F-007

学習機構を用いた迷惑メールの分類
A categorization of spam mails using machine learning

増田 明宏† 福本 淳一‡
Akihiro Masuda Jun'ichi Fukumoto

1. はじめに

送信されてくる e-mail の中には迷惑メールも多く存在する．各社で迷惑メール対策がなされてきたがその対策のうち NTTDoCoMo[4] では“i-mode のメール指定拒否・指定受信の設定件数を拡大 (2001/6)”や“「なりすまし迷惑メール」対策を実施 (2002/3)”などがある。迷惑メールの対策としてこのような e-mail のヘッダ情報を利用した分類は有効とは言えず，更に本文の内容パターンを利用した分類手法も有効ではなかった．本文内容学習することが出来れば，日々変わる形態に対応できる．

機械学習手法の一つである SVM(Support Vector Machine) は比較して汎化能力が高いとされ，画像処理や自然言語処理等の多くの分野で応用され，文書分類に関してもその有効性が報告されている.[1] SVM は Vapnik により提案された 2 値分類のための学習アルゴリズムであり，n 次元素性空間上の正 (+1) と負 (-1) から成る素性とラベルのペアの集合

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), x_i \in R^n, y_i \in \{-1, +1\} \quad (1)$$

に対して SVM は求める分離平面と並行で等間隔 2 つの平面間の距離 (マージン) を最大にするような分離平面を求める．

$$w \cdot x + b = 0, w \in R^n, b \in R \quad (2)$$

ここでマージンは $\|w\|/2$ であり，これを最大化するには以下の制限付き 2 次計画問題を解くことで得られる．

$$Minimize \frac{1}{2}\|w\|^2 \quad (3)$$

$$s.t. y_i(w \cdot x_i + b) \geq 1, \forall_i \quad (4)$$

この 2 次計画問題に対し Lagrange の未定乗数を用いて解くことにより最終的に，

$$f(x) = w \cdot x + b = \sum_i \lambda_i y_i x_i \cdot x + b \quad (5)$$

となる．ここで λ は (4) で示した Lagrange 乗数であり，テストデータは式 5 の符号にしたがって分類される．

本研究ではメールの分類に学習機構がどの程度有効かを実験的に検証した．また迷惑メールの特徴パターンを抽出し，学習と組み合わせることによって精度にどの様な変化が現れるか検証を行った．携帯電話に送られてきたメールは一般の文章より短いため，学習機構がどの程度精度を発揮できるかを調べた．また特徴パターンを抽出するために 2 章の述べる分析を行い，3 章にその抽出方法を示す．尚，ここでは学習機構に SVM を実装したパッケージツールの TinySVM[3] を利用し，パラメータにはメールを ChaSen[2] で分割した単語とその単語の出現回数を設定した．実験では，迷惑メールを細かく分類する実験を行った．一般のメールと迷惑メールを分類する実験は行っていない．

†立命館大学大学院理工学研究科
(a_masuda@nlp.is.ritsumei.ac.jp)
‡立命館大学情報理工学部 (fukumoto@media.ritsumei.ac.jp)
〒 525-8577 滋賀県草津市野路東 1 丁目 1-1

2. 迷惑メールの分析

分析に用いた迷惑メールは，2002 年 4 月 14 日から 2002 年 5 月 13 日までに NTTDoCoMo の携帯電話に送られてきた 1121 件を利用する．

2.1 統計的データの分析

携帯電話に送られてくるメールなので，平均文字数が 123 文字となっており，本文を短くまとめ，本文で使われている単語に特徴が現れていると考えられる．またサイトのアクセス手段である URL と e-mail アドレスの平均出現回数はそれぞれ，2.1 回と 0.1 回になっており，その中でも URL の非出現回数は分析データ 1121 件中の 58 件のみで，99%以上ものメールに見られることが分かった．

2.2 分類する種別ごとの特徴

本研究の目的である迷惑メールを分類するのに次の 2 種を分野として扱い，分類の対象としていくものとする．

- 出会い系メール … メール友達の募集サイト，“出会い系サイト”やアダルト情報を含むサイト案内のメール
- その他のメール … お金を貸す目的，融資をする会社からのサイト案内や着メロ，待ち受け画像の案内といった出会い系以外のメール

分析データ 1121 件を手で分類した結果，963 件が出会い系メールとなり，158 件がその他のメールとなった．出会い系メール，その他のメール共にアクセス手段である URL が出現し，また出会い系メールには言葉の一部を伏せた伏せ字が存在する．

そこで分類したメール中にこのような伏せ字がどの位の割合で出現するか調べた結果，出会い系メールでは平均 0.06 回出現するがその他のメールでは出現が見られなかった．またその他のメールに分類される融資の案内に関するメールでは融資やキャッシングの利用額を示した数字やお金に関する単語が多く含まれていた．これらの特徴を各分野の固有のものとして利用し，学習にうまく反映させるために次の 3 節に示す手法を用いる．

3. タグ付け (前処理)

分析結果から SVM で機械学習する前に迷惑メールの特徴である “URL”, “電話番号” 等にタグ処理を行い，それらの特徴ある表現を抽出する．このタグ処理を行わな

い場合 ChaSen により形態素解析を行うと半角英数字は“未知語”としてばらばらに区切られてしまい，また伏せ字は記号と文字に区切られてしまう．タグ付けを行った後に，そのタグをつけられた部分を一つの単語として認識するためにタグを以下のように置きかえる．

URL → <mwkURL>
電話番号 → <mwkTEL>
伏せ字 → <mwkFSJ>
数字 → <SUJI>

URL のタグ付けに関しては，“http”や“www”などの文字列が含まれている場合は URL と判断し，電話番号に関しては数字が 10 桁または 11 桁なら数字と判断している．また伏せ字に関しては，記号の前後にカタカナがあれば伏せ字であると判断している．伏せ字の判定において，記号の前後で文字又は文字列があるとすると何処まで伏せ字と設定するのかが難しいため現段階ではこれが最善であるとした．図 1 にタグ処理を行った後のメールを示す．URL や伏せ字といった特徴ある表現が抽出されているのがわかる．

なんて言うんですかね，人妻って凄いですよ。
夫との<mwkFSJ> ですんごいテク持ってるのに
夫は相手してしてくれないのか超欲求不満。
絡みつくような舌使いとか腰のグラインド!溜まりません!
<mwkURL>
取り合えずメール H でイカせればアポ取りやすいよ!がんばってね!
<mwkURL>
↓↓↓↓↓↓↓↓男性お試しポイント付き！女性安心の無料&メアド非公開！
<mwkURL>
今後 DM 不要の方はコチラからお願い致します。
<mwkURL>
このメールは広告です

図 1: タグ処理後のメール

4. 実験・評価

実験で用いたデータは次の 3 種である．

1. 分析に用いた 1121 件 (2002/4/14〜2002/5/13).
2. 評価用データ 1137 件 (2002/5/16〜2002/6/20).
3. 新しいデータ 4065 件 (2003/8/15〜2003/9/26).

評価のために全てのデータは 2 つの分野に手分けしておく．また 1,2 と 3 は集めた期間が異なり，1 と 2 はそれぞれ 2 種の分野のメールを含むが，3 のデータは全て出会い系メールである．

4.1 実験方法

1. 分析データ 1121 件をトレーニングデータとして評価データ 1137 件を分類する実験
2. 分析データ 1121 件をトレーニングデータとして，新しいデータ 4065 件を分類する実験
3. 分析データ 1121 件と評価用データ 1137 件あわせて 2258 件をトレーニングデータとし新しいデータ 4065 件を分類する実験

各実験では前処理を行った場合と行わない場合の比較をし，計 6 種の実験を行った．

4.2 評価方法

評価方法は *Accuracy*, *Precision*, *Recall* を用いて評価する．*Accuracy* とは SVM が出した答えと，User が出した答えがどの程度一致しているか，*Precision* とは SVM が正解であると返したメールがどの程度正解しているか，*Recall* とは正解とされるデータのうちどの程度正解しているかを示す値である．

$$Accuracy = \frac{A+D}{A+B+C+D} \quad (6)$$

$$Precision = \frac{A}{A+B} \quad (7)$$

$$Recall = \frac{A}{A+C} \quad (8)$$

A:システムがその分野のメールであると返したうち正しい数
B:システムがその分野のメールであると返したうち間違った数
C:システムがその分野のメールでないと返したうち間違った数
D:システムがその分野のメールでないと返したうち正しい数

4.3 評価実験の結果と評価

システムが分類した後の種別をそれぞれ“出会い系メール”，“その他のメール”とする．

まず迷惑メール 1137 件の分類実験について，システムが分類した正解データの分布を表 1 に示す．“Total”にシステムが出したその分野の正解データの件数を示し，それぞれシステムが返した結果の分野ごとの件数を表に示している．また，その評価を表 3 に示す．表 3 で示された

表 1: メールの分類結果 (前処理を行わない場合)

	出会いメール	その他のメール	Total
出会いメール	927 件	34 件	961 件
その他のメール	51 件	125 件	176 件

表 2: メールの分類結果 (前処理を行った場合)

	出会い系メール	その他のメール	Total
出会い系メール	964 件	37 件	1001 件
その他のメール	14 件	122 件	136 件

表 3: 分類評価 1

		deai		*other*	
	Acc.	*Pre.*	*Rec.*	*Pre.*	*Rec.*
前処理なし	92.5%	96.4%	94.7%	71.0%	78.6%
前処理あり	95.5%	96.3%	98.5%	89.7%	76.7%

ように前処理を行わない分類でも高い精度が示されており，また *Precision* や *Recall* などは前処理を行った場合多少，値の低下が見られるが全体的評価である *Accuracy* の値は上昇し前処理の有効性が示されている．

次に 4065 件のデータの分類実験に関して，まず，トレーニングデータ 1121 件で同じように実験した結果は前処理を行わない場合は 4065 件中 3288 件正しく分類され，前処理を行った場合は 4065 件中 3314 件正しく分類され，分類の評価値は表 4 に示される．表 4 から新しい

表 4: 分類評価 2(training data:1121 件)

		deai	
	Acc.	*Pre.*	*Rec.*
前処理なし	80.8%	100%	80.8%
前処理あり	81.5%	100%	81.5%

迷惑メールでも *Accuracy* は若干落ちるはものの 80%以上の精度を保ち, また新しい迷惑メールに対しても, 前処理を行うことにより精度が上昇することが実証できた.

次に, トレーニングデータを増やし 2258 件とした場合, 前処理を行わない場合は 4065 件中 3564 件正しく分類出来, 前処理を行うと 4065 件中 3801 件となった. 分類の評価値は表 5 に示される. 表 5 のトレーニングデータを増やした結果を見てみると, トレーニングデータが少なかったときよりも *Accuracy* が上昇し前処理を行った場合では 93%を超える分類精度が得られた.

表 5: 分類評価 3(training data:2258 件)

		deai	
	Acc.	*Pre.*	*Rec.*
前処理なし	87.6%	100%	87.6%
前処理あり	93.5%	100%	93.5%

5. 考察

本節では本研究で評価に用いた 2 つのデータセットに対する, タグ処理を行った場合と行わない場合の迷惑メールの分類の評価に対する考察を行う.

5.1 迷惑メール 1137 件の分類

まず前処理を行わず, 学習機構のみで分類をしたときでも表 3 から精度は 92%を超えており高い性能で分類できていることが分かった. そして前処理を行うことにより *Accuracy* の上昇を計ることが出来た. 一方, 出会い系の *Precision* やその他のメールの *Recall* が低下してしまった理由には, 前処理により出会い系特有の表現は多く抽出できたが, その他のメール特有のものが少なかったため, URL や e-mail アドレスを出会い系メールに見られる特徴であると誤って分類され, それらの値が低下したものと思われ, その為に各分野固有の表現を更に多く抽出する必要がある.

5.2 新しい迷惑メール 4065 件の分類

まず学習させる training データ数を変化させない場合, 表 4 に示されるように, 精度より 10%ほど下がったが 80%以上の精度が保て, training データも新しいものを使用すれば, およそ 90%の精度が出せるのではないかと考えられる. また前処理により若干ではあるが, このデータに対しても精度の上昇を計れた.

次に学習させるデータ数を増やした場合, トレーニングデータを増やし前処理を行うと精度の上昇が顕著に見られ前処理の有効性が示されている.

以上の結果から SVM を用いた迷惑メールの分類は高い分類精度があり, またタグ付けという前処理を行うことによりさらにその精度が上昇し, 本研究での hybrid の分類というものが有効であることが実証できた.

6. おわりに

実験から, 学習機構のみによる分類でも 90%を超える精度があったが, 従来のパターンを用いた分類を組み合わせることにより, *Accuracy* はおよそ 93%まで上昇した. また新しい迷惑メールに対しても同じく精度は 90%以上を保つことができた. 今後, 精度の更なる向上を計るためには“SVM に与えるデータの前処理の方法”として本研究で行った URL や伏せ字に対してタグ処理の他に方法が他にはないか考えるべきである. 次に“学習とパターンの組み合わせの割合”があり, 精度を更にあげるためにはどの割合の組み合わせが最も精度が高いか調べる必要がある.

参考文献

[1] 高村大也, 松本裕治.(2003) “SVM を用いた文書分類と構成的帰納学習法” 情報処理学会論文誌:データベース vol.44 No.SIG 3(TOD17) pp.1-10

[2] ChaSen. http://www.chasen.aist-nara.ac.jp/

[3] TinySVM. http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

[4] NTTDoCoMo. http://www.nttdocomo.co.jp/