反復度を用いた文字列の特徴選択によるスパム分類

尾上　徹[†]　　岡部　正幸[‡]　　梅村　恭司[†]　　阿部　洋丈[†]

[†] 豊橋技術科学大学情報工学系　　[‡] 豊橋技術科学大学情報メディア基盤センター

Feature Selection of String for Spam Classification based on Adaptation

Toru Onoe[†]　Masayuki Okabe[‡]　Kyoji Umemura[†]　Hirotake Abe[†]

[†] Information and Computer Science, Toyohashi University of Technology,

[‡] Information and Media Center, Toyohashi University of Technology

1. はじめに

日々大量に送られてくるメールからスパムメールを人の手で１つ１つ削除するのは面倒である．人によっては，あまりにも送られてくるメールの量が多く，手作業でスパムメールを削除するのが現実的ではないという問題がある．そのような場合，機械学習を用いてメールのスパム分類を行うのが１つの解決手段であると考えられる．スパム分類はメールの内容から Spam と Ham に分類する狭義のテキスト分類である．

テキスト分類ではテキストの特徴を表す文字または部分文字列の集合（特徴集合）として区切り文字で区切った単語を用いることが多いが，Spam では単語が改ざんされるケースもあるという問題がある．

一方，文字列を特徴とすると単語を用いる方法に比べて，連語情報を損なわず，区切り文字のないデータの処理に前処理を必要としないという利点がある．そして，単語が改ざんされても処理できる．しかし，部分文字列すべてを特徴集合とすると，その集合の大きさは単語数に比べ大変大きな数となり，文書の規模によっては機械学習に要する計算時間が現実的に計算できないほどに膨大になる可能性があるため，特徴集合を絞り込む必要がある．この特徴集合は分類の精度を左右するため，より文書の分類に役立つようなものを特徴集合として選ぶ必要がある．先行研究として，条件付確率を用いて類似した部分文字列をまとめることで特徴集合を改善する報告がある(Zhang et al., 2006 [1])．

本研究では，部分文字列を特徴集合とし，特徴集合の選択には反復度という統計量を用いる．そして，これにより得た部分文字列を用いてサポートベクターマシンによりスパム分類した結果を，条件付き確率を用いて特徴選択を行った場合の結果,単語を特徴集合とした場合の結果両方と比較を行う．

スパム分類では，Ham メールを Spam メールとみなす分類失敗は,その逆の失敗よりも一般に望ましくない.そこで，F 値による分類結果の比較だけでなく，フォールスポジティブによる評価も行った．

反復度を用いた実験として，平田らのロイターコーパス，20news コーパスという一般的な文書に対する文字列の特徴抽出(平田ら, 2007 [2])があるが，これは文字列を特徴集合とすることの必然性を示すことはできなかった．そこで，本研究では特殊な文書構造をもつスパムコーパスを特徴抽出の対象とし，条件付確率を用いる方法と単語を特徴集合とする方法に比べて結果が改善すること，すなわち，スパム分類において反復度を用いた特徴選択が有効であることを示すとともに，文字列を特徴集合とすることの利点について確認する．

スパム分類では，Ham メールを Spam メールとみなす分類失敗は,その逆の失敗よりも一般に望ましくない.そこで，F 値による分類結果の比較だけでなく，フォールスポジティブによる評価も行った．本稿の報告で用いる手法は，文献[4]に添ったものであるが，評価については，スパム分類に焦点をあて，フォールスポジティブの値を追加した．

2. Zhang らの特徴選択方法

まず，tf 値, df 値, tfidf 値を次のように定義する．

- tf(t,d): 文字列 t が文書 d に出現する頻度
- df(t,D): コーパス D 中で文字列 t が出現する文書数
- tfidf(t,d): tfidf(t,d) = tf(t,d)・log(|D|/df(t,D))（|D|はコーパス D の文書数を表す）

この選択法は，出現分布が同一または類似している文字列をまとめることで特徴選択を行う．ここで，出現分布が同一または類似している文字列とは，ある文字列のコーパス中におけるすべての出現場所をリストにしたとき，そのリストが別の文字列が持つ出現場所リストと等しいまたは類似している文字列のことを指す．

出現分布が同一な文字列は tf 値と df 値について同じ値を持つため，これをひとつの特徴としてまとめる．ただし，文字列をまとめるとき，最も文字列長が短い文字列のみを代表文字列として選択する．出現場所のリストが類似しているとき,そのような文字列の tf 値, df 値に大きな違いは生じない．これらの文字列をひとつの特徴にまとめても分類結果にあまり影響を与えることなく，特徴集合を減らすことができると考えられる．出現場所の類似性の判定の基準，すなわち類似した文字列を取り除くための条件を Zhang らは以下のように定めた．

[1] コーパス中である文字列の次に現れる文字の種類が b 種類未満の文字列を特徴集合から取り除く．
[2] ある文字列（S1）が現れたとき，この文字列から始まる文字列（S2）が出現する条件付確率 P(S2| S1)が p 以上であるならば，後者の文字列を特徴集合から取り除く．
[3] ある文字列（S3）が現れたとき，この文字列で終わる文字列（S4）が出現する条件付確率 P(S4| S3)が q 以上であるならば，特徴集合から後者の文字列を取り除く．

また，コーパス中で出現頻度が極端に多い文字列，少ない

文字列は分類に寄与しないと考え，最小頻度 l 未満の文字列，最大頻度 h 以上の文字列は特徴集合から除く．

これらの処理を行うには 5 つのパラメータ l, h, b, p および q を決定する必要があるが，これらは学習文書における交差検定法によって推定する．Zhang らは以上の処理を suffix tree を用いて効率的に行う方法を提案し，英語，中国語およびギリシャ語のコーパスを用いてテキスト分類の実験を行い，これまでに提案されてきた主な文字列ベースのテキスト分類手法よりも優れた性能を示したと報告している．

3. 提案手法

われわれの提案手法は，Zhang らの提案手法の類似した文字列をまとめる条件[1], [2], [3]を，反復度を利用した次の条件に変えたものである．ただし，パラメータ a は交差検定によって定める．

- 反復度が a 未満の部分文字列を特徴集合から取り除く

この提案手法は出現分布が同じものと反復度により出現分布が類似しているとみなされる文字列をまとめ，ある文書に偏って出現する文字列を特徴として抽出するものであるといえる．

反復度 adapt(t,D)は，文字列 t が出現した文書のうち，2 回以上繰り返し出現している文書の割合を示す統計量で以下のように定義される．

adapt(t,D) = df2(t,D)/df(t,D)

ここで，df2(t,D)はコーパス D 中で，文字列 t が 2 回以上出現する文書の数を表す．

反復度は語の意味の境界を越えたときに大きく減少する性質を持ち，キーワードの自動抽出(Takeda, Umemura, 2002 [3])に使用された．

4. 交差検定

本研究ではパラメータの決定を Zhang らと同様に交差検定法を用いて行った．交差検定法は既知のデータ（学習データ）から未知のデータ（テストデータ）に対するモデルのパラメータを推定する方法であり，本研究では 4 分割交差検定法を用いた．手順を以下に示す．

まず，学習データを 4 つのブロックに分割する．次に推定するパラメータを設定する．ブロックの 1 つをテスト文書，残り 3 つを学習文書としてテキスト分類するという操作をそれぞれのブロックに対して行う．この計 4 回のテキスト分類をパラメータの値を変えて繰り返し，もっともよい分類性能を示したものを最適なパラメータとみなす．

ただし，パラメータの設定は，ある範囲についてある一定間隔で行うため，刻み幅によっては必ずしも既知のデータから推測される最適のパラメータを得ることはできない．また，既知のデータから推測される最適なパラメータを得たとしても，それがテストデータに対してもっともよい分類結果を与えるとは限らないということに注意すべきである．

5. コーパス

本研究では，スパムコーパスとして TREC 2006 Spam Corpus[1]を用いた．このコーパスはスパム分類のために作られたもので，コーパスにはヘッダ情報が含まれ，一般的な英語文章とは異なる構造をしているという特徴がある．これを用いた実験で分類精度が向上すれば，実際のスパム分類においても分類精度が向上すると考えられる．

6. 分類実験

6. 1 実験方法

学習データとして，コーパスから Spam, Ham それぞれ 100 個をランダムに選ぶ．分類対象(テストデータ)として，Spam, Ham それぞれ 200 個をランダムに選ぶ．このような手順で 20 個の文書セットを作成して，この文書セットそれぞれに対して以下の分類実験を行う．

まず，学習データから次の 3 つの特徴集合を構成する．

- 反復度を用いて特徴選択した文字列からなる特徴集合(AS)
- 条件付確率を用いて特徴選択した文字列からなる特徴集合(CS)
- 単語からなる特徴集合(WS)

ただし，単語からなる特徴集合(WS)は，学習データの単語すべての集合のうち出現頻度が l 以上かつ h 以下の単語から構成される．l, h は交差検定により設定する．

各手法のパラメータは文書セットを変えた場合に設定し直す．ただし，条件付確率による手法のパラメータは先行研究(Zhang et al., 2006)と同様の値を用いることとし，反復度による手法のパラメータ l, h による部分文字列の除去操作は条件付確率の手法における除去操作と同様であるため l, h は条件付確率による手法のものと同じ値を用いることとする．

テキストの学習分類器には，線形カーネルを利用した SVM を用いることとし，本実験では LibSVM[2]を用いた．なお，このときソフトマージン法を用いて，パラメータはディフォルトのものを用いて分類した．

そして，文書セット 20 個に対するそれぞれの特徴集合を用いた場合の分類結果の比較を行う．

分類結果の評価は Spam と Ham の平均 F 値と，Ham メールを Spam メールと誤分類した数のそれぞれにより行うこととした．F 値は次式で表される．

$$適合率 = \frac{トピックに属すると分類した文書の正解文書数}{トピックに属すると分類した文書数}$$

$$再現率 = \frac{トピックに属すると分類した文書の正解文書数}{コーパス中の正解文書数}$$

$$F値 = \frac{2 \times 適合率 \times 再現率}{適合率 + 再現率}$$

[1] http://trec.nist.gov/data/spam.html

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

6.2　実験結果

表 1 に 20 個の文書セットに対する，各文書セットを用いた場合の分類結果を示す. この結果から, 平均 F 値について，AS を用いると F 値が CS を用いるより 2.63%改善され，WS を用いるよりも 1.00%だけ改善することが分かる．また，表の結果よりすべての文書セットにおいて反復度(AS)を用いると条件付確率(CS)を用いた場合の結果を上回ることが分かる．このことから，反復度を用いて選択した文字列を特徴集合とするのは条件付確率を用いる方法と比較して有効であると考えられる．AS を用いた場合と WS を用いた場合の F 値を比較すると，20 回のうち，AS が WS よりも良くなったのが 16 回，悪くなったのが 3 回，同値となったのが 1 回であった．この結果について符号検定を行い，両手法の F 値の間に有意な差があるかどうかを考える．まず，帰無仮説 H_0 と対立仮説 H_1 を以下に示すように定める．

H_0 :　反復度を用いる方法と単語を用いる方法の F 値の間に差がない

H_1 :　反復度を用いる方法は単語を用いる方法の F 値の間に差がある

両手法の結果が同じ値となった場合，単語を用いる方法の方が優れていると見なすと，両手法の F 値の分布が等しいという仮定の下で単語を用いる方法の結果が 20 回の内 4 回反復度よりも良くなる確率は，

$$\frac{1}{2^{20}}\left({}_{20}C_4+{}_{20}C_3+{}_{20}C_2+{}_{20}C_1+1\right)=0.0059089\cdots<0.01$$

となるため，有意水準 1%で帰無仮説は棄却され，対立仮説が採択される．このことから，反復度を用いる方法は単語を用いる手法よりも F 値において有意な差があると考えることができる．

文献[2], [4]では一般のテキスト分類として評価をしているが，スパムに焦点を当てた場合 Ham が Spam と判断されるケース（フォールスポジティブ）が重要である．

フォールスポジティブによる評価として**表 2** に Spam と誤分類された Ham の数を示す．表より AS を用いると，20 個すべての文書セットについて誤分類の数が CS を用いるよりも少なくなり，20 個中 16 個について WS を用いるよりも誤分類の数が少なくなることが分かる．F 値の場合と同様にして符号検定を行うと，AS を用いる場合と WS を用いる場合の結果の間に危険率 1%で有意差があることがわかる．よって，反復度を用いる方法は単語を用いる手法よりもフォールスポジティブによる評価（Spam と誤分類された Ham の数）において有意な差があると考えることができる．

7.　考察

結果から，反復度(AS)を用いたほうが条件付確率(CS)を用いるよりも分類結果を改善することが分かった．本章では，両特徴集合を比較し，その傾向と具体的にどのような文字列が分類を改善したかについて考える．

まず，文書セット 1 つを 6 章で述べたようにして作成し，文書セットから特徴集合 AS と CS を先に述べたようにして取り出しそれぞれを用いて分類を行う．その分類結果（F 値）は AS が 95.25%，CS が 90.25%となった．次に，AS と CS を比較し，AS にあって CS にない文字列集合(IS)を新たに作る．**表 3** にこれらの集合の大きさを示す．この文字列集合 IS を AS から取り除いて（AS と CS の積集合 AS∩CS で）分類を行うと，AS を用いるのに比べて 7%，CS を用いるのに比べて 2%だけ低い結果となった．この結果から，文字列集合(IS)の中に CS を用いた場合に比べて分類結果を改善する原因となった文字列が含まれていると考えられる．

次に反復度の特徴集合の内，サポートベクトルとして用いられた文字列の集合を取り出し，それぞれの重みを計算する．そして，重みが大きいものほど分類に寄与していると考え，その上位 50 個を取り出す．この 50 個の部分文字列の集合を見ると，message_id という文字列の一部と推測される部分文字列が 12 個見つかった（ただし文字 “_” は空白を示す）．

また, この 12 個の部分文字列すべては CS には含まれていないことが分かった. この 12 個の部分文字列を AS から取り除いて分類を行ったところ，F 値は 93.25%となり 2.00%低下する結果となった．このことから，これらの文字列は分類に役立っていることが分かる．

SV（反復度のサポートベクトル）全体から message_id の部分文字列を探したところ，26 個見つかり，そのうち 10 個は CS にも含まれ, 残りの 16 個は AS にのみ含まれることが分かった．ここで，この CS にも含まれる 10 個の部分文字列を AS（SV）から取り除き分類を行った場合，F 値（分類結果）は変化しないことを確認した．これより，AS にのみ含まれる 16 個の部分文字列は CS にも含まれる 10 個の部分文字列をカバーすると言える．

この message_id という文字列がコーパスの Spam, Ham メールのうちどれぐらい含まれるのかを調べたところ，Spam メールの約 81.9%，Ham メールの約 99.9%にこれが

表 1　各手法の平均 F 値

文書セット	反復度 (AS)	条件付き確率 (CS)	単語 (WS)
1	93.75%	90.25%	94.00%
2	93.00%	90.75%	92.25%
3	95.50%	92.75%	93.25%
4	92.75%	91.00%	92.75%
5	92.25%	91.25%	91.25%
6	93.50%	89.75%	92.75%
7	95.75%	92.50%	92.75%
8	93.00%	92.75%	91.50%
9	91.00%	90.50%	90.50%
10	94.00%	91.25%	91.00%
11	93.00%	85.75%	91.50%
12	93.50%	92.00%	93.25%
13	92.50%	91.50%	91.75%
14	87.75%	85.50%	88.00%
15	91.50%	89.50%	90.00%
16	95.00%	88.25%	93.00%
17	93.75%	90.00%	93.50%
18	92.50%	91.75%	91.75%
19	93.50%	93.25%	93.75%
20	93.00%	87.75%	92.00%
平均	93.03%	90.40%	92.03%

表 2　フォールスポジティブによる評価

文書セット	Spam と誤分類された Ham の数		
	反復度 (AS)	条件付き確率 (CS)	単語 (WS)
1	10	18	17
2	11	21	22
3	13	15	17
4	17	17	11
5	13	24	21
6	13	25	15
7	11	20	13
8	10	13	15
9	28	31	31
10	12	18	19
11	15	38	25
12	17	19	21
13	20	24	26
14	36	47	39
15	20	22	15
16	12	27	20
17	14	16	10
18	13	18	8
19	18	26	25
20	14	18	24

表 3　特徴集合の大きさ

文字列集合	記号	文字列数
反復度の特徴集合	AS	1988
条件付確率の特徴集合	CS	687
AS, CS の差集合(AS-CS)	IS	1633
反復度のサポートベクトル	SV	1976

含まれていることがわかった．よってこれが含まれていないとほぼ Spam と断定できる文字列であるということがわかり，これは分類に有用であるということは直感的に理解できる．

では，なぜ AS のみに含まれる 16 個の部分文字列が CS の 10 個の部分文字列をカバーしたのか，message_id の部分文字列集合を比較することで考察する．**表 4** に AS と CS それぞれに含まれる message_id の部分文字列を示す．**表 4** を見ると，条件付確率の手法を用いたほうは一見しただけでは何の部分文字列かわからないほど短い文字列である．これは別の意図しない文字列に対しても分類結果が影響を受けやすい，

つまり文字列 message_id を意図して me を選択しても member や meat などの別の文字の部分文字列と解釈される可能性があるということである．それに対して反復度で抽出した部分文字列は短い文字列もあるがかなり長い文字列も捉えており，age_i など間に空白が挟まった形も捉えているため別の意図しない文字列に影響されにくい部分文字列であるといえる．このような何を指しているのか分かり易いある程度長い部分文字列と，間に空白を挟んだ単語と単語を結ぶような形の部分文字列が分類結果を改善していると考えられる．

表 4　反復度と条件付確率の特徴集合の比較

反復度	条件付確率	反復度	条件付確率
ag	ag	id	id
age	age	id_	
age_		me	me
age_i		mes	
d_		mess	
e_i		message	
es	es	message_	
ess	ess	sa	sa
essa		sag	
g	g	sage	
ge		sage_	
ge_	ge_	ss	ss
ge_i		ssa	

8.　まとめ

スパム分類において，特徴抽出に反復度を用いると，条件付確率を用いる場合に比べて F 値を 90.40%から 93.03%に 2.63%だけ改善し，単語を特徴集合とする場合に比べて 1.00%だけ改善することが分かった．また，フォールスポジティブによる評価として Spam と誤分類された Ham の数を比較すると，反復度を用いると，20 個すべての文書セットについて条件付確率を用いる場合よりも誤分類の数が少なくなり，20 個中 16 個の文書セットについて単語を用いる場合よりも誤分類の数が少なくなることが分かった．これらの結果に有意差があることは危険率 1%の符号検定により確認した．これらのことから，反復度を用いた特徴抽出はスパム分類に有効であるといえる．

この結果の要因として，反復度を用いて抽出される部分文字列に，条件付き確率を用いる手法で抽出される部分文字列に比べて，特徴としてふさわしくないような文字列と部分一致しづらいような文字列が含まれていることが考えられる．単語を特徴集合とする場合よりも結果が改善したのは，間に空白を挟む単語と単語をつなぐ文字列を特徴とし，連語情報を用いることができたためであると推測できる．

参考文献

[1]D. Zhang, and W. S. Lee (2006). “Extracting Key-Substing-Group Features for Text Classification.” Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, pp.474-483.

[2] 平田，岡部，梅村 (2007). “文字列を特徴量とし反復度を用いたテキスト分類.” 情報処理学会研究報告，pp.121-126

[3] Takeda and Umemura (2002). “Selecting Indexing Strings using Adaptation.” Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.11-15

[4] 尾上，平田，岡部，梅村,"文字列を特徴量とし反復度を用いたテキスト分類", 自然言語処理，Vol. 17, No.1 掲載予定