機械学習によるスパムメールの特徴の決定木表現

杉井　学[1]、松野　浩嗣[2]
1 山口大学大学情報機構メディア基盤センター
2 山口大学大学院理工学研究科自然科学基盤系専攻

単語の出現頻度と語順解析を組み合わせた機械学習システムを用いて、スパムメールの特徴抽出を試みた。このシステムは、スパムメール群とそれ以外のメール群をそれぞれ正の学習例と負の学習例として与えると、学習例ごとに特徴的に出現する単語と文章中での出現パターンを解析して、二つの群を分ける規則を決定木として出力する。得られた決定木から、スパムメールの持つ特性について考察し、メールフィルターシステム構築の方策を検討した。

Decision Tree Representation of Spam Mail Features by Machine Learning

Manabu Sugii[1], Hiroshi Matsuno[2]
1 Media and Information Technology Center, Organization for Academic Information, Yamaguchi University
2 Graduate School of Science and Engineering, Yamaguchi University

We have tried to identify features of spam mails using machine learning system with a combination of word sequence analysis and the appearing rate of words. This machine learning system creates a decision tree as the classification rule from positive and negative examples by analyzing the distinctive features of words and its appearing patterns in a sentence. We discussed architecture and plan for constructing spam mail filter system on the basis of the decision tree functions getting from computational experiments of this research.

1．はじめに

迷惑電子メールであるスパムメールが増え続けている。現在では、さまざまな手口で宛先も無差別に送られるようになったスパムメールも、最初は特定のメールアドレスから通常のメールと同様に通常の経路で送信される、いわゆるダイレクトメールだった。From 行や Received 行に特定のサーバ名などを含んでいたため、比較的簡単に分類できたが[1]、次第にこのフィルター機能をかいくぐるスパムメールが増加し、現在では不正に SMTP サーバを利用したり、From 行を偽装したりしながら、電子メールのヘッダ情報に特定の文字列を含まないように細工をして送信されるようになっている[1]。スパムメール対策側は、正規表現などを用いて、スパムメールに特徴的に使わ

れる単語の存在でフィルターする仕組みを取り入れたが、特定の単語の有無だけで判断する方法は分類を誤ってしまう不具合を生んだ。また、人間には文章の判読が可能であるが、計算機には英単語の区切りがわからなくなるような処理として、スパムメール送信者が仕込ませた、英単語の中に英文のデリミタとなる空白文字を挿入するなどのやり方に、適切に対処することができなかった。この他、多くのフィルターシステムが開発され有効に機能しているが[2][3]、スパムメール送信者との「いたちごっこ」となってしまい、完全な対策方法は見出されていない。そこで我々は、次々に変化するスパムメールにも対応できるスパムメールフィルターシステムを開発するために、機械学習システムを用いてスパムメールの持つ特徴抽出を行い、最も有効なフィルター方法の検討とシステムの設計を試みた。

2．機械学習システム BONSAI

機械学習システム BONSAI は、概念学習の一つである確率的近似学習（Probably Approximately Correct Learning: PAC 学習）と呼ばれる学習パラダイムに基づき開発されたシステムである[4]。このシステムは、一次元の記号列データを対象にし、正の例と負の例の二つの学習例から、それらを区別する仮説を導き出すために、「インデックス化」と「決定木」を提示する機能を持つ。インデックス化は、学習例に含まれる要素をグルーピングし、要素数を減らすことで学習の効率化を行う。決定木は、インデックス化によって要素数を減少させた記号列を用い、正の例と負の例を最も高い確率で正しく分けることのできる規則を提示する。1992 年松野らは、BONSAI と Rough Reading という手法を用いて、生物系研究者が必要とする生物関連文献をより効率的に収集するシステムを開発した[5]。このシステムの特徴は、生物系研究者がキーワード検索でヒットしたいくつかの文献の abstract を読み、必要な文献と必要でない文献をある程度分類することで、これらを学習例にして次々に必要な関連文献を提示する機能にある。内部処理は、二つの学習例に分類された文献中の単語の出現頻度を算出し、その頻度を表す文字列で単語を置き換えたものを要素として BONSAI に学習させる方法である。

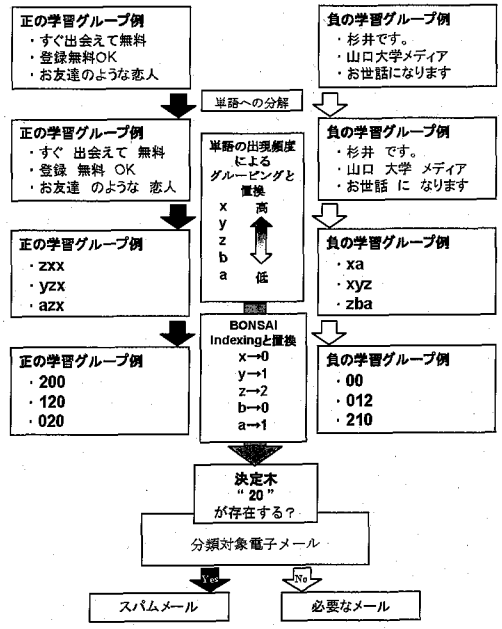

図1 BONSAIを用いたスパムメール分類の流れ

本実験では、BONSAI と Rough reading の手法を用い、図1に示すように、スパムメール群とそれ以外のメール群を学習例として BONSAI に入力して、二つのメール群を分類する特徴パターンの抽出を試みた。

3．スパムメール特性の決定木表現

2006 年 5 月 15 日～7 月 28 日に受信したメールの中で、送信日時の新しいものからスパムメールとそれ以外のメール 500 通を選定し、それぞれを正の例と負の例とした。同時に、2006

年7月29日から受信した1000通のメールを分類対象メール群とした。学習例としたメールは、メールヘッダおよびメール本文ともに、漢字→かな（ローマ字）変換プログラム「KAKASI(2.3.4-10)」を用いて分かち書きを行い、スペースで区切られた各単語に関して、それぞれの学習例内での出現数と二つの学習例での出現率の偏りを算出した。出現率の偏りは、正の例と負の例それぞれに特徴的で高頻度に出現する単語を特定するための指標とし、ある単語の学習例全体での出現数における、正の例での出現数の割合とした。

正の例での値が、0.8以上の単語をスパムメール群に高頻度に出現する単語として「x」で表し、0.8未満0.6以上の単語を「y」、0.6未満0.4以上を「z」、0.4未満0.2以上を「b」、0.2未満を「a」で表し、すべてのメールのすべての単語を「x, y, z, b, a」で置き換えた。分類対象の1000通のメールも同様に、すべての単語の置き換えを行い、学習例に一度も出現しなかった単語は「o」で置き換えた。

表1は、出現率の偏りにより分類された単語の一部を示しており、特に「x」と「y」にあたる単語は、直感的にスパムメールでよく見かける単語であることがわかる。また、「a」に分類された単語数が多いことは注目すべき点で、スパムメール以外のメールに特徴的に繰り返し用いられる単語が多数あることを示している。

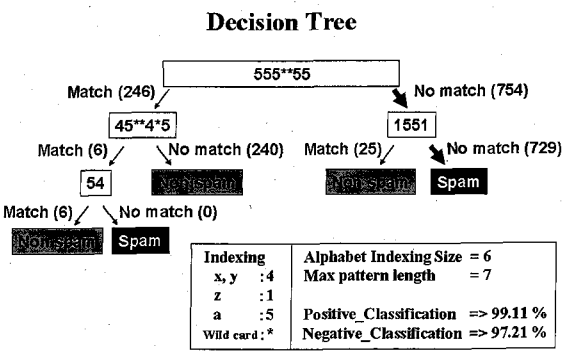

図2 BONSAIが出力したスパムメール判定の決定木

表1　出現率に偏りを持つ単語の例

出現率記号	抽出した単語
x	登録料
x	★
x	援助
	計 632
y	地域
y	好み
y	出張
	計 146
z	確実
z	理由
z	イベント
	計 252
b	ありがとうございました
b	個人情報
b	相談
	計 367
a	御中
a	XXXX@yamaguchi-u.ac.jp
a	お世話になります
	計 1710
※公開できない情報部分は「XXXX」と表記した。	

単語の出現率の偏りにより「x, y, z, b, a, o」で置き換えられたメール文章は、一次元の記号列データとしてBONSAIに入力した。BONSAIの学習パラメータとして、indexing sizeを6、pattern max lengthを7に設定した。Indexing sizeは、要素をインデックス化する時の最大分類数を、pattern max lengthは、決定木を構成するノードのパターンの最大長を定義している。図2は、BONSAIが、「x, y, z, b, a, o」をさらにインデックス化し、それぞれの学習例を最も効率よく分類することのできる規則を、決定木として出力した結果である。

BONSAIが決定木に明示的に用いた文字列は、「x, y, z, a」で、インデックス化して、「x」と「y」を「4」、「z」を「1」、「a」を「5」と定義した。残りの「o」と「b」はワイルドカードとして表現された「*」部分で許容される。つまり、決定木の根ノードであるパターン「555**55」は、「a」で始まり「a」で終わり、間の「*」部分は「x, y, z, b, a, o」すべての文字列が許容されることを意味している。言い換えれば、7つの単語で構成される文字列で、スパ

表２　BONSAIを用いたシステムと他のソフトのスパムメールの分類精度

学習例数	500			50			10		
ソフト名	BONSAI	Thunderbird	POPFile	BONSAI	Thunderbird	POPFile	BONSAI	Thunderbird	POPFile
分類対象メール数	1000								
Spam mail 数	726								
Non spam mail 数	274								
Spam mail 数	725	672	716	707	452	710	653	345	681
Spam mail 誤分類数	4	0	1	64	0	34	52	1	44
Non spam mail 数	270	274	273	210	274	240	222	273	230
Non spam mail 誤分類数	1	54	10	19	274	16	73	381	45
Spam mail 正解率	99.5	100	99.9	91,7	100	95.4	92.6	99.7	93.9
Spam mail 回収率	99.9	92.6	98.6	97.4	62.3	97.8	89.9	47.5	93.8
Non spam mail 正解率	99.6	83.5	96.5	91.7	50.0	93.8	75.3	41.7	83.6
Non spam mail 回収率	98.5	100	99.6	76.6	100	87.6	81.0	99.6	83.9
合計正解率	99.5	94.6	98.9	91.7	72.6	95.0	87.5	61.8	91.1

ムメール以外のメール群に高頻度に出現する単語が、1, 2, 3, 6, 7番目にあり、その他の位置には制限がないという文章パターンである。また、「x」と「y」をインデックス化して「4」と置き換えているので、決定木中のパターン内で「4」の部分は、「x」または「y」どちらが位置してもよいパターンとなる。

BONSAIの出力する決定木は、これらの文字列パターンが分類対象文字列に存在するか存在しないかで判定を行う。今回の結果では、ワールドカード表現を用いているために、決定木を構成するパターンは、わずか4つで、しかも高い精度で二つの学習例の分類に成功している。

{Positive_, Negative_} Classificationに示す値が、この決定木で学習例に用いたメールを再検定した結果で、正の例の分類正解率は 99.1%、負の例の分類正解率は 97.2%であった。また、パターン下括弧内の数字は、そのパターンの有無で分類されたメール数を示している。つまり、すべてのスパムメールはブロック矢印で示すルートをたどってスパムメールと判定され、決定木右半分の「555**55」と「1551 の」パターンがどちらも存在しない場合が、スパムメールという判定である。

得られた決定木を用いて、分類対象メール1000通を分類した結果が、表２の学習例数500、ソフト名 BONSAI の列の値である（「Thunderbird」、「POPFile」は、後述する比較実験に用いた別のメールフィルターソフト名）。分類対象メール1000通のうち、スパムメールは726通であったが、今回のシステムでは729通をスパムメールと判断し、そのうち725通が正しく分類されていた。つまり、4通のメールを誤ってスパムメール以外のメールと判断したことになる。一方、分類対象メール1000通のうち、スパムメール以外のメールは274通であったが、今回のシステムでは、スパムメール以外のメールとして分類した271通のメールのうち、270通を正しく分類し、1通のメールを誤ってスパムメールと判断した。メール全体の分類正解率は、99.5%である。

決定木の特徴として、根ノードのパターンがスパムメール以外のメール群に高頻度に出現する単語で構成されている点が挙げられる。このパターンの有無を調べることで、分類対象メールからほとんどのスパムメールの分類に成功している。データは示さないが、学習例を変えて同様の繰り返し実験を行っても、同じ傾向にあり、**スパムメール群に特徴的に出現するパターンで決定木を構成するのではなく、スパムメール以外のメールに存在するパターンを発**

見して、スパムメール以外のメールを分類している傾向が強い。

表３は、負の例として学習させたメールの中で、根ノードのパターン「555**55」が一致した単語列の一部を示している。予想に反してメールのヘッダ情報が多く、日常的にやり取りのある相手の名前や職名、所属名、メールアドレス、シグネチャ中の電話番号などが多かった。メール本文で該当するものの中には、「御中」「各位」「お世話になります」「よろしくお願いします」などの言葉が多く、スパムメールではあまり用いられないフレーズを利用して分類していた。

4. 他のメールフィルターソフトとの比較実験

次に、最近主流になりつつあるベイズ理論を応用したスパムメールフィルターシステムとの比較実験を行った。BONSAI を用いた今回のシステムも、過去のメールから学習するという点では、ベイズ理論を応用していると言えるが、単語の出現頻度と語順パターンを組み合わせた学習システムをエンジンにするスパムメールフィルターは過去に例がない。性能比較には、オープンソースの「POPFile(Ver. 0.22.4)」[6]と「Mozilla Thunderbird(Ver. 1.5.0.9)」[7]を用い、正の例と負の例を同数ずつ、それぞれ 500 通、50 通、10 通と変化させ、1000 通の分類対象メールを分類したときの分類精度と学習効率を検証した。表２はその結果を示し、正解率は「正しく分類したメール数／分類したメール総数×100」で表し、システムが分類したそれぞれのメール群がどれだけ正しく分類されているかを表す。回収率は、「正しく分類したメール数／分類するべきメールの総数×100」で表し、スパムメール 726 通、スパムメール以外 274 通で構成される分類対象メール 1000 通の中から、それぞれのメールをどれだけ回収することができたかという観点で表した値である。POPFile は「分類できないメール」という判定結果を出力するが、他のソフトとの比較のため、これらのメールは分類を誤ったとして計算した。

表３　負の学習例で「555**55」に当てはまる単語列

山口大学 情報 環境 部 情報 企画課 情報
tel 083-933-XXXX fax 083-933-XXXX email adress XXX@yamaguchi-uacjp
XXXXX@yamaguchi-u.ac.jp cc: XXX@yamaguchi-u.ac.jp XXXX@yamaguchi-u.ac.jp XXX@yamaguchi-u.ac.jp XXXXXX@yamaguchi-u.ac.jp XXXXXX@yamaguchi-u.ac.jp
いつもお 世話 になります 先 ほどご 依頼 ありました
メディア 基盤 センター 杉 井 先生 情報
----- 日本 バイオインフォマティクス 学会事務局 〒 108-8639 東京都
事務 担当 係 御中 いつもお 世話 になります
※公開できない情報部分は「XXXX」と表記した。

学習例が 500 通の場合、すべてのソフトが高い正解率と回収率を示した。しかし、学習例の数を減らすことによって、POPFile は、「分類できないメール」が増え、Thunderbird はスパムメールの回収率が下がり、スパムメールがもう一方のメール群に混入してしまう結果となった。BONSAI を用いたシステムは、高い分類性能を持つ POPFile に劣らない分類精度を有していることが明らかになった。

5. 考察

スパムメールの目的は「宣伝」である。内容はさまざまだが、あらゆる手段を使ってより多くの人に内容を読ませるための工夫を怠らない。しかし、電子メールとして送信する以上、最低限必要な電子メールの規則を逸脱することはできない。また、宣伝効果を高めるためには、他の電子メールよりも目に付くように工夫しなければ、スパムメール自身がスパムメールの中で埋もれてしまうことになる。スパムメールがこのような基本原則を破ることができない以上、機械学習やベイズ理論を応用したシステムを用いれば、目立つスパムメールであればあるほど、効率的に排除される方向へ導かれる

ことになる。

比較実験で用いたような、この原理に基づいて開発されてきたシステムは、非常に高い精度でスパムメールを排除することに成功している。しかし、本研究で開発したシステムは、この原理に基づきながらも、これまでとは発想を逆転させた新しいスパムメールフィルターの可能性を提示している。それは、スパムメールを排除するために、スパムメール群から特徴的なパターンを探し出して判断する方法と同様に、ユーザが必要とする電子メール群から特徴的なパターンを見出し、ネガティブモチーフとしてスパムメールを判断しても、効率的で正確な分類が可能なことである。

ユーザが必要とする電子メールよりもスパムメールの流通量の方が多くなってしまった今、さまざまな種類のスパムメールが考え出され、ユーザが通常やり取りするために必要な電子メールの種類を、はるかに超える多様なスパムメールを受信している。ネガティブモチーフとしてのスパムメール判定は、このようにスパムメールが多様化しても影響を受けない方法といえる。

6．おわりに

本実験システムでは、変化する電子メールの状況を学習することで、現状にあった規則でのスパムメールフィルター機能が期待できる。また、学習処理と分類処理を別サーバで実行することなどにより、正確かつ高速な処理が可能である。今後は、処理時間と性能のバランスを考慮した学習パラメータの最適化やスパムメール群に必要なメールを混入させないためのバイアス処理などを検討する必要がある。

謝辞

機械学習システム BONSAI のソースコードを提供いただいた九州大学大学院システム情報科学研究院 坂内英夫 助教授に深く感謝いたします。

また、本研究を行うに当たり、械学習システム BONSAI を動作させるため、山口大学大学情報機構メディア基盤センターの運用する、「計算機クラスターシステム」を利用いたしました。

参考文献

[1] 安東孝二, 世界の電子メールを spam 制御へ, 情報処理, Vol. 46, No. 7, pp. 741-746, 2005

[2] 山井成良・漣　一平・岡山聖彦・河野圭太・中村素典・丸山　伸・宮下卓也, SMTP セッションの強制切断による spam メール対策手法, 2006 FIT 情報科学技術フォーラム, 2006

[3] 西原　基夫, 「超高性能メールフィルターの研究開発と商品化」, 先端技術大賞フジサンケイビジネスアイ賞, 2004

[4] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., Knowledge acquisition from amino acid sequences by machine learning system BONSAI, Trans. Information Processing Society of Japan, 35 No. 10, 2009-2018, 1994.

[5] Shin-ichi, U., Kim, L., S., Tanaka, M., Nakazono, S., Matsuno, H., and Miyano, S., A Machine Learning Approach to Reducing the Work of Experts in Article Selection from Database: A Case Study for Regulatory Relations of *S. cerevisiae* Genes in MEDLINE, Genome Informatics Vol. 9, pp. 91-101, 1998.

[6] http://popfile.sourceforge.net/

[7] http://www.mozilla-japan.org/products/thunderbird/