知能と複雑系　132－23
(2003. 3. 15)

系列パターンを素性とした論文概要文の自動分類

山崎　貴宏†　新保　仁†　松本　裕治†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916–5
E-mail: †{takah-ya,shimbo,matsu}@is.aist-nara.ac.jp

あらまし　科学・技術論文の概要には研究の目的や背景，手法，結論などが述べられており，各文がそれらの役割を持って構成されている．概要中で各文の果たす役割を判別することは，文を単位としたテキスト分類問題とみなすことがでる．テキスト分類問題への機械学習の適用の際には，多くの場合単語を素性としたモデル（"bag-of-words"）用いられるが，各文の役割判定のタスクにおいては語順が大きな手掛かりとなると考えられ，bag-of-words モデルのみでは不十分と考えられる．本稿では，テキストマイニングの手法によって抽出されたパターンを素性として用い，医学論文データベース MEDLINE の概要文の分類に適用し，その有効性を報告する．
キーワード　テキスト分類，素性選択，系列パターン，MEDLINE アブストラクト

Takahiro YAMASAKI†, Masashi SHIMBO†, and Yuji MATSUMOTO†

† Graduate School of Information Science,
Nara Institute of Science and Technology
8916–5, Takayama, Ikoma, Nara 630–0192, Japan
E-mail: †{takah-ya,shimbo,matsu}@is.aist-nara.ac.jp

Abstract　We explore the use of (possibly non-contiguous) word sequence patterns as the features in the task of automatically classifying sentences in the MEDLINE abstracts. In this task, the categories to which the sentences are classified are determined not by their topics, but in accordance with the typical subsections in the abstracts; i.e., Background, Objectives, Conclusions, etc. The bag-of-words representation commonly used in text categorization is inadequate for this particular task, as the turns of phrase or expression characterize the categories better than the individual content words. An improvement of 3 to 11 points in terms of F-measure was observed when the sequential patterns mined with the PrefixSpan algorithm were used in addition to the bag-of-word features.
Key words　Text Categorization, Feature Selection, Sequential Pattern, MEDLINE Abstracts

1. はじめに

科学・技術論文の概要やあらましには研究の背景や目的，手法，結果など，論文の述べる趣旨となる事項が記述され，それぞれの事項の役割は大きく異なる．論文概要内で文の果たすこれらの役割の自動判別は，文を単位としたテキスト分類問題と考えることができる．

近年の機械学習手法の発展により，汎化性能の高い学習アルゴリズムが提案され，自然言語処理に適用されている．Joachims [2] は "bag-of-words" と呼ばれる単語を素性としたモデルを用いて，Support Vector Machine (SVM) [7] を新聞記事のカテゴリ分類に適用し，良い結果を残している．しかし，素性に単語のみを用いると，語順の問題が生じる．つまり，"The earth goes around the sun." という文と "The sun goes around the earth." という文は語順が異なるだけで，文を構成する単語が同じため，"bag-of-words" 素性空間上では同じ文とみなされる．記事の内容に基づく文書分類であれば，用いられる単語そのものが内容の手掛かりとなり，語順は大きな問題にはならない．しかし本稿で扱う，概要中で文の果たす役割の判定というタスクにおいては，内容語自体に意味はなく，むしろ特徴的な言い回しなどが文の識別の手掛かりとなると考えられる．このため，語順を無視することは情報の大きな損失となる．

そこで，本稿ではテキストマイニングの手法を用いて頻出系列パターンを抽出し，機械学習の素性として用いた．系列パターンとは連続または非連続の単語列のことを指し，単語のみを素性としたときに生じる語順の問題を避けることができる．具体的には，前の例では，連続 2 単語をパターンとしたとき，"earth goes" と "sun goes" というパターンにより 2 つの文を

識別することができる．さらに，単語列が必ずしも連続である必要がないため，“The aim of *the* study ...”，“The aim of *this* study ...”，“The aim of *the present* study ...” という文から “The aim of study” という共通のパターンを抽出することができる．このように，非連続パターンは冠詞や代名詞，形容詞などの表現の異なりを吸収できるため，独特の言い回しなど，特徴的なパターンの抽出が期待できる．

本稿では医学論文データベース MEDLINE(注1)の論文概要文から PrefixSpan [6] アルゴリズム用いて頻出系列パターンを抽出し，得られたパターンを素性として各文の果たす役割（背景，目的，手段，結論など）の判定を行なう．SVM を用いて，“bag-of-words” モデルとの比較を行ない，系列パターンの素性としての有効性を示す．概要中で文の果たす役割を以下ではクラスと呼ぶことにする．

本稿の構成は以下の通りである．まず 2 章でクラスの自動判定によって想定される応用事例を述べ，3 章で技術的な背景である PrefixSpan，SVM について説明し，4 章で実験手順と結果，考察を述べ，5 章で結論と今後の課題について述べる．

2. 想定される応用事例

文がどのクラス（背景，目的，手段，結論など）に属するか，というクラスの自動判定は，論文検索や知識獲得への応用が考えられる．

論文検索では，検索者の欲する情報に応じて検索のランキングの重み付けを変え，検索者にかかる負荷を軽減することが考えられる．具体的には，医師がある病気に対するある薬品の効果が知られているかどうかを調べたいとするとき．病名と薬品名の出現する結果クラスや結論クラスの文のみを見ればよい．しかし，現状では単にキーワードの追加のみの絞り込みで，文がどのクラスに属しているかということは考慮されていないため，クラスを限定して検索を行なうことはできない．あらかじめ文とクラスの対応がとれていれば，キーワードに加え，“結果” や “結論” といったクラスを指定して検索・絞り込みを行なうだけで目的文を探すことができる．

知識獲得においては，膨大なデータからの興味深い規則の発見に利用できると考える．データマイニングの手法では頻出する共起関係から相関ルールを得るが，このようにして得られた全ての頻出ルールが興味深く，重要なものとは限らない．大多数は常識的，あるいは既知の規則であろう．背景クラスの文にはそのような常識的な規則が記述されていると仮定すると，背景クラスから何らかの方法で規則を獲得し，その規則に一致しないもののみを重要な規則の候補とすることが考えられる．

3. 技術的背景

3.1 PrefixSpan

系列パターンの抽出に用いるアルゴリズムである PrefixSpan の概要について説明する．

系列パターンマイニングは，Agrawal [1] らによって以下のように定式化された．

$I = \{i_1, i_2, \ldots, i_k\}$ を**アイテム**集合とする．アイテムの列を，**系列**と呼び，$\langle u_i, u_2, \ldots, u_m \rangle$，と表す．ある系列 α 中のすべてのアイテムが，別の系列 β 中に存在し，その系列 β は系列 α を「含む」という．系列 id $sid \in N$ と系列 s のタプル $\langle sid, s \rangle$ の集合 $\{\langle 1, s_1 \rangle, \langle 2, s_2 \rangle, \ldots, \langle n, s_n \rangle\}$ を**系列データベース** S とし，S にあるタプルのうち，系列 α を含むものの数を，**サポート** $support_S(\alpha)$ とする．

系列パターンマイニングは，系列データベース S と任意の整数 ξ に対し，$support_S(\alpha) \geq \xi$ となるような系列 α を全て列挙することを指す．このときの ξ を *minimum support* という．

これを，アイテムを**単語**，系列を**文**，系列データベースを**テキスト**と置き換えて考えると，テキスト中に ξ 文以上に出現する（連続，非連続）単語列を漏れなく抽出する問題，とみなす．

この問題に対し，PrefixSpan は**射影**という操作を繰り返し行なうことで，深さ優先探索で多頻度パターンを抽出するアルゴリズムである．

射影とは，ある系列 $s = \langle a_1, a_2, \ldots, a_m \rangle$ とアイテム a に対し，$a_1 \neq a, a_2 \neq a, \ldots, a_{j-1} \neq a, a_j = a$ となるような整数 j $(1 \leq j \leq m)$ が存在する場合，系列 $\langle a_{j+1}, a_{j+2}, \ldots, a_m \rangle$ を作成し，それらの集合（**射影データベース**）$S|a$ を改めて系列データベースとする操作である．

PrefixSpan は系列中のある多頻度アイテム a について，minimum support 以下の頻度まで再帰的に射影を繰り返し，アイテム a から始まる全てのパターンの走査を行なう．これを全ての多頻度アイテムに適用することにより，効率的に系列パターンを抽出する．

3.2 Support Vector Machine

Support Vector Machine（SVM）は，二値分類を行なう教師付き学習モデルである．特徴ベクトルと正負のラベルのペアの集合

$$(x_i, y_i), \ldots, (x_l, y_l), \quad x_i \in R^n, y_i \in \{+1, -1\}$$

を訓練データとする．ここで，x_i は事例 i の n 次元素性ベクトルで，y_i は事例 i が正例（+1）か負例（−1）かを表すラベルである．

SVM は n 次元素性空間上の分離超平面

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

により訓練データを正例と負例に分類する．つまり，全ての i に対し，

$$y_i \cdot (w \cdot x + b) > 0$$

の成り立つ分離超平面を考えるが，このような超平面は一般に無数に存在する．SVM はこれらのうち，最も近い事例との距離（マージン）を最大にするような超平面（最適分類超平面）を求める．このようにして求められた超平面は，テストデータに対する分類誤差の期待値を最小にすることが示されている．そのため，素性空間の次元が大きい場合でも，過学習を起こし

(注1)：http://www.ncbi.nlm.nih.gov/PubMed/

にくいという性質を持つ[7]．このとき，分離平面に最も近い特徴ベクトル（事例）をサポートベクトルと呼ぶ．サポートベクトルによって事例の分類を行なう決定関数が定まるため，それ以外の事例は分類に影響を及ぼさない．

テスト事例 x が与えられた場合，以下に示す判定式の正負によって事例のラベル y を決定する．

$$y = \mathrm{sgn}(w \cdot x + b)$$

線形分離不可能な問題を扱う場合，特徴ベクトルをより高次元の空間に写像し，そこでの線形分離問題として取り扱う．SVM は学習，分類に特徴ベクトルの内積しか用いないという性質があるため，高次元空間上における写像の内積を求める関数を使用することにより，実際に高次元空間での計算をすることなく学習，分類を行なえる．このような関数（カーネル関数）を使用することで効率的な計算を行なうことができる．また，現実に分離可能な平面がない場合，soft margin と呼ばれる手法を用いて，若干の例外を認めることにより分離を行なう．どの程度の例外を認めるかは，パラメータ $C \in [0, \infty)$ を通じて制御する．C が小さい場合にはより多くの例外を認めて最適分離超平面を決定する．

4. 実　　験

系列パターンの分類における素性としての有効性を示すため，MEDLINE の論文概要を対象に PrefixSpan で抽出した系列パターンを素性とした分類実験を行なった．

4.1 デ ー タ

MEDLINE のデータは論文タイトルや著者，掲載雑誌とその巻，号，ページといった情報が XML 形式で記録されており，概要は`<AbstractText>`，`</AbstractText>`タグで区切られる中に記述されている．入手した MEDLINE データ 11,299,108 件のうち，概要を含んでいるものは 5,912,271 件あった．これらのうち，論文の著者がクラス名を見出しとして明記した書式に従って記述しているものが 374,585 件ある．このようにクラス名と，その示す部分の対応が既知のデータを本稿では実験に用いた．クラス名とその示す部分は，各クラスの先頭が "`BACKGROUND:␣...`" のように "（大文字）クラス名:␣内容の記述" という書式で記述されている．この書式を手がかりに，10 万件以上の概要テキストに出現した BACKGROUND,OBJECTIVE, METHODS, RESULTS, CONCLUSION, CONCLUSIONS の 6 つのクラスの文集合を抽出した．ここで，CONCLUSION と CONCLUSIONS は同一のクラスと考え，以下 5 クラスとして話を進める．また，見出しのない，一般の概要文を扱うことを想定して，SVM での学習，分類の際には各クラスの見出しは取り除いた．得られた文集合は文毎に区切られていないため，文集合を文に区切った(注2)が，このとき，3 単語以下からなる文は正しく文に区切られていないと考え，実験データに加えなかった．また，CONCLUSION(S) クラスの文に関しては，"International Journal of Obesity (2000)24, 101-107" のように雑誌の発行年などの情報が記述されているものが多数含まれているため，10 単語以下でかつ，数字の連続が 3 回以上あるものは実験データに加えないという条件を加えた．この結果，5 クラス合計 2,597,168 文を得た．各クラスの文数の内訳を表 1 に示す．

Class	# of sentences
BACKGROUND	264,589
OBJECTIVE	166,890
METHODS	540,415
RESULTS	1,378,785
CONCLUSION(S)	246,607
Total	2,597,286

表 1　各クラスの文数．
Table 1　Number of sentences in each class.

4.2 PrefixSpan を用いた系列パターンのマイニング

次に，得られた各クラスの文から PrefixSpan を用いて各クラスに特徴的な系列パターンを抽出する．このとき，前処理として得られた文から記号を除き，数字を単一の記号に置換した後，全てのアルファベット文字を子文字に変換した．数字の置換は METHODS，RESULTS クラスには数字が数多く出現し，数字の連続がそれらのクラスの特徴的なパターンとなると考えたためである．この後，単語 2-gram を単位（アイテム）として系列パターンの抽出を行なった．単語 2-gram とは，連続する 2 単語を 1 単語とみなしたものをいう．具体的には，"The earth goes around ..." という文は，"The_earth，earth_goes，goes_around ..." という単語 2-gram の列と考える．単に単語のみを単位としたときには "a a the ..." といった無意味な高頻度語のパターンが数多く抽出される．この問題を解決するため，隣接する 2 単語を 1 単語とする単語 2-gram を単位とすることで，単語種を増やし，その結果，高頻度で出現する無意味なパターンの頻度を相対的に下げ，意味のあるパターンを抽出することができる．パターン抽出のパラメータとして，minimum support を各クラスの文数の 0.05% と設定し，単語 2-gram の系列パターンを 5 クラス合計 242,005 パターン抽出した．

4.3 SVM を用いた学習，分類

SVM は二値分類器であるので，直接的に多値分類に適用することができない．本稿では，各クラスについてそのクラスかそれ以外かを分類する実験を行なった．

5 クラス計 2,597,168 文からランダムに訓練データとして 70,000 文選び，残りからテストデータとして 10,000 文をランダムに選び出したデータを使用した．

カーネル関数には線形カーネルを用い，素性として，(1) 各文に含まれる単語と，パターンマイニングによって得られた単語 2-gram の系列パターンを用いたもの，及び (2) 単語のみを用いたもの，の 2 種類について実験を行なった．各文を特徴付けるベクトルは素性の重みが 0 か 1，つまり文中に単語やパターンが出現すると，ベクトル要素を 1 とし，出現しない場合は 0 となるベクトルで表現する．抽出したパターンが一度も出現しない文も多数存在するため，今回はパターンのみを素性とする

(注2)：http://l2r.cs.uiuc.edu/~cogcomp/cc-software.htm

実験を行なわなかった．SVMの実装としてTinySVM(注3)を使用した．

4.4 結果と考察

表2はsoft marginパラメータCを{0.01, 0.065, 0.1, 0.65, 1}と変化させた中で，最もF-measureの高い結果のみを示した．表中の各欄の意味を以下に示す．pos/test行はテストデータ10,000件中の各クラスの文の占める割合を示し，w，w+p欄は学習に用いた素性を表し，wは単語のみ，w+pは単語とパターンを素性として用いたことを示す．C行は学習の際のsoft marginのパラメータを表す．SV，SV/sample行はサポートベクターの数と訓練データ70,000件に対する割合を表す．Accuracyテストデータのうち正しく正例，負例に分類されたものの割合，Precisionは分類器が正例と判定した事例のうち，正しく分類された割合，Recallは正例の事例のうち，分類器が正例と判定したものの割合，を示す．F-measureはPrecisionとRecallを同程度重視した評価尺度で，PrecisionをP，RecallをRとすると以下の式で表される．

$$\text{F-measure} = \frac{2 \cdot P \cdot R}{P + R}$$

また，±行は素性に単語のみを用いたものに対し，単語とパターンを素性に用いたもののF-measureの増加分を示す．

（RESULTSクラス以外は）正例と負例の割合に偏りがあるため，Accuracyだけでなく，PrecisionやRecallと，それらを考慮したF-measureを用いて評価を行なった．単語のみの素性に比べパターンの素性を追加することにより,F-measureに3から11程度の向上が見られた．

また，各クラス間の類似度を単語，単語とパターンについて測った（表3，表4）．表中のother行は各列のクラスと，それ以外の4クラスとの類似度を表す．類似度の尺度としてinformation radius [3]を用いた．分布p_nとq_nのinformation radius $D(p_n \| q_n)$は以下のように計算される．

$$D(p_n \| q_n) = \frac{1}{2}\Big[\sum_x p_n(x) \log \frac{p_n(x)}{(p_n(x) + q_n(x))/2} + \sum_x q_n(x) \log \frac{q_n(x)}{(p_n(x) + q_n(x))/2}\Big]$$

式から，p_nからq_nへの類似度とq_nからp_nへの類似度は同じ値をとることがわかる．このため，表は対角線を軸として対称の値をとる．Information radiusは分布間の距離を表す尺度のため，値が小さいほど類似度が高いことになる．分類精度の向上という観点からは，この値が大きいほど，クラス間の距離が離れ，精度の高い分類が期待される．ランダムに素性を追加すると分布間の距離が近付き，information radiusの値は下がるが，それに反してパターン素性を加えたときの値は単語のみ（表3）の場合に比べて大きくなっている（表4）．この面からも，パターンが有効な素性であると考えられる．もっとも，先述のとおり，素性として用いたパターンが1つも含まれていない文も数多く存在するため，全事例（文）をカバーするためには単語等の他の素性と組み合わせて用いざるを得ない．

参考までに，表5にパターンのみでのinformation radiusの値を示しているが，全クラスにおいて単語のみ，単語とパターンの素性よりも値が低い．これも，パターンの素性が有効であることを示唆している．

5. おわりに

本稿では，医学文献データベースMEDLINEの概要文を対象に，構造をもつ文書の各文が文書中で果たす役割の判別に，単語の他に系列パターンが有効な素性であることを確認した．

今後の課題として，次のような項目が挙げられる．系列パターンの抽出の際，機能語のみからなるような，無意味な高頻度パターンを除くためパターンの単位（アイテム）を単語2-gramとしたが，この方法によると，単語2-gramでは“the aim of {this, the} study {was, is} that ...”という文から“the aim of study that”という系列パターンは抽出できず，系列パターンの有効性が必ずしも発揮できているとはいえない．この問題を解決するため，単語を単位としたパターン抽出し，その後得られた系列パターンをinfomation gainなどの尺度で評価を行ない[5], [8]，有効なパターンの選別を行なうことが考えられる．また，minimum supportの設定を各クラスの文数の0.05%と設定したが，理論的な根拠はなく，改善の必要があると考えられる．

SVMでの学習，分類では事前にPrefixSpanを用いて頻出系列パターンの抽出を行なったが，パターンの抽出には，全文を走査するため計算量的な負荷が高い．そこで，カーネル関数としてstring kernelを導入することが考えられる [4]．String kernelは系列パターンを基底とする，素性空間中での内積の計算を行なうカーネル関数である．カーネル関数によって系列パターンの計算を陰に行なうため，事前に全文を走査して系列パターンの抽出を行なう必要がなくなり，計算量の軽減を図ることができると考える．また，今回の実験では素性の重みを0または1としたが，この重みを単語や系列パターンの頻度といった，統計量を用いた重みにしたとき，どのような変化が見られるか，確認する必要がある．

今回は，文中に現れる単語や系列パターンのみを素性として用いたが，論文の概要文は背景，目的，手法，結果，結論と，流れをもって書かれていることが多いため．したがって，背景部分の記述は概要中で冒頭に書かれていることが多く，同様に結論部分は概要の最終部分に書かれていることが多いと予想される．このように概要中での文の位置の情報というのも分類に有効な要因と考えられるため，このような情報も考慮することにより，精度の向上が期待される．

(注3)：http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

	BACKGROUND		OBJECTIVE		METHODS		RESULTS		CONCLUSION(S)	
pos/test(%)	10.27		6.39		21.58		52.19		9.57	
	w	w+p	w	w+p	w	w+p	w	w+p	w	w+p
C	0.1	0.1	0.1	0.1	0.065	0.065	0.065	0.065	0.65	0.1
SV	14,088	14,540	9,468	10,886	16,800	15,159	24,883	21,106	14,142	13,275
SV/sample(%)	20.13	20.77	13.53	15.55	24.00	21.66	35.55	30.15	20.20	18.96
Accuracy	92.33	92.84	95.66	95.92	92.02	93.96	88.14	91.34	92.13	94.10
Precision	68.57	67.85	81.16	81.47	84.73	88.16	88.70	91.93	61.97	77.27
Recall	46.73	57.55	41.78	46.79	76.88	83.18	88.56	91.44	45.98	54.34
F-measure	55.59	62.28	55.17	59.44	80.61	85.60	88.63	91.68	52.79	63.80
±	6.69		4.27		4.99		3.05		11.01	

表 2　実験結果.

Table 2　Experimental result.

class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)	other
BACKGROUND	0	0.0438	0.1209	0.1132	0.0447	0.0784
OBJECTIVE	0.0438	0	0.1046	0.1143	0.0639	0.0754
METHODS	0.1209	0.1046	0	0.0715	0.1254	0.0675
RESULTS	0.1132	0.1143	0.0715	0	0.0898	0.0616
CONCLUSIONS	0.0447	0.0639	0.1254	0.0898	0	0.0671

表 3　Information radius: 単語.

Table 3　Information radius: word.

class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)	other
BACKGROUND	0	0.0617	0.2827	0.5038	0.1792	0.4090
OBJECTIVE	0.0617	0	0.3093	0.5177	0.2580	0.4253
METHODS	0.2827	0.3093	0	0.3624	0.2808	0.3209
RESULTS	0.5038	0.5177	0.3624	0	0.5100	0.4135
CONCLUSIONS	0.1792	0.2580	0.2808	0.5100	0	0.4386

表 4　Information radius: 単語+パターン.

Table 4　Informatiion radius: word+pattern.

class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)	other
BACKGROUND	0	0.0659	0.3997	0.5737	0.2995	0.4941
OBJECTIVE	0.0659	0	0.4274	0.5870	0.3864	0.5130
METHODS	0.3997	0.4274	0	0.3853	0.4341	0.3580
RESULTS	0.5737	0.5870	0.3853	0	0.6040	0.4588
CONCLUSIONS	0.2995	0.3864	0.4341	0.6040	0	0.5548

表 5　Information radius: パターン.

Table 5　Information radius: pattern.

文　献

[1] R. Agrawal, R. Srikant. "Mining sequential patterns," Proc. 11th Int. Conf. Data Engineering, ICDE, pp. 3-14. IEEE Press, 6-10 1995.

[2] T. Joachims. "Text categorization with support vector machines: learning with many relevant features," Technical Report LS-8 Report 23, Computer Science Department, University of Dortmund, Dortmund, Germany, 1997.

[3] L. Lee. "Measures of distributional similarity," In Proc. of the 37st ACL, pp. 25-32, 1999.

[4] H. Lodhi, J. Shawe-Taylor, N. Cristianini, and, C. Watkins. "Text classification using string kernels," Technical Report NC-TR-2002-079, NeuroCOLT, 2002.

[5] S. Morishita. "On Classification and Regression," In Proc. of the First International Conference on Discovery Science, 1998.

[6] J. Pei, B. Mortazavi-Asl, J. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," In Proc. of International Conference of Data Engineering, pp. 215-224, 2001.

[7] V. Vapnik. "Statistical Learning Theory," John Wiley & Sons, 1998.

[8] Y. Yang and J.O. Pedersen. "A comparative study on feature selection in text categorization. In Machine Learning: Proc. 14th Int. Conf. ICML, pp. 412-420. 1997.