文字列を特徴量とし反復度を用いたテキスト分類

尾上 徹[†]・平田 勝大[†]・岡部 正幸^{††}・梅村 恭司[†]

テキスト分類における特徴抽出とは、分類結果を改善するためにテキストの特徴たる単語または文字列を取捨選択する手続きである。ドキュメントセットのすべての部分文字列の数は、通常は非常に膨大であるため、部分文字列を特徴として使用するとき、この操作は重要な役割を果たす。

本研究では、部分文字列の特徴抽出の方法に焦点を当て、反復度と呼ばれる統計量を使って特徴抽出する方法を提案する。反復度は、高確率でドキュメントに二度以上出現する文字列は文書のキーワードであるはずだという仮定に基づく統計量であり、この反復度の性質は、テキスト分類にも有効であると考ええる。実験では、Zhang ら (Zhang et al. 2006) によって提案された、条件付確率を用いることで分布が類似した文字列をまとめるという手法（以下、条件付確率の方法と記す）と我々の提案する手法の比較を行う。結果の評価には適合率と再現率に基づく F 値を用いることとした。ニュース記事とスパムメールの分類実験の結果、我々の提案する反復度を用いた特徴抽出法を用いると、条件付確率の方法を用いるのに比べて、ニュース記事の分類では分類結果を平均 79.65% から平均 83.39% に改善し、スパムメールの分類では分類結果を平均 90.23% から平均 93.15% に改善した。提案手法である反復度を用いる特徴抽出法は Zhang らの提案する条件付確率を用いる特徴抽出法に比べて、ニュース分類記事の分類では平均 3.74%、スパムメールの分類では平均 2.93% だけ結果を改善しており、その両方の実験において結果に有意差があることを確認した。また、反復度を用いる特徴抽出方法を用いると、単語を特徴集合とする方法を用いる場合と比べて、ニュース記事の分類では分類の結果を平均 83.88% から平均 83.39% と平均 0.49% 低下させることとなったものの、スパムメールの分類では分類の結果を平均 92.11% から平均 93.15% と平均 1.04% 改善した。ニュース記事の分類においては反復度を用いる特徴抽出方法と単語を特徴集合とする方法に有意差は本実験では認められず、スパムメールの分類の結果においては有意差があることを確認した。

この結果が得られた要因について考察すると、条件付確率の方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を抽出する傾向にあることが分かった。これは不特定多数の文字列の一部として出現しやすいことを意味しており、文書の特徴になりえないような文字列がこれを含んでいたとき、分類結果がその文字列の影響を受けることを意味する。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ものの、長い文字列や間に空白が挟まった単語をつなぐ部分文字列も捉えているため、特定のものをさす文字列の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わかりやす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。

キーワード：テキストマイニング、テキスト分類、機械学習、特徴選択

[†] 豊橋技術科学大学情報工学系, Information and Computer Science, Toyohashi University of Technology

^{††} 豊橋技術科学大学情報メディア基盤センター, Information and Media Center, Toyohashi University of Technology

Extracting String Features with Adaptation for Text Classification

TORU ONOE[†], KATSUHIRO HIRATA[†], MASAYUKI OKABE^{††} and KYOJI UMEMURA[†]

Feature selection for text classification is a procedure that categorizes words or strings to improve the classification performance. This operation is especially important when we use substrings as a feature because the number of substrings in a given data set is usually quite large.

In this paper, we focus on the substring feature selection technique and describe a method that uses a statistic score called “adaptation” as a measure for the selection. Adaptation works on the assumption that strings appearing more than twice in a document have a high probability of being keywords; we expect this feature to be an effective tool for text classification. We compared our method with a state-of-the-art method proposed by Zhang et al. that identifies a substring feature set by removing redundant substrings that are similar in terms of statistical distribution. We evaluate the classification results by F-measure that is a harmonic mean of precision and recall. An experiment on news classification demonstrated that our method outperformed Zhang’s by 3.74% (it improves Zhang’s result from 79.65% to 83.39%) on average based on the classification results. In addition, an experiment on spam classification demonstrated that our method outperformed Zhang’s by 2.93% (it improves the Zhang’s result from 90.23% to 93.15%). We verified existence of significant difference between the results in each experiment.

An experiment on news classification shows that our method is worse than a method of using word for feature by 0.49% (although there is no significant difference) on average based on the classification results. In addition, an experiment on spam classification demonstrated that our method outperformed the word method by 1.04% (our method improves its result from 92.11% to 93.15%). We verified that there is a significant difference between the results in spam classification experiment.

Zhang’s method tends to extract substrings that are so short it is difficult to understand the original phrases from which they are extracted. This degrades classification performance because such a substring can be a part of many different words, some or most of which are unrelated to the original substring. Our method, on the other hand, avoids this pitfall because it selects substrings containing a limited number of original words. Selecting substrings in such a manner is the key advantage of our method.

Key Words: *Text Mining, Text Classification, Machine Learning, Feature Selection*

1 はじめに

テキスト分類学習は、スパムメールの除去、Web コンテンツのフィルタリング、ニュースの自動分類など様々な応用分野をもつ重要な技術である。

一般の分類学習と同様に、テキスト分類学習においても特徴集合の選択は学習性能を決定する重要な要素である。通常、英文であればスペースによって区切られた語、日本語文であれば形態素解析によって分割された語を特徴として用いることが多いが、このような方法では二語が接続していることの情報が欠落するので、分類に役立つ熟語・複合語などの情報を取りこぼす可能性が高い。このため、この情報についてはあきらめるか辞書から得るかしなければならない。さらにこの情報を利用する場合は言語モデルの利用や string kernel などの特殊なカーネルを利用することにより学習アルゴリズム側で接続を考慮するといった対応を行う必要が生じる。

一方、特徴選択の方法として文を文字列と見なし、全ての部分文字列を考慮することで、接続を特徴選択の際に取り込もうとするアプローチがある。このアプローチでは、熟語・複合語を取り込むための辞書や接続を考慮した学習アルゴリズムを使用する必要がないという利点があるが、部分文字列数のオーダーはテキストデータの全文字数の 2 乗のオーダーという非常に大きな値になってしまうため、取捨選択してサイズを縮小する必要がある。部分文字列を考慮した特徴選択の代表的なものに、Zhang らが提案した方法がある (Zhang et al. 2006)。彼らは suffix tree を利用して、出現分布が同一または類似している文字列を一つにまとめることによって特徴集合のサイズを縮小する方法を提案した。そして、この選択方法による特徴集合とサポートベクターマシンを利用したテキスト分類実験において、接続や文字列を考慮した他の代表的な方法よりも高い性能を与えることを示した。

これに対して、本研究ではすべての部分文字列を考慮する点は同じものの、反復度と呼ばれる統計量を利用して、Zhang らの方法と異なる部分文字列の選択方法を提案する。反復度は文書内で繰り返される文字列は文書内容を特徴づける上で重要な語であるという仮定に基づく統計量であり、これまでキーワード抽出などに利用されている (Takeda and Umemura 2002)。Zhang らの方法は部分文字列の出現分布が類似したものを一つにまとめるという操作のみを行い、選択した部分文字列の文書内容を特徴づける上での重要性は学習アルゴリズムによって決めるというアプローチであるといえるが、反復度では特徴選択時にも部分文字列の重要性を考慮しており、分類に寄与しない特徴を予め取り除く効果が期待できる。本研究では、この反復度を用いた部分文字列からの特徴選択の効果を、ニュース記事を用いた分類実験、スパムメールのデータセットを用いた分類実験において検証する。そして、ニュース記事の分類実験では、提案手法である反復度を用いた特徴抽出方法が Zhang らの特徴抽出方法よりも優れた結果を示し、単語を特徴集合とする方法との間には有意差が認められなかったことを報告する。一方、スパムメールの分類実験において提案手法は Zhang らの方法、単語を特徴集合とする方法よりも優れ

た結果を示し、有意差が確認されたことを報告する。

以下、2章では Zhang らの方法について詳しく説明する。また3章では本研究で利用する反復度と交差検定によるパラメータの設定方法について説明する。4章では実験方法と実験結果について述べ、5章でその結果について考察し、6章でまとめを行う。

2 Zhang らの特徴選択方法

ここではテキスト分類における特徴選択の先行研究として、Zhang らが提案した方法について詳しく説明する。

テキスト分類に用いる機械学習アルゴリズムの多くは、学習の際のデータ表現に文書ベクトルを用いるが、通常このベクトルの値として、以下の tf 値, df 値を元に計算した tfidf と呼ばれる値が使用される。

- $tf(t, d)$: 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
- $tfidf(t, d)$: $tfidf(t, d) = tf(t, d) \cdot \log(|D|/df(t, D))$

($|D|$ はコーパス D の文書数を表す)

Zhang らは、膨大な部分文字列を削り込むために出現分布が同一または類似している文字列をまとめ、上で述べた値に違いのあるものをなるべく特徴として選択するというアプローチを採用した。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、学習において区別する必要はなく、ひとつの特徴としてまとめてしまう。具体的な手続きとしては、出現場所のリストが等しい文字列のうち、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。また、出現場所のリストが厳密に等しくなくても類似していれば、そのような文字列の tf 値, df 値にも大きな違いは生じないため、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余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。

ただし出現場所の類似性の判定には基準が必要なので、Zhang らは類似した文字列を取り除くための条件を以下のようにした。

- (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 種類未満の文字列は特徴集合から取り除く。
- (2) ある文字列 (S_1) が現れたとき、この文字列から始まる文字列 (S_2) が出現する条件付確率 $P(S_2 | S_1)$ 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 (S_3) が現れたとき、この文字列で終わる文字列 (S_4) が出現する条件付確率

$P(S_4|S_3)$ 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 1 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.

以上の処理により, 特徴集合の大きさを, 全部分文字列を特徴集合とした場合に比べ大幅に小さくすることができる. 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ったところ, 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, 例えば, 言語モデルを利用した生成アプローチ (Peng 2004) や string kernel を利用した識別アプローチ (Lodhi 2001) などの方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

3.1 反復度による特徴量抽出

本研究では, 出現分布が同一または類似した文字列をまとめることに加え, ある文書に偏って出現する文字列をテキスト分類の重要な特徴として残すことを考え, Zhang らが用いた手法の条件 (1), (2), (3) の代わりに反復度と呼ばれる統計量を用いることによって文字列を選択する方法を提案する.

反復度 $\text{adapt}(t, D)$ は, 語 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.

$$\text{adapt}(t, D) = \frac{\text{df}_2(t, D)}{\text{df}(t, D)}$$

ここで, $\text{df}_2(t, D)$ はコーパス D 中の文書で, 文字列 t が 2 回以上出現する文書数を表す.

表 1 は, ある英文中における反復度の変化の様子を示したものである. ただし, 「_」は空白を表す. この例では, 「natural_gas_」に 1 文字追加して「natural_gas_s」となったときに反復度が急激に減少している様子が示されている. 表 1 に見られるように, 反復度はある境界を境にそれまではほぼ一定だった値が急激に減少する統計量であり, $\text{df} / |D|$ で計算される出現確率とは異なり, 意味的に一塊の語の境界で減少することが多いことから, キーワードの自動抽出 (Takeda and Umemura 2002) などに利用されている. 表 1 の例においても, 「natural_gas_」で語が区切られることは, その意味を考えると妥当であるといえる. 提案する方法では, 出現分布が等しい文字列をその代表文字列だけにまとめることは Zhang らと同じであるが, 2 章で説明した (1), (2), (3) の条件の代わりに, 以下の条件を用いる.

表 1 語の境界における反復度の変化

文字列	df	df ₂	出現確率	反復度
n	9565	9492	0.996	0.992
na	5564	4074	0.579	0.732
nat	2820	1201	0.294	0.426
natu	2820	1201	0.294	0.426
natur	206	63	0.021	0.306
natura	172	55	0.018	0.32
natural	170	55	0.018	0.324
natural_	162	53	0.017	0.327
natural_g	162	53	0.017	0.327
natural_ga	162	53	0.017	0.327
natural_gas	116	37	0.012	0.319
natural_gas_	115	37	0.012	0.322
natural_gas_s	12	1	0.001	0.083

- 反復度が最小反復度 a 未満の文字列は特徴集合から取り除く

Zhang らの (1), (2), (3) の条件を用いていないため, 出現分布が類似していてもひとつにまとめず別の特徴として扱う. ただし, 出現分布が等しい文字列はひとつの特徴にまとめるため, 表 1 のような 1 文字ずつ増加させたような文字列が必ず選ばれるわけではなく, 単語中の語幹や連語単位の文字列などを特徴集合に含めることができる (平田 他 2007). 表 1 の例では, 「nat」と「natu」の 2 つ, 「natural_」, 「natural_g」, 「natural_ga」の 3 つは出現場所のリストが等しく, 統計量も同じなので, それぞれ「nat」と「natural_」だけの特徴として選択する. また, 「natural_gas」のような連語も特徴として選択される.

3.2 交差検定法によるパラメータの設定

提案手法では, 最小頻度 l , 最大頻度 h といったパラメータに加え, 最小反復度 a を決定する必要があるが, 本研究では Zhang らと同じく交差検定法によって推定する. 交差検定法とは, 未知のデータに対するモデルのパラメータを推定する方法のひとつである. 本研究で用いる 4 分割交差検定法は, 学習文書を 4 つのブロックに分割し, それぞれのブロックをテスト文書とし, テストに使用していない残りのブロックを学習文書に使用する. この 4 回のテキスト分類において最も分類性能が良くなるパラメータを最適なパラメータとして推定する. 以上のように, パラメータの推定のテスト文書に学習文書とは別の文書を使用し, 元の学習文書のすべての文書を順にテスト文書として使用することで, 過学習を防ぐパラメータを決定することができ, 学習における汎化性能の向上が期待される.

4 実験

実験は、ニュースのトピック分類と Spam 分類の 2 種類のタスクについて行い、それぞれのタスクについて、関連研究の手法の分類結果と反復度による特徴集合を用いた場合の分類結果の比較を行う。

4.1 ニュースのトピック分類実験

この実験には、Reuters-21578¹と 20newsgroups²の 2 つの英語コーパスを使用し、各文書には前処理として、アルファベット以外の文字を空白に変換し、2 文字以上空白が続く場合は空白 1 文字に変換するという処理を行う。

テキスト分類に適用可能な分類学習器は数多くあるが、本実験では特徴選択方法の比較を行うので、適切と考えられる分類学習器について実験を行った。分類学習器には Zhang らの先行研究と同じく線形カーネルを利用した SVM を用いる。線形カーネルを利用した SVM の識別関数は次式のように表される。

$$f(\mathbf{x}) = \sum_{j=1}^d w_j x_j + b$$

ここで、 $\mathbf{x}$ は識別対象となる文書ベクトル、 x_j は $\mathbf{x}$ の要素 j の値、 w_j は重みベクトル $\mathbf{w}$ の要素 j の値、 d は $\mathbf{x}$ の要素数、 b はバイアス項である。 $\mathbf{w}$ と b は学習によって決定される。

我々の提案手法では、 $\mathbf{x}$ の要素集合として、反復度によって特徴選択した文字列集合を用いる。また、比較対象として、2 章の先行研究において説明した条件付確率によって特徴選択した文字列集合を $\mathbf{x}$ の要素集合とした方法をベースラインに用いる。また、単語を特徴集合とする方法との比較として、最小出現頻度 1 と最大出現頻度 h で選択した単語集合を $\mathbf{x}$ の要素集合とした方法とも比較する。 x_j の値は 3 手法とも tfidf によって計算された値を用いる。tfidf の計算式は 2 章で記述した式を用いる。

実際の SVM の学習には、SVM ツールのひとつである SVMlight³を使用し、すべてデフォルトのパラメータで学習を行う。複数トピックの分類に対しては、ターゲットとするトピックに属する文書を正例、そのトピックに属さない文書を負例とした 2 クラスによる学習を各トピックについて行う。結果の評価には次の 3 つの尺度を利用する。

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

¹ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

² <http://people.csail.mit.edu/jrennie/20Newsgroups/>

³ <http://svmlight.joachims.org/>

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

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

4.1.1 Reuters-21578

Reuters-21578 は、英語のテキスト分類の標準的なコーパスであり、先行研究でも用いられている。このコーパスのうち、「ModApte」学習・テストセットを使用し、本文のうち TITLE タグと BODY タグのついた文書を使用する。さらに、Reuter-21578 の文書に含まれるトピックのうちの文書数の多い上位 10 トピックについてテキスト分類を行う。ただし、各文書は複数のトピックに属することがあり、その場合、正例として使用される文書は負例として同時に使用しないこととした。分類は、ひとつの文書が各トピックに対して属するか属さないかを SVM を用いて判定することによって行う。学習には、学習用文書セットの全 9,603 文書を使用し、テストには学習文書とは異なるテストセットの全 3,299 文書を使用する。表 2 に学習セットおよびテストセットにおける上位 10 トピックの正例の文書数を示す。後述の実験において、正例の文書数が少ないときに提案方法が優位であることを述べるために、このデータを示した。このコーパスに対して、次の 3 つの特徴集合を用いた場合のテキスト分類を行い、その結果を比較する。

- 反復度：提案手法である、反復度を用い文字列を選択した特徴集合
- 条件付確率：Zhang ら (Zhang and Lee 2006) が提案した条件付確率を用いて文字列を選択した特徴集合。ベースラインとして用いる。
- 単語：スペースを区切りとした単語からなる特徴集合

また、各特徴集合のパラメータ設定と選択された特徴数を以下に示す。

- 反復度：l=80, h=8000, a=0.3（3 章参照）として 7,099 文字列が選択された。

表 2 学習セットの文書数

トピック	学習セット	テストセット
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	369	117
interest	347	131
wheat	212	71
ship	197	89
corn	181	56

- 条件付確率： $l=80$, $h=8000$, $b=8$, $p=0.8$, $q=0.8$ （2章参照）として 8,438 文字列が選択された。
- 単語： $l=10$, $h=8000$ として 6,581 単語が選択された。

条件付確率による手法のパラメータは先行研究と同様のパラメータを使用し、反復度のパラメータおよび単語の特徴選択のパラメータは学習用文書セットでの 4 分割交差検定法において F 値の平均が最も良くなる値を調べ決定している。ただし、文字列を特徴集合とする場合は、空白で始まる文字列は特徴集合からは除く。これは、英語の単語が空白で区切られているためである。以上の実験結果を図 1, 図 2, 図 3, 表 3 に示す。

図 1 および図 2 に注目して文字列に対する特徴選択を比較すると、トピック上位 3 つの earn, acq および money-fx については条件付確率による特徴集合を用いたトピック分類との差が適合

図 1 適合率

図 2 再現率

図 3 F 値

表 3 F 値

トピック	反復度	条件付確率	単語
earn	97.84%	97.92%	95.31%
acq	93.27%	93.42%	87.38%
money-fx	75.74%	75.52%	86.89%
grain	87.46%	85.09%	98.62%
crude	87.44%	80.23%	90.85%
trade	77.68%	74.33%	69.37%
interest	72.07%	69.69%	76.11%
ship	75.68%	63.77%	74.67%
wheat	82.44%	76.92%	76.32%
corn	84.32%	79.59%	83.33%
マクロ平均	83.39%	79.65%	83.88%
マイクロ平均	90.60%	89.21%	89.23%

率、再現率ともにほとんどないが、これら以外の7トピックについては、反復度による特徴集合を用いたトピック分類の方が結果が良くなり、特に再現率が大きく改善された。図3、表3のF値でも同様に、上位3トピックのearn, acq および money-fx では条件付確率を用いた場合と比べてF値はあまり変わらないが、これら以外のトピックでは反復度を用いた方が改善された。

10個のトピックの内、反復度が優れた結果を出したものは8個であり、劣った結果であってもF値は0.20%しか差が出ない。一方、マクロ平均ではF値に3.74%の差があり、反復度の方が良い性能を示した。マイクロ平均ではF値に約1.39%の差があり、反復度の方が良い性能を示した。ただし、このコーパスにおいてはearnとacqで全体の約65%のドキュメントがあり、文書数が偏っている。これはこのテストセットに特殊なケースであり、カテゴリごとの平均で

比較する方が実際の性能を反映すると考えられる。単語を特徴集合とした方法と比較すると、図 3、表 3 のように、結果の優劣はトピック毎に異なる。10 トピック中 6 トピックについては反復度による方法の方が、F 値が良くなるという結果になった。F 値のマクロ平均は反復度による方法と単語による方法を比較するとほぼ等しくなった。

4.1.2 20newsgroups

20newsgroups は 20 のトピックに属する文書からなる。分類は、Reuters-21578 と同様にひとつの文書が各トピックに属するか属さないかを SVM を用いて判定することによって行う。このコーパスには、学習文書 11,314 とテスト文書 7,532 文書が含まれており、トピック間での文書数の違いはあまりない。実験では、文書中の From, Subject およびニュース本文の文書から特徴に使用する部分文字列を選択し、テキスト分類を行う。この実験における、各特徴集合のパラメータ設定と選択された特徴数は以下の通りである。

- 反復度： $l=100$, $h=200000$, $a=0.1$ として 69,653 文字列が選択された。
- 条件付確率： $l=100$, $h=200000$, $b=8$, $p=0.8$, $q=0.8$ として 24,732 文字列が選択された。
- 単語： $l=5$, $h=10000$ として 26,573 単語が選択された。

l , h , a のパラメータは学習文書での 4 分割交差検定法によって再設定した。ただし、このコーパスにおいても Reuters-21578 と同様に、先頭が空白で始まる文字列は特徴として用いる文字列から除外した。また、学習時に負例の文書数が正例と比べ非常に多いことが SVM のモデルに大きな影響を与えてしまったため、正例と負例の判定エラーに対するコスト比の値を文書数の比に近い値である 20 に設定し学習を行う。以上の条件において 20 のトピックに分類した結果、F 値のマクロ平均は反復度による方法で 76.05%、条件付確率による方法で 74.75%、単語による方法で 76.71% となった。このように、このコーパスにおいても反復度による特徴選択の方が条件付確率による特徴選択よりも良い性能を示した。トピックごとの比較では、20 トピック中 10 トピックにおいて反復度による方法が単語よりも良くなるという結果になった。しかしながら、その差はほとんどないといえる。

4.2 Spam 分類実験

本実験では TREC 2006 Spam Corpus⁴をコーパスとして用いた。このコーパスは Spam 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。コーパスの資料には正確な定義はないが、コーパス作成者が主観的に有用なテキストを Ham、それ以外を Spam としたものと考えられる。これを用いた実験で分類精度が向上すれば、実際の Spam 分類においても分類精度が向上すると考えられる。

⁴ <http://trec.nist.gov/data/spam.html>

4.2.1 実験方法

トレーニングデータとして Spam, Ham それぞれ 100 個, 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. 記号, マルチバイト文字は前処理段階でカットし, 分類に用いない. このようにして 20 個の文書セットを構成し, 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)
- スペースを区切りとした単語からなる特徴集合 (WS)

各手法のパラメータは文書セットごとに交差検定により設定する. 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する. また, 反復度による手法のパラメータ l , h は条件付き確率と同様の値に設定する. テキストの学習分類は 4.1 節と同様に行い, 評価も 4.1 節と同様に F 値を用いて行うこととする.

4.2.2 実験結果

4.2.1 節で述べたようにトレーニング, テストデータのセットを 20 個作り, それぞれに対して, 特徴集合として AS, CS, WS それぞれを用いて分類を行う. このようにして得られた分類結果を表 4 に示す. 表 4 において, この F 値は (Spam の F 値+Ham の F 値) / 2 として得た平均値である. 表 5 には各文書セットに対する反復度の手法のパラメータ a と, 単語の手法のパラメータ l を示す. ここで, 反復度の手法のパラメータは a を交差検定で求め, l , h は $l=80$, $h=8000$ で固定し, 条件付確率の手法のパラメータは $l=80$, $h=8000$, $b=8$, $p=0.8$, $q=0.8$ で固定する. また, 単語の手法のパラメータ l は交差検定で求め, h は $h=8000$ で固定とした. これは, $h=8000$ を設定すると各文書セットで良い分類結果を示し, その付近で変化させても分類結果に影響がなかったためである. 表では固定されたパラメータについては表記を省略したが, 本文中に記したパラメータを使った.

表 4 を見ると, 文書セットを変えたときには平均的に反復度が優れた分類結果を示し, 条件付確率をもっとも悪い結果を示していることがわかる. 反復度を用いた結果は単語を用いた結果より平均 1.04%, 条件付確率を用いた結果より平均 2.93%だけ F 値が高い. 表から, 全 20 の文書セットすべての分類結果において, 反復度を用いた方が条件付き確率を用いるよりも良い分類結果を示していることが分かる. 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. 反復度を用いる方法と単語を用いる方法の F 値を比較すると, 20 回の分類実験のうち, 反復度が単語よりも良い結果となったのが 16 回で, 悪い結果となったのが 3 回, 同じ値となったのが 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,

表 4 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	95.00%	88.25%	93.50%
19	93.50%	93.25%	93.75%
20	93.00%	87.75%	92.00%
平均	93.15%	90.23%	92.11%

表 5 設定したパラメータ

文書セット	a	l	文書セット	a	l
1	0.15	10	11	0.2	6
2	0.375	30	12	0.3	10
3	0.1	10	13	0.25	84
4	0.3	1	14	0.25	7
5	0.4	5	15	0.35	60
6	0.475	1	16	0.2	10
7	0.25	1	17	0.15	1
8	0.2	40	18	0.35	2
9	0.325	3	19	0.48	1
10	0.3	10	20	0.2	20

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,

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

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られる. 表 5 をみると, 単語の手法のパラメータ 1 はほとんどが 10 以下で, まれに大きな値をとることがわかる. また, 反復度の手法のパラメータ a は 0.2 から 0.4 程度の値をとることがわかる.

5 考察

5.1 ニュースのトピック分類

まず, 文字列に対する 2 つの特徴選択方法, 提案手法である反復度による方法とベースラインである条件付確率による方法を比較する. 4.1 節の実験において, 学習文書とテスト文書に同じ文書集合を用いてみると, F 値のマクロ平均は, 反復度を用いた方法では 92.87%, 条件付確率を用いた方法では 95.17% となり, 条件付確率による特徴集合の方が全体的に F 値が高くなる. 学習文書とテスト文書に異なる文書集合を用いる本来の評価では, 4.1 節で説明したように反復度による特徴集合の方が F 値が高いことから, 条件付確率を用いた特徴集合では, 反復度を用いた場合に比べ, 過学習してしまう傾向があると考えられる.

ここで, 各手法で選択された文字列を比較すると, 共通して選択されたのは 1,400 文字列で, 特徴集合全体に比べて小さい. 2 つの手法で選択される文字列の差を直感的に理解しやすい例をこの文字列から一つ示す. トピックのひとつである ship に注目し, このトピックに含まれる学習文書を見るとトピック名である「ship」という単語が含まれていることがわかった. それぞれの手法で選択された文字列のうちこの単語に関連する文字列を表 6 に示す.

条件付確率による特徴選択に比べて, 反復度による特徴選択では, 直前に現れる単語の最後の 1 文字加えた文字列や統計的には類似している文字列が追加で選択されている. これらのうち共通していない 2 文字列を特徴集合から取り除いて実験を行ったところ, ship の分類結果

表 6 特徴文字列

条件付確率	反復度
ships	Ships
shi	Shi
	n ship
	ship

が 75.68% から 73.10% に減少した。これは, ship に含まれる文書中の「foreign ships」や「own shipping」のような文字列の特徴をテキスト分類に使用したためだと考えられる。連語そのものを検出しているとはいえないが, 連語の情報を利用できていることが示唆される。

また, 単語を特徴集合とする方法に比べ, 提案手法は同等の性能を示したものの有意差は認められなかった。しかしながら, 提案手法は区切り文字のないデータにおいて, 単語抽出を行うための事前処理が必要なく, また上記の連語などのような情報を損なうことのないといった利点がある。

5.2 Spam 分類

実験結果から, 部分文字列を特徴集合とする 2 つの方法を比較すると, 反復度で特徴選択した場合の方が, 分類結果が良いことがわかる。そこで, ここでは両者の特徴集合を比較し, どのような文字列によりこの差が生まれたのかについて考察する。

この考察のために, 4.2.1 節で生成した 20 の文書セットの内の一つに相当する別の文書セット 1 個を生成した。これ一つについて分類を行い, 反復度と条件付確率それぞれによる特徴集合を取り出す。さらに反復度について, 特徴集合のうちサポートベクトルとして使用された文字列を抽出する。この分類実験の結果として表 7 に示すデータが得られた。

このとき, 各手法のパラメータは次のように設定した。

- 反復度 : $l=80, h=8000, a=0.3$
- 条件付確率 : $l=80, h=8000, b=8, p=0.8, q=0.8$

ここで, 表 8 のように記号を定義する。IS は AS にあって CS にはない文字列の集合であるため, この文字列の中に条件付確率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。IS がどれほど分類に寄与しているかと, たとえばどのような文字列が寄与しているかを調べるため以下の 2 つの実験 ([実験 1], [実験 2]) を行い, その結果を用いて考察する。

[実験 1]

ここでは, AS - IS を特徴集合として分類を行う。この分類の結果として表 9 の実験 1-a に示される F 値を得た。結果を見ると, 反復度で特徴抽出した場合よりも 7.00%, 条件付確率で特

表 7 手法ごとの F 値

手法	F 値
反復度	95.25%
条件付き確率	90.25%

表 8 文字列集合の記号との対応と大きさ

文字列集合	記号	文字列集合の大きさ
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	$IS = AS - CS$	1633
反復度のサポートベクトル	SV	1976

徴抽出した場合よりも 2.00%だけ F 値が下がっていることがわかる。このことから、IS の文字列は F 値を 7.00%上昇させることがわかる。また、このときの F 値が CS で分類したときよりも下がっていることから、反復度では捉えることができなかったが条件付確率では捉えることができた分類に役立つ文字列があったことがわかる。ただし、IS のうち実際に分類に使われるものは IS $\cap$ SV (大きさは 1628) であるから、IS $\cap$ SV を AS から取り除いた場合と IS を取り除いた場合の結果は同じである。

[実験 2]

実験 1 から IS $\cap$ SV が分類結果を改善しているということがわかった。ここでは、実際にどのような文字列が分類に寄与しているのかについて調べる。

まず、考察のために作成した文書セットの分類において反復度が選んだ特徴集合 (AS) の内サポートベクトルとして用いられた部分文字列 (SV) の重み w_j (4.1 節参照) を計算する。そして、 w_j が大きいほど分類に寄与していると考え、その上位 50 の文字列をとりだす。その集合と IS の積をとり、それを AS から取り除いて分類を行う。この結果として表 9 の実験 2-a に示される F 値を得た。この結果を見ると、反復度で特徴抽出した場合よりも 2.50%だけ F 値が下がっていることがわかる。この 50 個の文字列を調べると、message_id という文字列の一部と推測できる部分文字列 12 個が含まれていることが分かった。これはたとえば表 10 に示されるような文字列である。ここで「_」は空白を意味することとする。

ただし、この 12 個の部分文字列はすべて CS に含まれていないことが分かった。これらを AS から取り除いて分類すると F 値は表 9 の実験 2-b に示される値となった。このように、これを除去することで F 値が下がるという結果から、明らかにこれらの部分文字列は分類に役立っていることがわかる。

SV 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、IS との積をとると 16 個の文字列が得られた。この 16 個の文字列を AS から取り除き分類すると F 値は表 9 の実験 2-c に示される値となった。この結果からも、message_id の部分文字列群は役立っていることが示唆される。ここで、CS にも含まれている 10 個の message_id の部分文字列を除去した場合、F 値は変化しなかった。よって、CS に含まれない 16 個の部分文字列は CS に含まれる 10 個の部分文字列をカバーするといえる。

表 9 条件ごとの F 値

実験	F 値
1-a	88.25%
2-a	92.75%
2-b	93.25%
2-c	92.75%

表 10 見つかった message_id の部分文字列の一部

age.i
message_
sage_
message
mess

表 11 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	

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.

message_id という文字列の一部が CS にも含まれており (26 個中 10 個), CS に含まれない 16 個の部分文字列を AS から取り除き分類すると分類結果が悪くなることは先に述べた. ではなぜ 10 個の文字列は CS に含まれない 16 個をカバーできなかったのか, それらの文字列の違いについてここでは考える. 考察のために, 表 11 に反復度, 条件付確率それぞれの手法が捉えた message_id の部分文字列を示す.

表 11 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引きずられやすい, 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, 不特定多数の文字列の一部となりえない特定のものをさす文字列の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わかりやす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.

6 まとめ

文字列によるテキスト分類において、条件付確率を用いて文書の特徴集合を選択する代わりに、反復度を用いて特徴選択を行い、ニュース記事のコーパスである Reuters-21578, 20newsgroups と、スパムメールのコーパスである TREC 2006 Spam Corpus のテキスト分類の結果の比較を行った。反復度によって特徴選択した特徴集合を用いると条件付確率による特徴集合を用いた場合に比べて、ニュース記事の分類では平均 79.65% から平均 83.39% と、平均 3.74% だけテキスト分類の結果を改善することを報告し、スパムメールの分類では分類結果を平均 90.23% から平均 93.15% と平均 2.93% だけ結果を改善することを報告した。このとき、その両方の実験において、提案する反復度を用いる手法と条件付確率を用いる Zhang ら手法の間に有意差があることを確認した。

また、本実験では提案手法である反復度を用いて特徴集合を選択する方法と単語を特徴集合とする方法との比較についても Zhang らの手法との比較と同様にして行った。Reuters-21578, 20newsgroups を用いたニュース記事の分類においては両手法の間に有意差は確認できなかった。しかし、TREC 2006 Spam Corpus を用いたスパムメールの分類においては、反復度による特徴抽出法を用いると、単語を特徴集合とする場合に比べて分類結果を、平均 92.11% から平均 93.15% と平均 1.04% だけ改善するということを報告した。そして、このとき危険率 1% の検定を行い両手法の間に有意差があるということを確認した。この結果の一つの要因として、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別の部分文字列と解釈されにくい部分文字列や、単語による方法では抽出できない単語と単語を結ぶような文字列が含まれているということが考えられる。よって、本研究は意味ある結果となったといえる。

謝 辞

この研究は、住友電工情報システムとの共同研究の成果です。データの解析には、戦略的情報通信開発推進制度 (SCOPE) の課題「実空間情報処理のためのインターユビキタスネットワークの研究」の成果の分析技術を利用しました。

また、多くの有益なご指摘を頂いた査読者の方々に感謝致します。

参考文献

Manning, C. and Schutze H. (1999). "Foundations of Statistical Natural Language Processing." MIT Press, Cambridge.

- Zhang, D. and Lee, W. S. (2006). “Extracting Key-Substing-Group Features for Text Classification.” In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pp. 474–483.
- Peng, F., Shuurmans D., and Wang, S. (2004). “Augmenting Naive Bayes text classifier with statistical language models.” *Information Retrieval*, **7** (3-4), pp. 317–345.
- Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2001). “Text Classification Using String Kernels.” *Journal of Machine Learning Research (JMLR)*, pp. 419–444.
- Goodman, J. (2001). “A bit progress in language modeling, extended version.” Technical report, Microsoft Research, pp. 403–434.
- Church, K. W. (2000). “Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2 .” In *Proceedings of 18th International Conference on Computational Linguistics*, **1**, pp. 180–186.
- Cristianini, N. and Shawe-Taylor, J. (2000). “An Introduction to Support Vector Machines.” Cambridge University Press, Cambridge.
- Dumais, S., Platt, J., Hecherman, D., and Sahami, M. (1998). “Inductive learning algorithms and representations for text categorization.” In *Proceedings of the 7th ACM International Conference on Information and Knowledge Management*, pp. 148–155.
- Mitchell, T. (1997). “Machine Learning.” McGraw Hill, international edition.
- Geng, Xiubo, Liu, Tie-Yan, Qin, Tao, and Li, Hang (2007). “Feature Selection for Ranking.” *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 407–414.
- Takeda, Y. and Umemura K. (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.
- Yiming, Yang and Pedersen, Jan O. (1997). “A comparative study on feature selection in text categorization.” *Proceedings of ICML-97 14th International Conference on Machine Learning*, pp. 412–420.
- 平田勝大, 岡部正幸, 梅村恭司 (2007). 文字列を特徴量とし反復度を用いたテキスト分類. 情報処理学会研究会報告, **76**, pp. 121–126.

付録

表 4 には文書セットごとの各手法の平均 F 値を示した. 表 12 には文書セットごとの各手法の Spam, Ham それぞれの F 値を示す.

表 12 表 12 各手法の Spam, Ham それぞれの F 値

文書セット	反復度 (AS)		条件付き確率 (CS)		単語 (WS)	
	Spam	Ham	Spam	Ham	Spam	Ham
1	92.50%	95.00%	89.50%	91.00%	96.50%	91.50%
2	91.50%	94.50%	92.00%	89.50%	95.50%	89.00%
3	97.50%	93.50%	93.00%	92.50%	95.00%	91.50%
4	94.00%	91.50%	90.50%	91.50%	91.00%	94.50%
5	91.00%	93.50%	94.50%	88.00%	93.00%	89.50%
6	93.50%	93.50%	92.00%	87.50%	93.00%	92.50%
7	97.00%	94.50%	95.00%	90.00%	92.00%	93.50%
8	91.00%	95.00%	92.00%	93.50%	91.00%	92.00%
9	96.00%	86.00%	96.50%	84.50%	96.50%	84.50%
10	94.00%	94.00%	91.50%	91.00%	91.50%	90.50%
11	93.50%	92.50%	90.50%	81.00%	95.50%	87.50%
12	95.50%	91.50%	93.50%	90.50%	97.00%	89.50%
13	95.00%	90.00%	95.00%	88.00%	96.50%	87.00%
14	93.50%	82.00%	94.50%	76.50%	95.50%	80.50%
15	93.00%	90.00%	90.00%	89.00%	87.50%	92.50%
16	96.00%	94.00%	90.00%	86.50%	96.00%	90.00%
17	94.50%	93.00%	88.00%	92.00%	92.00%	95.00%
18	96.00%	94.00%	95.00%	81.50%	96.00%	91.00%
19	93.50%	93.50%	95.50%	91.00%	91.50%	96.00%
20	95.00%	91.00%	88.50%	87.00%	96.50%	87.50%
平均	94.18%	92.13%	92.35%	88.10%	93.95%	90.28%

略歴

尾上 徹：2009 年豊橋技術科学大学工学部情報工学課程卒業. 同年, 同大学院入学, 現在に至る.

平田 勝大：2009 年豊橋技術科学大学大学院工学部情報工学専攻修士課程修了. 同年, NTT データ (株) 入社.

岡部 正幸：2001 年東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了. 博士 (工学). 同年科学技術振興機構 (CREST) 研究員, 2003

年 豊橋技術科学大学情報メディア基盤センター助教. 知的情報検索の研究に従事. 人工知能学会会員.

梅村 恭司: 1983 年東京大学大学院工学系研究系情報工学専攻修士課程修了. 博士 (工学). 同年, 日本電信電話公社電気通信研究所入所. 1995 年豊橋技術科学大学工学部情報工学系助教授, 2003 年教授. 自然言語処理, システムプログラム, 記号処理の研究に従事. ACM, ソフトウェア科学会, 電子情報通信学会, 計量国語学会各会員.

(2009 年 4 月 30 日 受付)

(2009 年 6 月 24 日 再受付)

(2009 年 9 月 15 日 再々受付)

(2009 年 10 月 2 日 採録)