社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (19)
2007－BIO－11 (19)
2007／12／20

迷惑メールフィルタのためのベイジアンフィルタの改良

谷岡広樹[†1] 中川尚[†1] 丸山稔[†2]

本論文では，従来法の 1 つであるベイジアンフィルタを用いた spam メールフィルタの精度を改善する方法について提案する．これまでの学習型 spam メールフィルタとしては，ベイジアンフィルタがよく利用されており，一定の成果が得られている．しかしながら，ベイジアンフィルタを利用した方法においても，誤検出 (false positive) の低減や，さらなる精度向上といった課題が残されている．我々は，単語ごとの spam 確率の分布及びメールごとの spam 確率の分布状況を分析し，誤検出を押さえながらも，高い判定精度を実現する方法について提案し，その精度について，従来方式と比較して評価する．

An Improved Bayesian Filter for Spam Filtering

Hiroki Tanioka,[†1] Takashi Nakagawa[†1]
and Minoru Maruyama[†2]

We propose an improved baysian filter for spam mail detection. Bayesian filter was used on existing learning spam filters which achieved some positive results. Although we expect them to improve accuracy while keeping the false positive rate low. Therefore, it is based on a thorough review of distribution for each word and mail that our means of spam mail detection shows an impressively higher accuracy than ever.

1. はじめに

我々は，ビジネスシーンでの利用において，十分に耐えることを目的とした spam メールフィルタの開発を目指している．実用化に際して，ユーザのフィードバックから学習できること，学習の速度が高速であること，加えて，誤検出 (false positive) が少ないことが求められる．

文書フィルタリングに応用可能な分類器には，さまざまなアルゴリズムや学習モデルが提案されているが，学習速度が高速であるためには，フィルタリングのアルゴリズムを適切に選択する必要がある．

そこで我々は，従来法の中で最も学習コストが低いアルゴリズムの 1 つであるベイジアンフィルタを学習モデルとして選択し，さらなる精度向上を目指す．

1.1 ベイジアンフィルタ

spam メールフィルタのためのベイジアンフィルタについては，Paul Graham[2] の文献をきっかけに，広く普及することとなったが，PaulGraham 方式及びその改良版である Robinson[4] 方式ににおいても，実用面においては未だ十分とは言えず[10]，誤検出は低いままに，より高い spam 判定精度を実現することが望まれる．

そこで我々は，まずメールに含まれる単語の spam 確率 (尤度) 及びメールの spam 確率 (尤度) についての分布状況を分析し，spam 確率の計算に加味すべき単語の選択方法について再考する．その結果から，spam 度を計算するのに適切な単語の選択方法について検討し，誤検出を低く抑えたまま，spam 判定精度の向上を図る．以下に，提案方法の特徴について列挙する．なお，ham とは非 spam(non spam) を意味する．

- spam 確率及び ham 確率の高い順に単語を選択．
- spam 確率と ham 確率を比較し，spam 度を判定．

以上のような特徴により，本アルゴリズムは従来法の欠点を補い，誤検出を低く押さえたまま，さらなる精度向上が可能となる．

本稿ではまず，ベイジアンフィルタを用いた従来方式について概観し，それらの精度を調べる．その後，我々が提案する方式について詳しく説明する．

2. 従来方式

ベイジアンフィルタを用いた迷惑メールフィルタ

†1 株式会社ジャストシステム 本社，徳島市
JustSystems Corporation, Tokushima, Japan
†2 信州大学工学部情報工学科，長野市
Shinshu University, Nagano, Japan

には，PaulGraham 方式を初めとして，いくつかの方式が存在するが[11]，本稿では，NaiveBayes 方式，PaulGraham 方式，Robinson 方式，Robinson-Fisher 方式の 4 つの方式について，単語の spam 確率に対する分布状況や，メールの spam 確率に対する分布状況について概観し，それぞれの特徴を分析する．

2.1 NaiveBayes 方式

NaiveBayes 方式 (naive Bayesian classifier)[1),9)] は，各単語の特徴量 w_i が，クラス c が決まっているという条件で独立であるとき，次式を最大化するカテゴリ $\hat{c}$ を選べばよい．

$$\hat{c} = \arg\max_{c} P(c) \prod_{i=1}^{M} P(w_i|c) \tag{1}$$

具体的には，ある単語 w_i が spam として登場した回数を s_i，not spam として登場した回数を h_i とし，spam メールの総数が S，not spam メールの総数が H としたとき，spam 確率 $P(w_i)$ を，

$$P(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \tag{2}$$

とすると，単語の分布は，spam 確率が 1.0 付近に spam メールに含まれる単語が多く，spam 確率が 0 付近では not spam メールに含まれる単語が多く分布する．また，演算誤差を防ぐために対数値で spam 確率 $P(S)$，ham 確率 $P(H)$，及び spam 度 $P(M)$ を，

$$\begin{aligned} P(S) &= \sum_{i=1}^{n} \log(P(w_i)), \\ P(H) &= \sum_{i=1}^{n} \log(1 - P(w_i)), \\ P(M) &= \frac{P(S)}{P(S) + P(M)} \end{aligned} \tag{3}$$

とすると，$P(M)$ の分布から，閾値 t を 0.55 程度として，$P(M)$ がこの値を上回った場合に，メール M を spam メールであると判定すればよい．

2.2 PaulGraham 方式

PaulGraham 方式は，Paul Graham 氏によって提案された方式[2),3)] であり，ある単語 w_i の spam 確率 $P(w_i)$ を次式で表す．

$$P(w_i) = \frac{s_i/S}{a \cdot h_i/H + s_i/S} \tag{4}$$

このとき a は誤検出を低減することを狙ったバイアスであり，$a = 2$ とする．PaulGraham 方式では，spam 確率が 0.5 から離れている単語を順に n 語を選ぶため，単語を $|P(w_i) - 0.5|$ の大きい順にソートし，上位 $n(= 15)$ 語を抽出すると，その分布から，うまく抽出できていることがわかる．さらに，メールごとの spam 確率を，

$$P(M) = \frac{\prod_{i=1}^{n} p(w_i)}{\prod_{i=1}^{n} p(w_i) + \prod_{i=1}^{n} (1 - p(w_i))} \tag{5}$$

とすると，$P(M)$ の分布は，はっきりと分離している．ここで閾値 t を 0.9 とし，$P(M)$ がこの値を上回った場合，メール M を spam メールであると判定すればよい．なお，単語が未知語の場合は $p(w_i) = 0.4$ とする．

2.3 Robinson 方式

Robinson 方式は，Gary Robinson 氏が PaulGraham 方式を改良した方式[4)] である．まず，$s(w_i) = s_i/S$，$h(w_i) = h_i/H$ とし，単語 w_i ごとに spam メール確率 $P(w_i)$ をバイアスをかけずに計算し，尤度 $f(w_i)$ を次式のように計算する．

$$P(w_i) = \frac{s(w_i)}{h(w_i) + s(w_i)} \tag{6}$$

$$f(w_i) = \frac{s \cdot + n_i \cdot p(w_i)}{s + n_i} \tag{7}$$

このとき，x は未知語の spam 確率，s は x の予測に与える強さ (strength)，n_i は単語 w_i の出現回数 $(h_i + s_i)$ とし，$x = 0.5$，$s = 1$ とすると，尤度 $f(w_i)$ はである．また，spam 性 S を

$$\begin{aligned} P &= 1 - \{\prod_{i=1}^{n} (1 - f(w_i))\}^{\frac{1}{n}}, \\ Q &= 1 - \{\prod_{i=1}^{n} f(w_i)\}^{\frac{1}{n}}, \\ S &= (P - Q)/(P + Q) \end{aligned} \tag{8}$$

とすると，閾値 t を 0.55 とすると，単語 $w_1, \cdots, w_n$ を含むメール M は，spam メールであると判定できる．

2.4 Robinson-Fisher 方式

Robinson-Fisher 方式は，Thunder Bird[5)] や POPFile[6)] といったフリーウェアで採用されている方式として定評がある．本方式は Robinson 方式と同様に，spam 確率 $P(w_i)$ と尤度 $f(w_i)$ を求め，

$$\begin{aligned} H &= C^{-1}(-2 \ln \prod_{w_i}^{n} f(w_i), 2n), \\ S &= C^{-1}(-2 \ln \prod_{w_i}^{n} (1 - f(w_i)), 2n) \end{aligned}$$

のように non spam 性 H と spam 性 S を計算する．さらに指標 I を次式で計算すると，

$$I = \frac{1 + H - S}{2} \tag{9}$$

I は求まる．ここで，C^{-1} は逆 χ^2 関数とし，I が任

意の閾値を超えたとき spam メールであるとする．但し，計算コストの問題から，bsfilter[7] 等と同様に，

$$H' = 1 - C(-2\ln\prod_{w_i}^{n} f(w_i), 2n),$$
$$S' = 1 - C(-2\ln\prod_{w_i}^{n}(1 - f(w_i)), 2n),$$
$$I' = \frac{1 + H' - S'}{2} \qquad (10)$$

のように近似し，I' が閾値 t を上回った場合，単語 $w_1, \cdots, w_n$ を含むメール M は，spam メールであると判定する．

3. 提案方式

PaulGraham 方式では，spam 確率の高い単語を多く含んでいる場合でも，その確率を 0.5 からの差の絶対値で上回るような ham 確率の単語を同数以上混入することで，spam 判定されることを避けることが可能である．このような場合は，閾値による調整はまったく役に立たない．

Robinson 方式による場合も，意図的にメール内の単語の分布を ham メールに似せた spam メールに対して弱い．例を挙げると，「Via_g_r_a」「ばいアグ RA」「VI@GRA」などの単語を混入すると，未知語が多くなり，逆に，「Yahoo!」「Amazon」「楽天」といった，ham メールに含まれていそうな単語を大量に混入することで，判定を誤らせることができる．

3.1 Bipolar 方式

我々は，従来の方式には，単語の抽出方法あるいはベイズ推定の際の主観確率分布に改良の余地があるものと考え，Bipolar 方式と呼んでいる新しい方式を提案する．基本的なアイディアは，spam 確率の高い単語と，ham 確率の高い単語を，同数抽出するというものである．

この方式によると，spam メールらしさは，一部の spam 単語が含まれる場合に判断できる．逆に，ham メールらしさについても特定の ham 単語が含まれる場合に判断できる．

同様に，一部の単語を抽出する方式である PaulGraham 方式は，0.5 から遠い spam 確率の順に選ぶため，抽出された単語が，運悪く全て spam 単語または ham 単語となってしまう場合があったが，提案方法によると，ham と spam の双方から単語を選択するため，メール内の単語の分布状況も加味でき，spam 確率に対するメールの分布状況は，Robinson 方式に似た分布となり，閾値による制御にも効果があると考えられる．

Bipolar 方式では，まず，ある単語 w_i が spam として登場した回数を s_i，ham として登場した回数を h_i とし，spam メールの総数を S，ham メールの総数を H としたとき，spam 確率 $P(w_i)$ を次式で表す．

$$P(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \qquad (11)$$

なお，単語が未知語の場合，$p(w_i) = 0.4$ とする．ここで，メール M に含まれる n 語の単語を $P(w_i)$ の大きい順にソートし，spam 確率の高い単語 n_s 語による spam 確率 $P(S)$ と，ham 確率の高い単語 n_h 語による ham 確率 $P(H)$ を，$n'_h = n - n_h + 1$ から，

$$P(S) = \sum_{i=1}^{n_s} P(S|w_i) = \sum_{i=1}^{n_s} P(w_i),$$
$$P(H) = \sum_{i=n'_h}^{n} P(H|w_i) = \sum_{i=n'_h}^{n} (1 - P(w_i))$$

と表して，メール M の spam 度 $P(M)$ を，

$$P(M) = \frac{P(S)}{P(S) + P(H)} \qquad (12)$$

とすると，$P(M)$ が閾値 t を 0.55 程度として，その値を上回った場合に，メール M を spam メールであると判定できる．

Bipolar 方式では，spam 確率に対する単語の分布状況から，全体的に spam と ham をうまく分離しつつ，0.5 付近までなだらかに分布している．また，spam 度に対するメールの分布状況を見ると，広い範囲で分布しながらも，他のどのグラフよりも，spam メールと ham メールはうまく分離できているのがわかる．

3.2 従来方式の精度

従来方式と提案手法について，どの程度正しく spam メールを判別できるのかを，以下のメールデータ[*1]をコーパスとして用い，2-folds の交差検定 (cross validation) を 3 回試行したときの平均で計測した．なお，本文，ヘッダ情報，添付ファイルを含むすべてのメールデータを形態素解析し，その結果得られた単語の表記を，特徴データとした．

- spam メール: 1,671 通
- ham メール: 3,260 通

表 1 は，各方式を利用した場合の精度である．ベースラインと考える NaiveBayes 方式と比較すると，同等かそれ以上の精度であり，従来手法の中では，PaulGraham 方式が最も高い精度となった．提案手法は，他の方式よりも若干高い精度を示しており，PaulGraham

*1 2007 年 8 月の約 2 週間の間に，個人のメールアドレスで受信したメールのすべてであり，一部の spam メールを除いて，大半が日本語と英語のメールである．

表 1 交差検定の結果 (1)
Table 1 Cross validation(1)

Method	Threshold	Accuracy
NaiveBayes	0.5	0.952
PaulGraham	0.9	0.973
Robinson	0.55	0.968
Robinson-Fisher	0.55	0.952
Bipolar	0.55	**0.981**

There are the avarage accuracies of three results with 2-folds cross validation for each method.

表 2 交差検定の結果 (2)
Table 2 Cross validation(2)

Method	Threshold	FPR	Accuracy
NaiveBayes	0.600	0.006	0.950
PaulGraham	$*0.9\cdots$	0.008	0.952
Robinson	0.578	0.008	0.955
Robinson-Fisher	0.995	0.009	0.949
Bipolar	0.60	0.006	**0.981**

There are the avarage accuracies of three results with 2-folds cross validation for each method. FPR means the false positive rate. $*0.9\cdots = 0.999999999999999$.

方式の 0.973 に対しても，0.008 ポイント上回った．

次に，実用上の課題である誤検出率 (FPR; false positive rate) を低減するための設定を施した場合の精度について調べる．まず，前もって ham メールの判定精度が平均で 0.99 以上 (FPR: 0.01 以下) になるように各方式の閾値 (threshold) を調整した．

表 2 は，誤検出率の低減を優先する設定で cross validation を計測した結果である．すると，従来手法では，平均の判定精度は誤検出率を優先しない場合と比較して，やや下回った．最も高い精度であった Robinson 方式の場合も，若干ではあるが，0.013 ポイント下回った．一方，提案手法では，誤検出率を 0.99 以上に保ちながらも判定精度は 0.981 となっており，非常に高い精度となった．

4. おわりに

本稿では，ビジネスシーンでの利用において，十分に耐えることを目的とした spam メールフィルタの開発を目指し，従来法の 1 つであるベイジアンフィルタを用いた spam メールフィルタの精度を改善する方法について提案した．

提案方法によると，ベイジアンフィルタの特性の 1 つである学習及び判定の高速性を維持しつつ，誤判定率を一定以下に保ったままでも，従来方式以上の spam メール判定精度を実現することが可能であった．

本提案手法の特徴としては，

- PaulGraham 方式と同様に，特徴的な単語を単語の spam 確率を基準に選択することに加えて，ham 確率の視点からも同数の特徴的な単語を選択して，メール内の単語の分布状況を加味する．
- Robinson 方式及び Robinson-Fisher 方式の弱点と思われる，意図的に混入された雑音に対する脆弱性に対して，spam 性にも ham 性にも寄与しない，雑音となりうる単語を大幅に除去することで，頑健性を実現する．

といったことが挙げられる．

但し，本提案手法においては，メールがすべて自然文で記述されていることを前提に特徴抽出を行っており，HTML 本文のメールや画像メール，PDF メールといった spam メールに対してはなんら対策を施していないため，さらなる改良の余地がある．また，実際のビジネスシーンに於いては，日々進化する spam メール[8] に対してさらなる対応が必要となる．

参考文献

1) Duda, R. O., Hart, P. E. and Stork, D. G.: Pattern Classification Second Edition, pp.61–62 John Wiley & Sons Inc (2000).
2) Graham, P.: A Plan for Spam (2002). *http://paulgraham.com/spam.html*
3) Graham, P.: Better Bayesian Filtering (2003). *http://www.paulgraham.com/better.html*
4) Robinson, G.: A Statistical Approach to the Spam Problem, Linux Journal, Vol.107 (2003).
5) ThunderBird, Mozilla Foundation. *http://www.mozilla.com/en-US/thunderbird/*
6) Graham, C. J.: POPFile. *http://popfile.sourceforge.net/*
7) Nabeya, K.: bsfilter. *http://bsfilter.org/index-e.html*
8) The State of Spam, A Monthly Report, Symantec Corporation. *http://www.symantec.com/enterprise/security_response/weblog/security_response_blog/spam/*
9) 麻生英樹，津田宏治，村田昇：パターン認識と学習の統計学新しい概念と手法，統計科学のフロンティア 6, Chapter3.8, 岩波書店 (2003).
10) 特集 spam メールの現状と対策の動向，IPSJ Magazine Vol.46 No.7, pp.739–791 情報処理学会, July (2005).
11) 渡部綾太，愛甲健二：スパムメールの教科書，pp. 106–114, DATA HOUSE (2006).