DF threshold*，*Assumed probability* はこれまでに得た最適値を用い，その他のパラメタについては表4に示す Graham が用いた値を使った．

図から，すべての $\sharp Tr$ に対して，$Features \leq 16$ では分類性能はほぼ安定しているが，*Features* が16を超えたあたりから悪化し始め，32〜64あたりで最

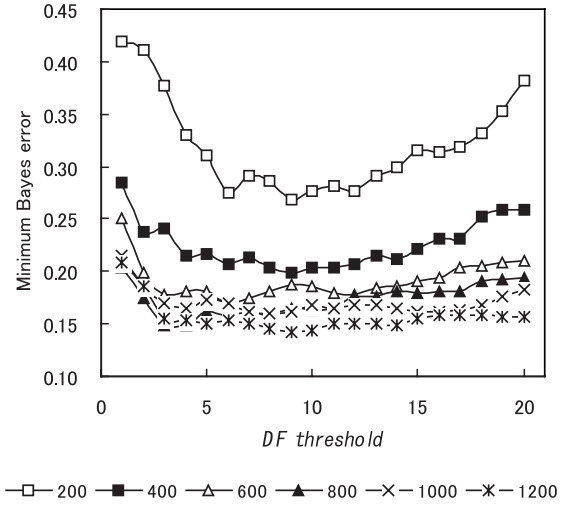

図2 *DF threshold* **依存性**
Graham **モデルにおいて，分類性能の** *DF threshold* **依存性を訓練データ数** $\sharp Tr$ **別にプロットしたもの．***Assumed probability* **は予備実験で得た最適値（図1参照）を，その他は** Graham **の既定値（表4参照）を用いた．**

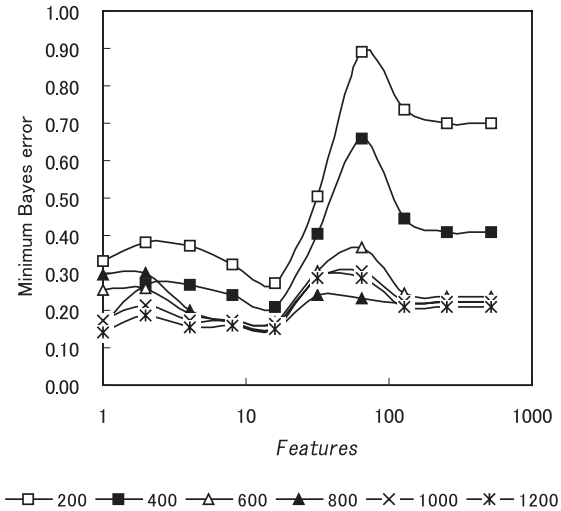

図3 *Features* **依存性**
Graham **モデルにおいて，分類性能の** *Features* **依存性を訓練データ数** $\sharp Tr$ **別にプロットしたもの．***Assumed probability, DF threshold* **は予備実験で得た最適値（図1，2参照）を，***TSPCO* **は** Graham **の既定値（表4参照）を用いた．**

も悪くなっていることがわかる．この結果は，特徴選択がモデルの分類性能を向上させることを示すものである．Graham は *Features* の値として15を用いたが，今回の予備実験の結果（すべての$\sharp Tr$で$Features = 16$が最適値）はそれとほぼ一致している．*Features* が128を超えると，いずれの$\sharp Tr$の場合も最小 Bayes エラーがそれぞれある値に収束しているように見えるが，コーパス中の平均語数が116語（表3参照）であることを考えるとこれは自明な結果である．

1．4．*TSPCO* の最適値

図4は，*TSPCO*=0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1に対して最小 Bayes エラーを計算してプロットしたものである（横軸は対数目盛である）．なお，このとき，その他のパラメタについては，これまでに得た最適値を用いた．

図から，すべての$\sharp Tr$に対して，$TSPCO \leq 0.01$では最小 Bayes エラーはそれぞれある一定の値に収まっており，それを超えると増大することがわかる．*TSPCO* は計算の便宜上導入されたパラメタなので，当然のことながら，十分な訓練データがある場合（$\sharp Tr$が大きい場合）は，さほど重要な役割はないことを示す結果である．

2．Robinson–Fisher モデルのパラメタの最適値

Robinson–Fisher モデルのパラメタは*robx* と *robs* の2つだけなので，*robs*=0.2, 0.6, 1.0, 1.4, 1.8, 2.2, 2.6の各値に対して，*robx* を0.2から0.8まで0.2ずつ増やしながら最小 Bayes エラーを計算した．訓練用データ数$\sharp Tr$のそれぞれについて分類実験を行った結果を図5に示す．

図から，訓練データ数$\sharp Tr$が少ない場合は，分類性能に与えるパラメタ*robx* と *robs* の影響は大きいが，$\sharp Tr$の増加に伴ってその影響は小さくなっていくことがわかる．もともと*robx* と *robs* は訓練データの不完全さに由来するゼロ頻度問題（低頻度問題）を補うために Robinson が導入したパラメタであったので，この結果は当然のことである．ただ．興味深いのは，*robx* に対応する Graham モデルのパラメタ *Assumed probability* はすべての$\sharp Tr$に対して0.2であったのに対して，Robinson-Fisher モデルでは *robx* の最適値は0.5以上になっている点である（表4参照）．つまり，Robinson–Fisher モデルでは低頻度語のスパム確度にはスパム側にバイアスがかかっている．(4) 式で与えられる語wのスパム確率$p(w)$が小さくても，(6) 式で与えられる Robinson–Fisher モデルのスパム確度は高めに見積もられることになる．この傾向は，語の出現頻度$n(w)$が少ないほど顕著になる．

ちなみに Robinson は初期値として*robx*=0.5, *robs*=1.0から開始して，最適値を決定すればよいと述べている[15]．

予備実験で得られた各モデルの訓練用データ数$\sharp Tr$別に求めたパラメタの最適値を表4にまとめた．

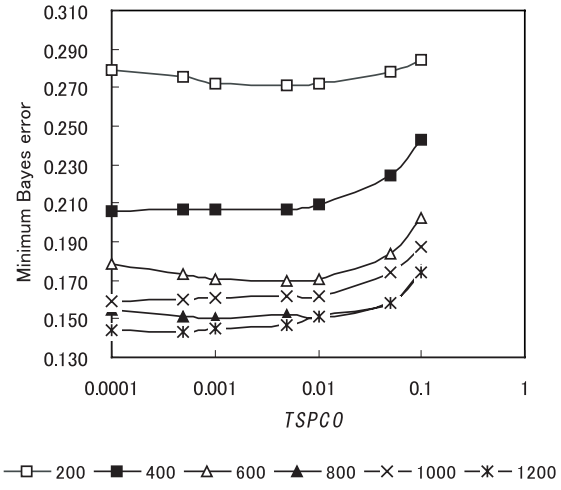

図4　*TSPCO* 依存性
Graham モデルにおいて，分類性能の *TSPCO* 依存性を訓練データ数$\sharp Tr$別にプロットしたもの．他のパラメタは予備実験で得た最適値（表4参照）を用いた．

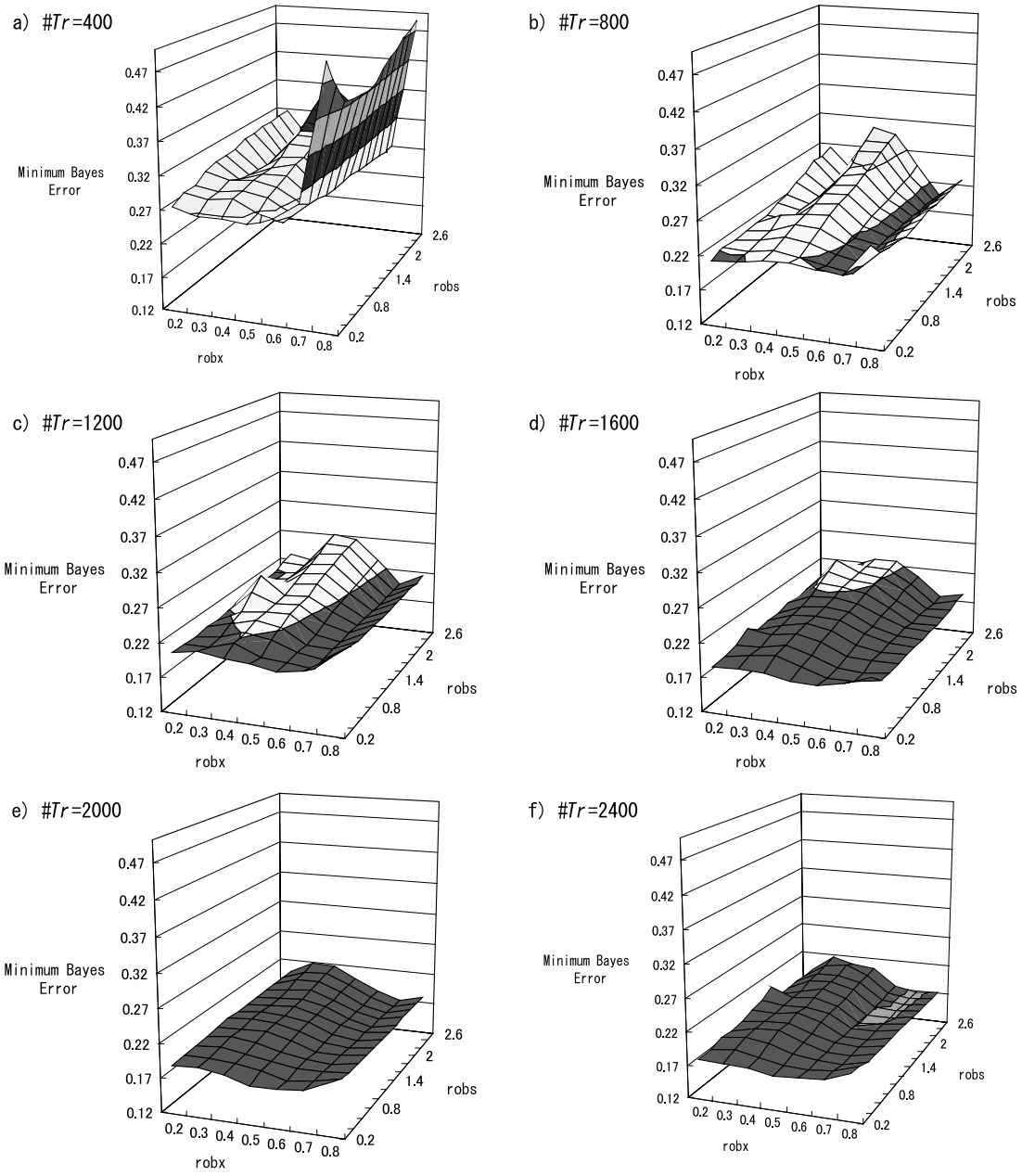

図5　Robinson-Fisher モデルのパラメタ依存
最小 Bayes エラー（縦軸）を指標とする分類性能（小さいほど性能は高い）の$robx, robs$ 依存性．a）〜f）は，それぞれ訓練データ数$\sharp Tr$が400，800，1200，1600，2000，2400の場合に対応している．

結　　果

予備実験で得られたパラメタの最適値（表4）を用いて Graham モデルと Robinson-Fisher モデルの分類性能を比較した結果を図6に示す．図6aは予備実験における最適パラメタ値での分類性能をプロットしたもので，図6bは，検証用データとして予備実験で用いたものを除いた2,244件（Spam, Nonspam とも1,122件）を使い，各$\sharp Tr$に対して予備実験で求めた最適パラメタ値を使って分類実験を行った結果である．

図が示すように，両モデルとも訓練用データ数$\sharp Tr$の増加とともに分類性能が向上するが，$\sharp Tr$がある程度のサイズを超えるとほぼ横ばいとなる．また，$\sharp Tr$が小さい場合（$\sharp Tr \leq 800$）は Graham モデルよりも Robinson-Fisher モデルの方が高い性能を示しているが，$\sharp Tr$が大きくなると（$\sharp Tr \geq 1200$）両モデルの性能は予備実験と逆転している．

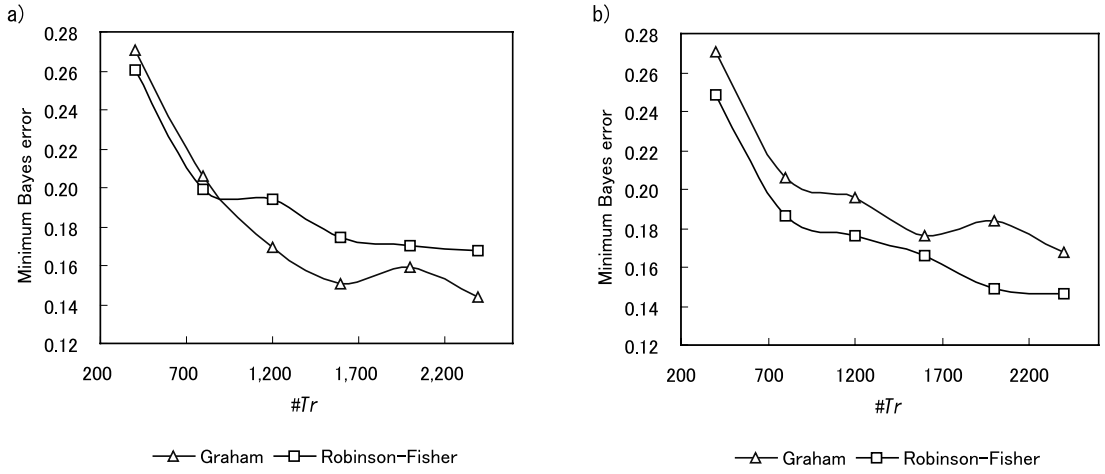

図 6　分類性能の比較

Grahama モデルと Robinson-Fisher モデルの分類性能（最小 Bayes エラー）の訓練用データ数$\sharp Tr$ 依存性を描いた図．a）は最適パラメタ値を求める予備実験でのもの，b）は検証用データから予備実験で用いたものを除いたデータで分類性能を検証したもの．検証用データは Spam，Nonspam とも時系列（論文発表順）に並べ替え，奇数番目を予備実験に，偶数番目を検証実験に用いた．各々，2244件，2245件である．このように検証用データを時系列で並べ替え，交互にサンプリングしたのはデータセットに「時間（発表時期）」の偏りが生じないようにするため．

本検証実験（図 6 b）で最も性能が高かったのは$\sharp Tr = 2{,}400$の Robinson–Fisher モデルで，最小 Bayes エラーが0.146（閾値=0.80125），Recall が91.3%，Precision が93.9%，F–measure は92.6%であった．ちなみに，予備実験（図 6 a）では$\sharp Tr$=2,400 の Graham モデルが最も性能が高く，最小 Bayes エラーが0.143（閾値=0.68625），Recall が93.3%，Precision が92.4%，F–measure は92.9%であった．

考　　察

迷惑メールの判定を目的として考案された BSF を医学文書の自動分類に適用したところ，Precision，Recall ともに90%以上の性能を示した．これは，BSF の本来の目的である Spam の判定（Precision:99.5%，Recall:96.6%）[25] には遠く及ばないものの，改善次第では実用的な域に達していると言える．そもそも BSF は，文書を構成する単語と文書が帰属すべきカテゴリとの間の関連の強さに基づいて文書を分類しているので，この結果は当然である．何故なら，BSF はその原理において，迷惑メールの判定という問題固有の知識をなんら利用していないからである．では，何故，英文 abstract の分類では迷惑メールの分類ほど高い性能を示さなかったのか．これは扱っている問題領域，即ち分類対象のコーパスの違いとしか考えられない．つまり，“liver cancer” をキーワードとする英文 abstract とそれ以外の abstract の違いは，Spam メールと通常のメールの違いほど顕著ではなかったということである．

次に，Robinson が Graham モデルに対して加えた改良についての考察を行う．まず，低頻度語の扱いについてであるが，訓練用データ数$\sharp Tr$ が少ない場合，分類対象の文書中に訓練用コーパスに現われなかった未知語が数多く出現する（図 7 参照）．その場合，ゼロ頻度問題が発生するが，$\sharp Tr$ が小さい領域で予備実験（図 6 a），検証実験（図 6 b）ともに Robinson-Fisher モデルの性能の方が優れていたのは，（6）式による低頻度語の扱いが Graham モデルよりも洗練されていたためと考えられる．

Robinson が行った 2 つ目の改良は，文書を構成する語に付与した TSP から文書の Spam 度を示す結合確率を計算する方法に Fisher の計算方法を採り入れた点である．Fisher の計算方法が Graham や他の多くの Naive Bayes 理論に基づく方法と異なる点は独立性の仮定を前提としないという点である．文書中の単語が互いに独立に出現するというありそうにない独立性の仮定をおきながら Naive Bayes モデルがそれなりの性能を示すのは，特徴選択によって特徴空間の次元削減を行っているからである．つまり，特徴選択によって文書中の単語が疎らに抽出されると，それだけ単語間の依存性が低くなると予想されるからである．実際，Graham モデルにおいても *Features* で与えられるパラメタで特徴選択を行っている．一方，Robinson-Fisher モデルは特徴選択を行わず，文書中に出現する全ての語の TSP を用いて結合確率を計算している．したがって，特徴選択を行わなくても Graham モデルに匹敵する性能を示した今回の実験結果は Robinson の改良が功を奏したことを裏付けている．

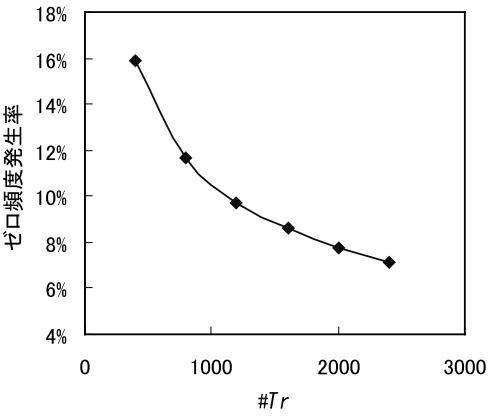

図 7　ゼロ頻度発生率の$\sharp Tr$ 依存性
検証用の abstract 中に訓練用コーパスに出現しない未知の単語が含まれている割合（ゼロ頻度発生率）が訓練用文書数$\sharp Tr$ によってどのように変化するかを描いた図．$\sharp Tr$ の増大に伴ってゼロ頻度発生率は減少する．

では，Robinson–Fisher モデルに特徴選択を行うとどうなるであろう．表 5 はχ^2 による特徴選択[9,21]を行った場合の Robinson–Fisher モデルの性能を調べたものである（使用したパラメタは$robx$=0.7, $robs$=2.6, $\sharp Tr$=2,400）．

選択された特徴数が16のとき，最小 Bayes エラーが0.113で最小となり，そのときの Recall が96.0%，Precision が92.9%，そして F-measure は94.4%で，Recall と F–measure は特徴選択を行わなかったときに比べて向上している．これは，予想に反して Robinson–Fisher モデルが特徴選択によって改善したことを示している．

そもそも Robinson が適用した Fisher の計算方法には，独立性の仮定が帰無仮説の中に取り込まれていた．そして Spam 確度を与える（7）式のS やH は帰無仮説を棄却するための尺度であった．Robinson のシナリオは，独立性の仮定を含む帰無仮説を棄却して対立仮説を支持する度合いを計算し，それを Spam 確度として利用するというものである．つまり，Robinson–Fisher モデルは独立性の仮定を前提としないというだけであって，特徴選択によって性能が改善しないと言っているわけではない．

さらに，表 5 において注目すべき点は，選択された特徴数がわずかに 1 語や 2 語の場合でもかなりの分類性能を示す点である．事実，2 語の場合において，特徴選択を行わなかった場合に匹敵する分類性能が得られた．しかし，ここで，次のような疑問が生じる．分類を試みた abstract が，一方は “liver cancer” という 2 語をキーワードとして検索したものであった．ということは，特徴選択にこの 2 語を選んだ場合，それだけで分類できることにならないか．

表 6 は ，“liver　cancer”，“liver”，“cancer”，“liver”＋“cancer” といった語（あるいは語の組合せ）を含む文書が各カテゴリの文書中にどの程度出現するかを調べたものである．また，表 7 は，特徴選択の語数が 1 ，2 , 16のそれぞれの場合に，選択された単語とその構成比率を調べたものである（$\sharp Tr$=2,400で行った）．これらの結果から，必ずしもキーワードに指定した 2 語だけで分類が行われたわけではないことが分かる．

とはいえ，PubMed の検索に用いたキーワードを分類カテゴリとし，その結果得られた abstract を分類するという本研究のシナリオはあまり適切とは言えなかったかもしれない．これに対して，Wang 等は，本研究と同様に PubMed から収集したコーパスに対する分類実験を，分類手法に kNN 法を用いて行っているが，彼らは分類対象の論文が収載されているジャーナルを分類カテゴリとして使用している[25]．テキスト分類手法の性能向上を目的とした研究であれば，そういったシナリオがより妥当と考えられるが，本研究は，退院時サマリの自動病名コーディングなどのように，ある特定の具体的なテーマ（たとえば病名など）に文書を自動分類することに関心があったため，このようなシナリオを選んだ．

ところで，表 2 を見ると “of” や “in” そして “and” などのように，どのようなカテゴリの文書にも出現すると思われる単語が，特に Nonspam コーパスにおいて高頻度で選択されている．これらの単語は，分類能力があまり期待できそうもない，いわゆる stop words と呼ばれる単語で，通常のテキスト分類では前処理で除去されることが多い．また，“patients” や “types” など，語尾変化（この場合は複数形）を伴う単語も前処理で語幹に変換することが多い．本

表5 χ^2による特徴選択

Features	BayesError	Recall	Precision	F-measure
1	0.180	86.9%	94.7%	90.6%
2	0.161	91.6%	92.2%	91.9%
4	0.158	93.1%	91.2%	92.2%
8	0.149	93.2%	92.0%	92.6%
16	0.113	96.0%	92.9%	94.4%
32	0.119	95.3%	93.0%	94.1%
64	0.129	95.8%	91.7%	93.7%
128	0.154	91.6%	92.9%	92.3%
256	0.158	90.9%	93.2%	92.0%
512	0.158	91.0%	93.1%	92.0%

Robinson-Fisherモデルに対してχ^2による特徴選択を行って分類実験を行った結果（$robx = 0.7, robs = 2.6, \sharp Tr = 2400$）．Robinsonの文献[15,16]では特徴選択の使用は言及されていない．今回の実験で，*Features*=16に対してRecallが96.0%，Precisionが92.9%，F-measureが94.4%となり，特徴選択を行わなかった場合のRobinson-Fisherモデルの最高パフォーマンスに比べて，Recall，F-measureはそれぞれ4.7ポイント，1.8ポイント向上し，Precisionは逆に1ポイント劣化しているが，F-measureでの総合評価では性能が向上している．

表6 キーワードおよびその構成単語を含む文書の割合

カテゴリ		"liver cancer"		"liver"		"cancer"		"liver"+"cancer"	
Spam	訓練用	210	9.4%	1408	62.7%	832	37.1%	584	26.0%
	検証用	196	8.7%	1371	61.1%	797	35.5%	556	24.8%
Nonspam	訓練用	1	0.0%	102	4.5%	142	6.3%	7	0.3%
	検証用	0	0.0%	96	4.3%	158	7.0%	10	0.4%

キーワードを構成する語およびその組合せを含む文書の数とコーパス中でのその割合を示す．なお，“liver”＋“cancer” は “liver” と “cancer” の両方の単語を含むという意味である．当然のことながら，Spamコーパスにはキーワードを構成する単語を含む文書が多く存在するが，それだけで90%を超えるRecallを達成することはできない．

研究ではGrahamに倣ってそういった前処理は行わなかったが，これらが分類性能にどのような影響を及ぼすかを検討してみる必要はある．

最後に，本研究では，訓練用データのSpamとNonspamの割合を等しいとして実験を行ったが，現実的な問題ではNonspamに比べてSpamが圧倒的に少ない場合が多いと考えるのが自然であろう．たとえば，病院情報システムに蓄積された膨大な退院時サマリからある病名（たとえば “liver cancer”）に関するサマリを分類するといった場合である．訓練用データの数が少ないとゼロ頻度問題が顕著になり，分類性能が低下することが予想される．今回はSpamとNonspamの比率による影響までは調べることができなかったが，実用化を考えた場合は検討する必要があろう．

ま と め

本研究では，迷惑メールのフィルタとして考案されたベイジアンスパムフィルタを医学文書の自動分類に適用し，その可用性を検討した．その結果，最大でRecallが96.0%，Precisionが92.9%という分類性能を得ることができ，ベイジアンスパムフィルタは医学文書にも適用できることを示した．また，特徴選択によって分類性能が向上することも確認した．とりわけ，わずか2語でも本来の分類性能をほとんど低下させることなく次元削減できることを確認できたのは大きな成果である．今回利用したコーパスの1文書あたりの平均語数は116語なので，これは約1/50の次元削減に相当する．Yang等は，10分の1の次元削減でも全く性能が落ちないこと，そして，100分の1の次元削減でわずかに性能低下が

表7　特徴選択によって選択された単語とその構成比率

カテゴリ	*Features* =1			*Features* =2			*Features* =16		
	単語	頻度	構成比率	単語	頻度	構成比率	単語	頻度	構成比率
Spam	liver	1344	59.9%	liver	1344	29.9%	of	2154	6.0%
	hepatocellular	486	21.7%	hepatocellular	1179	26.3%	in	2086	5.8%
	tumor	95	4.2%	carcinoma	581	12.9%	and	2028	5.7%
	of	73	3.3%	tumor	291	6.5%	with	1832	5.1%
	hepatic	61	2.7%	cancer	228	5.1%	to	1624	4.5%
	publication	60	2.7%	hepatic	211	4.7%	was	1480	4.1%
	cancer	56	2.5%	publication	202	4.5%	liver	1344	3.7%
	carcinoma	30	1.3%	of	163	3.6%	were	1270	3.5%
	hepatitis	27	1.2%	types	68	1.5%	carcinoma	1223	3.4%
	hcc	3	0.1%	with	63	1.4%	hepatocellular	1179	3.3%
	その他	9	0.4%	その他	158	3.5%	その他	19636	54.8%
	総異なり語数	16			24			743	
Nonspam	of	1109	49.4%	of	1628	36.3%	of	1885	5.5%
	patients	379	16.9%	with	743	16.5%	in	1763	5.2%
	publication	158	7.0%	patients	438	9.8%	and	1750	5.1%
	cancer	103	4.6%	in	257	5.7%	to	1552	4.6%
	in	71	3.2%	types	204	4.5%	the	1477	4.3%
	tumor	50	2.2%	publication	173	3.9%	with	1379	4.1%
	authors	47	2.1%	cancer	143	3.2%	for	1281	3.8%
	carcinoma	43	1.9%	and	94	2.1%	a	1223	3.6%
	liver	43	1.9%	authors	80	1.8%	was	973	2.9%
	types	42	1.9%	tumor	70	1.6%	university	959	2.8%
	その他	200	8.9%	その他	660	14.7%	その他	19770	58.1%
	総異なり語数	55			124			2717	

Robinson-Fisher モデルでχ^2による特徴選択を行った場合（$\sharp Tr$=2,400）に選択された単語を$Features = 1, 2, 16$の場合について頻度順にリストした．キーワードを構成する単語（“liver” や “cancer”）が必ずしも上位にランクされているとは限らないことがわかる．

見られることを示したが[21]，今回の結果もそれを裏付けている．

一方，医学文書を対象とした今回の試みでは，その性能指標は，迷惑メールのそれには遠く及ばなかったが，これは分類性能がコーパスに依存することを示唆している．今回は PubMed から “liver cancer” を入力して得た abstract と何もキーワードを入力しないで得た abstract をコーパスとして分類を行ったが，これは様々な内容の abstract から肝癌に関する abstract を抽出する問題に対応する．これ以外に，よく似た内容，たとえば肝癌と肝硬変に関する abstract の分類，また，全く異なった内容，たとえば肝癌と高血圧症に関する abstract の分類など，コーパスを変えてパラメタの最適値や分類性能がどう変わるかを調べることにより，モデルのコーパス依存性を明らかにする必要がある．

文　　献

1）Weiss SM，Indurkhya N，Zhang T and Damerau FJ：*Text Mining*. Springer，USA，52–53，2005.

2）岩橋佑佳，大江和彦：インシデント自由入力文からのインシデント種別の自動分類の試み．第24回医療情報学連合大会論文集，p10205，2005.

3）今井健，小野木雄三：格フレームを用いた放射線読影レポートの文型分類と所見抽出．第24回医療情報学連合大会論文集，p10125，2005.

4）Larkey LS and Croft WB：Combining Classifiers in Text Categorization. *Proceedings of SIGIR–96，19th*

ACM International Conference on Research and Development in Information Retrieval, 289–297, 1996.

5) Luciano RS de Lima, Alberto HF Laender and Berthier A Ribeiro-Neto：A hierarchical approach to the automatic categorization of medical documents. *Proceedings of the seventh international conference on Information and knowledge management*, 132–139, 1998.

6) Larkey LS and Croft WB：Automatic Assignment of ICD9 Codes To Discharge Summaries. `http://ciir.cs.umass.edu/pubfiles/coding.html`, 1995.

7) Lewis DD：Naive Bayes at Forty：The Independence Assumption in Information Retrieval. *Lecture Notes In Computer Science*, **1398**, 4–15, 1998.

8) Iwayama M and Tokunaga T：A probabilistic model for text categorization：Based on a single random variable with multiple values. *Proceedings of ANLP–94, 4 th Conference on Applied Natural Language Processing*, 162–167, 1994.

9) Sebastiani F：Machine Learning in Automated Text Categorization. *ACM Computing Surveys*. **34**(1), 1–47, 2002.

10) Sahami M, Dumais S, Heckerman D and Horvitz E：A bayesian approach to filtering junk E-mail. *Learning for Text Categorization*, Papers from the 1998.

11) Androutsopoulos I, et al. ：An evaluation of Naive Bayesian anti-spam filtering. *Proceedings of the workshop on Machine Learning in the New Information Age*, 9–17, 2000.

12) Androutsopoulos I, et al.：Learning to Filter Spam E-Mail：A Comparison of a Naive Bayesian and a Memory-Based Approach. *Proceedings of the workshop "Machine Learning and Textual Information Access", 4 th European Conference on Principles and Practice of Knowledge Discovery*, 1–13, 2000.

13) Graham P：A Plan for Spam. `http://www.paulgraham.com/paulgraham/spam.html`, 2002.

14) Graham P：Better Bayesian Filtering. `http://www.paulgraham.com/better.html`, 2003.

15) Robinson G：A Statistical Approach to the Spam Problem. `http://www.linuxjournal.com/node/6467/print`, 2003.

16) Robinson G：Spam Detection. `http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html`, 2003.

17) National Library of Medicine：*PubMed*. `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`.

18) Combining Probabilities. `http://www.mathpages.com/home/kmath267.htm`.

19) 北研二：確率的言語モデル．第3版，東京大学出版会，東京，46–56，2004.

20) Lewis DD：Feature selection and feature extraction for text categorization. *Proceedings of Speech and Natural Language Workshop*, 212–217, 1992.

21) Yang Y and Pedersen JO：A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420, 1997.

22) Little R and Folks J：Asymptotic Optimality of Fisher's Method of Combining Independent Tests. *Journal of the American Statistical Association*, **66**, 802–806, 1971.

23) Duda RD, Hart PE and Stork DG：*Pattern Classification*. 2 nd edition, New York, Wiley-Interscience, 20–83, 2000.

24) Weiss SM, Indurkhya N, Zhang T and Damerau FJ：*Text Mining*. Springer, USA, 77–81, 2005.

25) Louis G：Greg's Bogofilter Page. `http://www.bgl.nu/bogofilter/`, 2004.

26) Wang BB, Mckay RI, Abbass HA and Barlow N：A comparative study for domain ontology guided feature extraction. *Proceedings of the 26th Australian computer science conference*, 69–78, 2003.

（平成17年11月20日受理）

Trial on the Automatic Classification of Abstracts Using the Bayesian Spam Filter

Masaaki TANAKA

(Accepted Nov. 20, 2005)

Key words : bayesian spam filter, text classification, machine learning, feature selection, bayes theory

Abstract

In the healthcare domain, unstructured written medical records such as inpatient discharge summaries must be dealt with. In order to utilize them effectively, they need to be encoded to facilitate archiving and later retrieval. However, it is not an easy task to classify and encode them manually. An automated system, using text classification technology cultivated in the field of machine learning is needed.

In the present study, the author applied the Bayesian Spam Filter to the automatic classification of a medical text and examined its feasibility. A Bayesian Spam filter breaks down a text into its constituent words, assesses the degree of their relevance to classification categories from a corpus that has been developed beforehand, and then classifies the novel text into relevant categories.

For medical texts, the author collected abstracts from PubMed. As Bayesian Spam filters, models were utilized that were first devised by Graham and improved by Robinson. Some preliminarily experiments were performed to determine the optimal parameters, followed by an examination of classification performance.

The results show that, the model achieved 96.0% recall and 92.9% precision at the maximum, which is considered to be acceptable for practical use.

Correspondence to : Masaaki TANAKA
Department of Health Informatics
Faculty of Health and Welfare Services Administration
Kawasaki University of Medical Welfare
Kurashiki, 701-0193, Japan
E-Mail: mtanaka@mw.kawasaki-m.ac.jp
(Kawasaki Medical Welfare Journal Vol.15, No.2, 2006 539–552)