Support Vector Machine によるテキスト分類における属性選択

平 博順[†] 春野雅彦[††]

本論文では Support Vector Machine（SVM）を使ったテキスト分類における属性選択手法について述べる．我々は最適な属性選択を調べるため相互情報量を基準とした属性選択と品詞を基準とした属性選択を比較した．前者の実験では相互情報量の大きい単語を順に追加して属性を増やし，後者の実験では普通名詞のみの単語属性から始めて固有名詞，サ変名詞，未定義語，動詞を追加して属性を増やした．その結果，1) 最適な属性数はカテゴリごとに異なるが，2) 平均すると品詞基準の属性選択で普通名詞，固有名詞，サ変名詞，未定義語，動詞の 5 品詞の単語をすべて使用したときに最高の精度が得られた．この結果から SVM の汎化能力は非常に高く，高い分類精度を得るためには品詞によるフィルタリングという単純な処理のみを行い，後は全単語を入力として用いればよいことが明らかになった．

Feature Selection in SVM Text Categorization

Hirotoshi Taira[†] and Masahiko Haruno[††]

This paper investigates the effect of prior feature selection in Support Vector Machine (SVM) text categorization. The input space was gradually increased by using mutual information (MI) filtering and part-of-speech (POS) filtering, which determine the portion of words that are appropriate for SVM learning from the information-theoretic and the linguistic perspectives, respectively. The experimental results are that 1) the optimal number of features differed completely across categories, and 2) the average performance for all categories was best when all of the words were used. In addition, a comparison of the two experiments clarified that POS filtering consistently outperformed MI filtering, which indicates that SVMs cannot find irrelevant parts of speech. These results suggest a simple strategy for using a full number of words found through a rough filtering technique like part-of-speech tagging.

1. はじめに

インターネットの急速な成長とともにオンライン情報が増加し，目的とする情報を容易に取得することが困難になってきている．情報を容易に取得するための技術の 1 つとしてテキスト自動分類技術が重要視されている．これまでテキスト自動分類では，人手で書いたルールに基づく分類手法と機械学習による分類手法が用いられてきた．前者は，分類対象が限定されたり，分類すべきカテゴリ数が少ないなどの場合には高い精度を得ることができるが，分類対象が変わるごとに人手で新たにルールを作成しなければならずコストがかかる．また，カテゴリ数が増えルール数が多くなった場合，新たに追加したルールによる副作用を考慮することが非常に困難になる．一方，後者の機械学習手法による方法は，多数のカテゴリがありデータ数が大規模な場合や，ユーザの要求によって頻繁に分野が変わるような場合に前者より優れている．

このような背景から，k-近傍法[12]，決定木[6]，Naive-Bayes[6]など様々な学習手法が適用されてきた．これらの学習手法を適用する際の問題に属性選択がある．高い分類精度を達成するために各分野のキーワードとなる単語属性を大量に用いると，過学習が起こりやすく計算時間も増大してしまう．そこで，分類に必要な単語属性だけをあらかじめ選ぶ属性選択が用いられるが，学習手法によって最適な単語属性数が異なるため，学習手法ごとに適切な属性選択基準を選ぶ必要があった[6]．また，過学習の問題から単語属性数が大規模な場合の属性選択の効果についてこれ

† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
†† ATR 人間情報通信研究所
ATR Human Information Processing Research Laboratories

まであまり調べられてこなかった．たとえば文献 3) では FindSimilar，NaiveBayes，BayesNets，決定木を用いたテキスト分類が比較されている．相互情報量で属性選択を行い，決定木に関しては属性数 300，FindSimilar，NaiveBayes，BayesNets に関しては属性数 50 で比較を行い，break-even point の精度評価で FindSimilar 64.6%，NaiveBayes 81.5%，BayesNets 85.0%，決定木 88.4%である．300 程度の属性数で決定木は他の手法と比べて遜色のない性能を示している．そこで以下では従来手法の代表として比較実験に決定木を用いた．

Support Vector Machine（SVM）[1),11)]は与えられたデータを超空間上で正例集合と負例集合へと分離する際，マージンを最大にすることによって最適な分離超平面を得る学習手法である☆．SVM は手書き文字認識[11)]，人の顔の検出[8)]といった高次元の入力属性が必要であるような多くの分類問題において，優れた汎化能力を持つことが明らかになっている．テキスト分類問題においても文献 3)，5) の研究がある．これらの研究で SVM が非常に高い分類能力を持つことが示されているが，使われている単語属性選択の方法はまったく異なっている．文献 5) では訓練データにおいて少なくとも 3 回以上出現した単語から ‘and’ や ‘or’ といった，いわゆる stop word を除いた属性のみを選択し，属性値としては 1 テキスト中での単語の出現頻度を加味する IDF [10)]を用いている．一方，文献 3) では各カテゴリに対して高い相互情報量[2)]を持つ上位 300 語のみを使用しており，属性値は 1 テキスト中での単語の出現有無を表す 0，1 で表現されている．このように両者の研究では異なる属性選択方法がとられており，SVM を使ったテキスト分類において異なる属性選択をとった場合の影響については知られていない．

日本語や中国語のような膠着言語では単語の分割自体難しく，単語の分割誤りや未知語による影響も考慮しなければならないため，属性選択は重要な問題である．特に未知語には分類に不要な単語だけでなく個人名，企業名といった分類に関係の深いキーワードも含まれているため，適切な未知語を属性として選ぶことがきわめて重要である．

本論文では日本語新聞記事コーパスを使用して SVM によるテキスト分類における最適な属性選択について調べた．属性選択では相互情報量を基準とした選択手法と品詞を基準とした選択手法の比較を行った．どちらの選択手法も入力属性数を徐々に増やして分類に最適な属性集合を調べた．なお，相互情報量基準は情報理論的な観点から，品詞基準は言語学的な観点から，各々単語を選択するものである．前者の実験では相互情報量の大きい順に属性数を 300 から 15,000 まで増やした．後者の実験では属性を 1) 普通名詞，2) 1) + 固有名詞，3) 2) + サ変名詞，4) 3) + 未定義語，5) 4) + 動詞，の順に増やし最適な単語属性集合を求めた．また比較のため，相互情報量による選択と品詞による選択の効果を決定木（C4.5）でも評価した．

本論文の構成は以下のとおりである．次章で SVM について簡単に説明し入力属性空間が大規模でも過学習を回避できる理論的根拠およびテキスト分類への適用について概略を述べる．3 章で SVM に対して相互情報量を基準とした属性選択と利用する品詞を基準とした属性選択の結果について C4.5 の場合と比較して述べる．4 章で 2 つの属性選択手法に対する考察を行い，最終章で結論を述べる．

2. Support Vector Machine

2.1 理論的背景

SVM は最小の汎化誤差を保証する仮説 h を見つける構造的リスク最小化[11)]に基づく手法である．

$$error_g(h) \leq error_t(h) + 2\sqrt{\frac{\lambda(\ln\frac{2l}{\lambda}+1) - \ln\frac{\eta}{4}}{l}} \tag{1}$$

式 (1) は仮説 h の汎化誤差 $error_g(h)$ が $1-\eta$ 以上の確率で，訓練データにおける仮説 h の誤差と仮説 h の複雑さの和で抑えられることを表す[11)]．ここで，l は訓練データの数，λ は仮説空間の表現力を表す VC 次元[11)]である．

式 (1) の右辺第 1 項 $error_t(h)$ は訓練データにおける誤差，第 2 項は仮説空間の複雑さを表し，両者はトレードオフの関係にある．つまり仮説空間が単純な場合（すなわち λ が小さい場合）訓練データを精度良く近似する関数を含みにくいため，第 1 項の訓練誤差が大きくなる．逆に仮説空間の表現力が高い場合（すなわち λ が大きな場合）は，訓練誤差は小さいが第 2 項は大きくなる，いわゆる過学習を起こす．

SVM では入力ベクトルを $\boldsymbol{x}$ としたとき，次の関数が仮説 h を表すと仮定する．

$$\begin{aligned} h(\boldsymbol{x}) &= \mathrm{sign}\{\boldsymbol{w}\cdot\boldsymbol{x}+b\} \\ &= \begin{cases} +1, & \text{if } \boldsymbol{w}\cdot\boldsymbol{x}+b>0 \\ -1, & \text{else} \end{cases} \end{aligned} \tag{2}$$

ここで $\boldsymbol{w}$，b はパラメータである．入力ベクトル $\boldsymbol{x}$ の次元 n と VC 次元 λ の関係については次の補助定

☆ ここでマージンとは直観的にはデータ点から分離超平面までの距離である．

理が知られている．

補助定理 1 (Vapnik) 仮説 $h(\boldsymbol{x})$ として超平面 $h(\boldsymbol{x}) = \text{sign}\{\boldsymbol{w}\cdot\boldsymbol{x}+b\}$ を仮定する．l 個の訓練データ $\boldsymbol{x}=\boldsymbol{x}_i$（$i$ は 1 から l までの整数）すべてを含む半径 R の球が存在し，各 $\boldsymbol{x}_i$ に対して

$$|\boldsymbol{w}\cdot\boldsymbol{x}_i+b|\geq 1$$

が成り立つならば，$||\boldsymbol{w}||$ を $\boldsymbol{w}$ のノルムとしたとき，VC 次元 λ について

$$\lambda \leq \min([R^2||\boldsymbol{w}||^2], n)+1 \tag{3}$$

が成り立つ．

ここで，式 (3) から VC 次元は入力ベクトルの次元ではなく，$||\boldsymbol{w}||$ に依存する場合があることが分かる．つまりテキスト分類において単語属性数を大規模にしても SVM は高い汎化能力を持ちうることが分かる．ただし，式 (1) と式 (3) では緩い上限が与えられているにすぎず，実験的な評価が必要となる．

SVM は基本的に訓練データを正例と負例に分け，正負例間のマージンが最大，すなわち $||\boldsymbol{w}||$ が最小になるような超平面を求める．この問題は Lagrange 乗数を導入し 2 次最適化問題として扱うことができる．2 次最適化問題については，一般最適解を得るためのアルゴリズムが存在する．具体的には式 (4) が最小となるような係数 α_i の集合を求める最適化を行い，式 (5) により α_i から $\boldsymbol{w}$ を構成し，マージンを最大にする超平面を得ることができる．

$$-\sum_{i=1}^{l}\alpha_i+\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\cdot\boldsymbol{x}_j$$
$$\text{s.t.}\sum_{i=1}^{l}\alpha_i y_i=0,\quad \forall i:\alpha_i\geq 0 \tag{4}$$

$$\boldsymbol{w}=\sum_{i=1}^{l}\alpha_i y_i \boldsymbol{x}_i \tag{5}$$

ここで，y_i は $\boldsymbol{x}_i$ の正負例を表す変数であり，$\boldsymbol{x}_i$ が正例のとき $+1$，負例のとき -1 の値をとる．

また，SVM では式 (4) 中の内積を kernel 関数で置き換えることによって非線形の仮説 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$☆を扱うことができる．kernel 関数には多くの種類があるが本論文では d 次の多項式関数，

$$K_{poly}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+1)^d \tag{6}$$

を使用した．

2.2 テキスト分類への適用

SVM をテキスト分類へ適用するためにはテキストから入力ベクトルを作る必要がある．まず訓練データとなる記事について形態素解析を行い，分割された単語ごとに品詞を特定する．単語が出現したときに 1，出現しなかったときに 0 を要素に持つ入力ベクトル $\boldsymbol{x}_i$ を作成する．同時に，対象記事が分類対象カテゴリに属せば $+1$，属さない場合には -1 となるラベル y_i も作成する．ここで i は記事の通し番号である．たとえば，記事がスポーツカテゴリに属するか否かの分類法を学習しているときに，

「私のパソコンのメールボックスに毎週届く、
心温まるメール。」

というコミュニケーションカテゴリに属する記事が入力されたとすると，「私（普通名詞）の（名詞接続助詞）パソコン（普通名詞）の（名詞接続助詞）メールボックス（普通名詞）に（格助詞）毎週（時相名詞）届く（動詞）、（読点）心（普通名詞）温まる（動詞）メール（サ変名詞）。（句点）」と形態素解析する．ここで括弧内は前の単語の品詞名である．ついで属性選択を行う．たとえば，属性として普通名詞，固有名詞，サ変名詞，未定義語，動詞の 5 品詞のみを選択する場合には，「私」「パソコン」「メールボックス」「届く」「心」「温まる」「メール」を抽出する．そして，たとえば，入力ベクトルの要素が，第 1 要素（「愛」の出現有無），第 2 要素（「温まる」の出現有無），第 3 要素以下，同様に「カラス」「心」「サッカー」「シュート」「届く」「パソコン」「メール」「メールボックス」「野球」「塁」「ワイド」「私」... の出現有無を表すとすると，

$$\boldsymbol{x}_i=(0,1,0,1,0,0,1,1,1,1,0,0,0,1,\ldots)$$

という入力ベクトルが作成できる．また，記事は，コミュニケーションカテゴリに属し，スポーツカテゴリには属さないため，変数 y_i は -1 の値をとる．同様にテストデータについても入力ベクトルを作り，SVM で学習，分類を行う．

3. 実験結果

本章では SVM による日本語テキスト分類における属性選択手法として相互情報量フィルタリング，品詞フィルタリングを用いた場合の実験結果について述べる．なお，比較のため同じ実験を決定木 C4.5 [9]（パラメータはデフォルト値）に対しても行った．

3.1 実験設定

分類対象として毎日新聞（1994 年発行分）の 30,207 記事からなる RWCP コーパス[14]を使用した．このコーパスには各記事が属するカテゴリを表す複数の UDC コード[13]が付与されている．この中からスポーツ，犯

☆ ただし Mercer 条件[11]を満足していなければならない．

表 1 対象データのカテゴリ別内訳
Table 1 RWCP corpus for training and test.

カテゴリ	訓練データ	テストデータ
スポーツ	161	147
犯罪	156	148
政府	135	142
教育システム	110	124
交通	112	103
軍事	110	118
国際関係	96	97
コミュニケーション	76	83
演劇	86	95
農業	78	72

罪，政府，教育システム，交通，軍事，国際関係，コミュニケーション，演劇，農業の 10 のカテゴリ☆について訓練記事とテスト記事，それぞれ 1,000 記事ずつ選び，実験を行った．**表 1** は各カテゴリに属する記事数を示す．

これらの記事を日本語形態素解析システム Chasen[7] によって単語分割，形態素解析を行い，20,490 語の異なり語を得た．得られた単語にはすべて品詞がつけられている．相互情報量フィルタリングの実験では全品詞を用い，品詞フィルタリングにおいては普通名詞，固有名詞，サ変名詞，未定義語，動詞の 5 品詞のみ☆☆を用いた．実験にはフィルタリング後の単語の集合を入力空間として使い，属性値は各単語の純粋な影響を見るため，単語の 1 テキスト中での出現有無を表す 1, 0 とした．

3.1.1 相互情報量フィルタリング

単語 T とカテゴリ C の間の相互情報量（MI）は式 (7) で定義される．

$$MI(T;C) = \sum_{t\in\{T,notT\}} \sum_{c\in\{C,notC\}} P(t,c)\log\frac{P(t,c)}{P(t)P(c)} \quad (7)$$

ここで $P(t)$，$P(c)$，$P(t,c)$ はそれぞれ，全記事中での単語 t を含む記事の割合，カテゴリ c に属する記事の割合，単語 t を含みかつカテゴリ c に属する記事の割合，である．相互情報量は T の出現頻度が 1 つのカテゴリ C とその他のカテゴリの間で偏りがあるときに大きな値をとる．したがって，相互情報量の高い単語はそのカテゴリにおいてキーワードになっている単語であると考えられる．

表 2 に出現単語を相互情報量の高い順に並べたときの各カテゴリにおける 300 番目，500 番目，1,000 番目，5,000 番目，10,000 番目の単語を掲げる．この表を見るとスポーツカテゴリでの「変化球」，「ゴルファー」，軍事カテゴリでの「平和」，「モスクワ」のように，1,000 番目までの単語は各カテゴリのキーワードになっている．しかし，5,000 番目以降の単語ではカテゴリ特有のキーワードはほとんど見られない．また，軍事カテゴリの中での「カザフスタン」のように品詞は未知語になっているが，高い相互情報量を持っているものがあることが分かる．

3.1.2 品詞フィルタリング

品詞フィルタリングの実験では次の 5 つの属性集合を用いた．

集合 1：普通名詞
集合 2：集合 1 + 固有名詞
集合 3：集合 2 + サ変名詞
集合 4：集合 3 + 未定義語
集合 5：集合 4 + 動詞

各品詞における異なり単語数は**表 3** のようにまとめられる．これらの異なり語の総数は 18,111 である．

3.2 実験結果

表 4 と**表 5** に相互情報量フィルタリングにおいて相互情報量の高い順に単語属性を増やしたときの SVM, C4.5 それぞれの手法における再現率と適合率の平均を示す．また**表 6** と**表 7** に品詞フィルタリングにおいて上記 5 属性がそれぞれ使われたときの SVM と C4.5 の再現率と適合率の平均を示す．いずれの場合も SVM では kernel 関数として 1 次と 2 次の多項式を使用した．また，表において太字の部分は各カテゴリにおける最高の精度を表している．

3.2.1 SVM と C4.5 の分類性能の比較

表 4 と表 5 において，相互情報量フィルタリングの結果をカテゴリ平均で見ると，C4.5 では少数の単語で最高値の精度が得られているのに対し，SVM では kernel 関数の次元によらず単語数が多いほど高い精度を示すことが分かる．具体的には C4.5 では 500 属性で急に精度が落ちるのに対し，SVM では単調に精度が高くなり属性数が 15,000 のときに平均の精度が最高になる．

一方，表 6 と表 7 から品詞フィルタリングの実験においては SVM，C4.5 両者ともカテゴリごとに最高値をとる属性が大きく違っていることが分かる．しかし SVM では品詞を順に加えていっても精度は急激には減少せず，SVM が大規模な属性空間でも有効に働くことを示している．カテゴリ平均での精度が最大になるのは 5 品詞すべての単語が使われた場合（属性集

☆ 他のカテゴリはこの 10 カテゴリに近いものが多い．
☆☆ 5 品詞に属する単語は全部で 18,111 語であった．

表 2 相互情報量によって選択された単語
Table 2 Words selected with MI.

	単語				
相互情報量（順位）	300 番	500 番	1,000 番	5,000 番	10,000 番
スポーツ	変化球	応援	ゴルファー	アンケート	目安
犯罪	疑惑	送検	地下	売る	増進
政府	議会	運輸省	約束	根幹	さえぎる
教育システム	塾	文相	理想的だ	涙	即
交通	大型車	配達	速さ	池	双方向
軍事	平和	モスクワ	カザフスタン	実際	降下
国際関係	有事	各国	大筋	年内	裁く
コミュニケーション	会議	衛星通信	伝送	正常	慎重
演劇	台本	終演	賞	要素	ロイ
農業	イモ	砂糖	飼料	改善	変貌

表 3 訓練データにおける品詞の分布
Table 3 POS distribution of training data.

	品詞				
	普通名詞	固有名詞	サ変名詞	未定義語	動詞
単語数	8,629	2,725	2,829	1,634	2,294
割合（%）	47.6	15.0	16.0	7.4	12.7

表 4 SVM での相互情報量フィルタリングによる再現率，適合率の平均値
Table 4 Average of recall and precision with MI on SVM.

	多項式関数の次元 $d = 1/d = 2$					
属性（単語）数	300	500	1,000	5,000	10,000	15,000
スポーツ	**91.9**/91.9	89.5/89.5	90.9/90.9	90.8/90.0	90.0/89.6	90.4/89.6
犯罪	71.5/70.7	69.2/71.0	68.2/70.3	72.2/73.0	74.3/74.1	75.5/**76.4**
政府	66.6/66.1	68.4/68.2	74.4/76.4	79.3/79.0	76.8/78.0	78.2/**79.8**
教育システム	68.4/68.2	69.1/69.7	71.7/73.5	78.1/77.8	80.0/79.8	**80.1**/79.6
交通	66.6/66.6	70.5/71.6	**72.1**/71.8	70.7/68.3	71.0/69.1	71.0/71.1
軍事	66.3/68.3	71.3/71.9	74.5/75.7	74.6/74.7	75.6/75.9	**77.1**/76.3
国際関係	54.3/56.9	60.1/61.9	62.9/**63.5**	61.6/60.4	61.0/59.2	57.1/58.9
コミュニケーション	64.0/64.9	65.7/**66.6**	59.3/59.3	55.7/53.3	53.6/50.0	58.2/50.0
演劇	83.9/84.0	**88.7**/83.9	86.2/88.2	83.6/86.2	83.8/82.2	83.8/82.4
農業	85.9/85.2	**87.5**/86.6	85.7/85.7	85.0/83.2	85.9/85.0	84.1/84.1
平均	71.9/72.2	74.0/74.0	74.5/75.5	75.1/74.5	75.2/74.2	**75.5**/74.8

表 5 C4.5 での相互情報量フィルタリングによる再現率，適合率の平均値
Table 5 Average of recall and precision with MI on C4.5.

属性（単語）数	300	500	1,000	5,000	10,000	15,000
スポーツ	**87.5**	86.2	85.2	83.6	83.6	83.6
犯罪	67.9	**70.8**	68.9	68.8	68.8	68.8
政府	**65.5**	63.0	58.0	57.9	57.9	57.9
教育システム	**72.0**	69.2	70.1	70.1	70.1	70.1
交通	**63.0**	61.0	61.0	61.0	61.0	61.0
軍事	**75.9**	73.3	69.1	68.8	68.8	68.8
国際関係	**50.0**	45.6	42.4	42.4	42.4	42.4
コミュニケーション	**52.7**	50.3	50.3	50.3	50.3	50.3
演劇	**80.9**	80.9	79.5	79.5	79.5	79.5
農業	**84.4**	84.4	84.4	83.8	83.8	83.8
平均	**70.0**	68.5	66.9	66.6	66.6	66.6

表 6 SVM での品詞フィルタリングによる再現率，適合率の平均値
Table 6 Average of recall and precision with POS filtering on SVM.

	多項式関数の次元 $d = 1/d = 2$				
属性	集合 1	集合 2	集合 3	集合 4	集合 5
スポーツ	92.2/91.4	**93.2**/92.4	92.9/92.4	92.0/92.0	90.5/90.8
犯罪	74.0/73.0	73.3/73.6	72.5/73.3	73.0/73.3	**75.2**/74.9
政府	76.9/76.7	78.4/78.7	79.3/79.0	78.9/78.2	**79.6**/79.2
教育システム	81.4/81.4	80.8/80.3	81.4/80.9	81.4/**81.8**	81.2/80.3
交通	72.8/72.0	76.0/74.5	74.8/**76.0**	74.8/75.2	73.0/72.2
軍事	80.1/77.3	76.1/76.1	78.8/76.2	77.0/77.0	**80.1**/77.9
国際関係	54.5/54.6	59.2/60.2	60.7/61.4	61.2/62.2	**64.0**/64.0
コミュニケーション	65.7/67.6	63.8/62.3	**69.3**/68.0	68.9/67.1	65.7/63.2
演劇	83.8/83.8	82.4/82.4	85.2/85.2	85.2/85.2	**87.0**/85.0
農業	87.5/87.5	**88.3**/88.3	87.5/86.6	86.6/86.6	85.0/84.8
平均	76.9/76.5	77.5/77.2	78.0/77.9	77.9/77.8	**78.1**/77.2

表 7 C4.5 での品詞フィルタリングによる再現率，適合率の平均値
Table 7 Average of recall and precision with POS filtering on C4.5.

属性	集合 1	集合 2	集合 3	集合 4	集合 5
スポーツ	**84.7**	82.9	83.4	83.0	83.4
犯罪	61.5	59.3	**71.3**	71.3	71.3
政府	58.0	**62.8**	62.7	62.7	60.4
教育システム	60.2	63.5	62.8	**70.2**	70.2
交通	58.2	56.4	58.1	58.1	**59.4**
軍事	**75.5**	71.8	71.8	71.8	71.8
国際関係	**49.3**	44.1	48.9	46.4	46.4
コミュニケーション	49.6	48.5	**51.0**	44.6	44.6
演劇	**79.7**	71.3	79.2	79.2	79.2
農業	81.2	**81.4**	81.4	81.4	81.4
平均	65.8	63.9	**67.1**	66.9	66.8

合 5）である．

SVM では相互情報量フィルタリング，品詞フィルタリングどちらの実験においてもカテゴリごとに振舞いは異なるが，カテゴリ平均で見ると単語数が多いほど精度が高くなることが分かる．これは C4.5 では単語数が多くなると精度が下がってしまうのとは対照的である．

さらに同じデータにおける適合率，再現率について詳しく考察する．**図 1** は相互情報量フィルタリングの場合の $d = 1$ の SVM（上）と C4.5（下）における適合率である．C4.5 では適合率は 300 属性，500 属性といった小さな属性数のときに最高値を示し，それ以上属性数を増やしても精度が下がった後，上昇しない．一方，SVM の適合率は国際関係カテゴリが属性数 15,000 で悪くなっているのを除けば，属性数が多くなるとすべてのカテゴリで増加傾向を示している．

図 2 は相互情報量フィルタリング実験における $d = 1$ の SVM（上）と C4.5（下）における再現率を示したものである．C4.5 では適合率と同様，1,000 属性以降では再現率は低いままである．C4.5 では，相互情報量が 1,000 番目以降の単語はカテゴリ間の特徴の違いを認識する力がないことを表している．一方，SVM での再現率の変化は適合率の場合とは異なり複雑である．しかし，属性数を増やしても再現率は急激に下がってはいない．これらの結果より 15,000 属性といった大規模な属性数でも SVM の汎化能力が高いのは適合率が高く，再現率もそれほど低くならないためであることが分かる．

図 3，**図 4** は品詞フィルタリング実験における SVM と C4.5 の適合率，再現率を各々プロットしたものである．一見，相互情報量フィルタリングの場合とは挙動が大きく異なっているように見える．これは集合 1（普通名詞）の数が 8,629 あり，個数だけで見ると図 2 における 10,000 属性付近に相当することによる．個数だけで比較すれば，品詞フィルタリングと相互情報量フィルタリングの傾向にそれほど違いはない．

以上のように，2 つのフィルタリング実験から SVM が C4.5 より大規模な単語属性を扱うのに適していることが分かる．

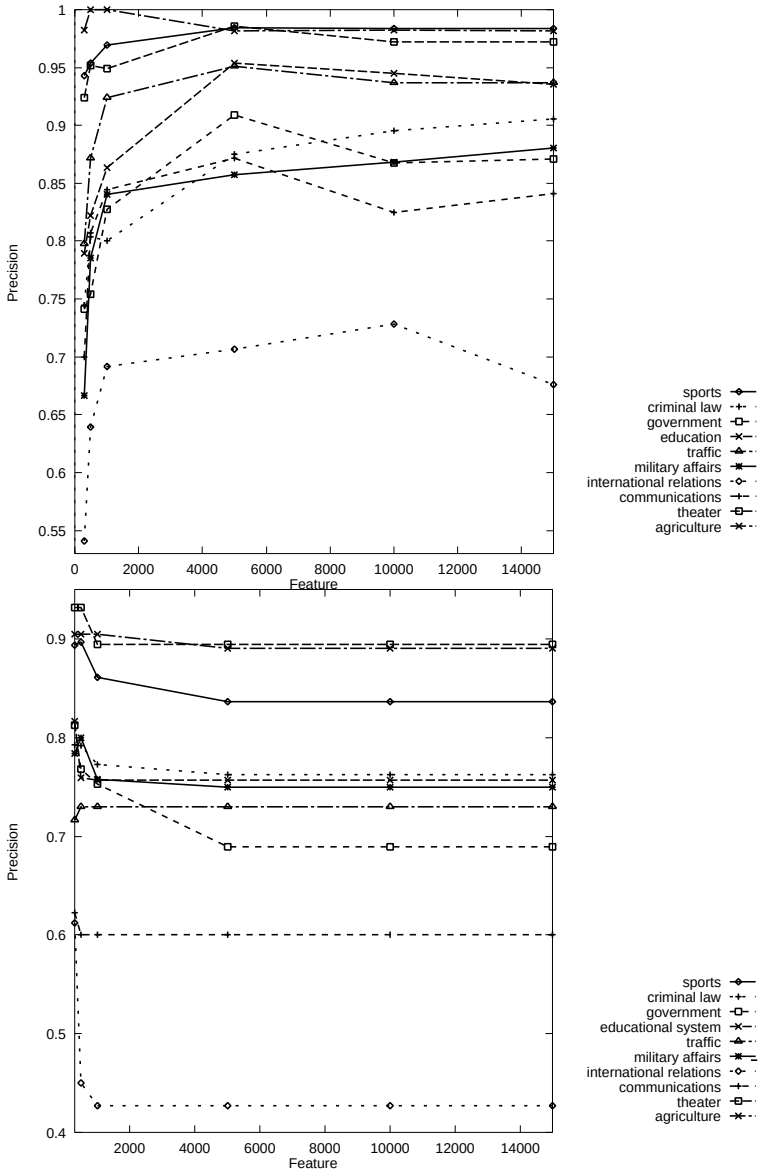

図 1 相互情報量フィルタリング使用時の適合率〔SVM（上），C4.5（下）〕
Fig. 1 Precision with MI features on SVMs (top) and on C4.5 (bottom).

3.2.2 両フィルタリングの比較

表 4 と表 6 より SVM におけるカテゴリ平均での精度の最高値（相互情報量フィルタリング 15,000 属性と品詞フィルタリング集合 5）を比較すると相互情報量フィルタリングより品詞フィルタリングの方が精度が高いことが分かる．これは，相互情報量だけでは分類に無関係な品詞の単語を除くことができないことを意味している．C4.5 の場合（表 5，表 7）では，単語属性数の近い，相互情報量フィルタリングにおける 10,000 属性，15,000 属性と品詞フィルタリングにおける集合 1〜5 とを比較すると，精度に目立った差はない．これは C4.5 においては単語属性の増加による過学習の影響が大きく，品詞フィルタリングの効果が打ち消されているためだと考えられる．SVM では大規模な単語属性数でも高い精度が得られるが，さらに精度を高めるために品詞フィルタリングが有効であることが分かった．

また，表 6 から SVM を使ったテキスト分類において各品詞がどのような役割を果たしているかが読みとれる．普通名詞は他の品詞を一緒に使わなくても最高

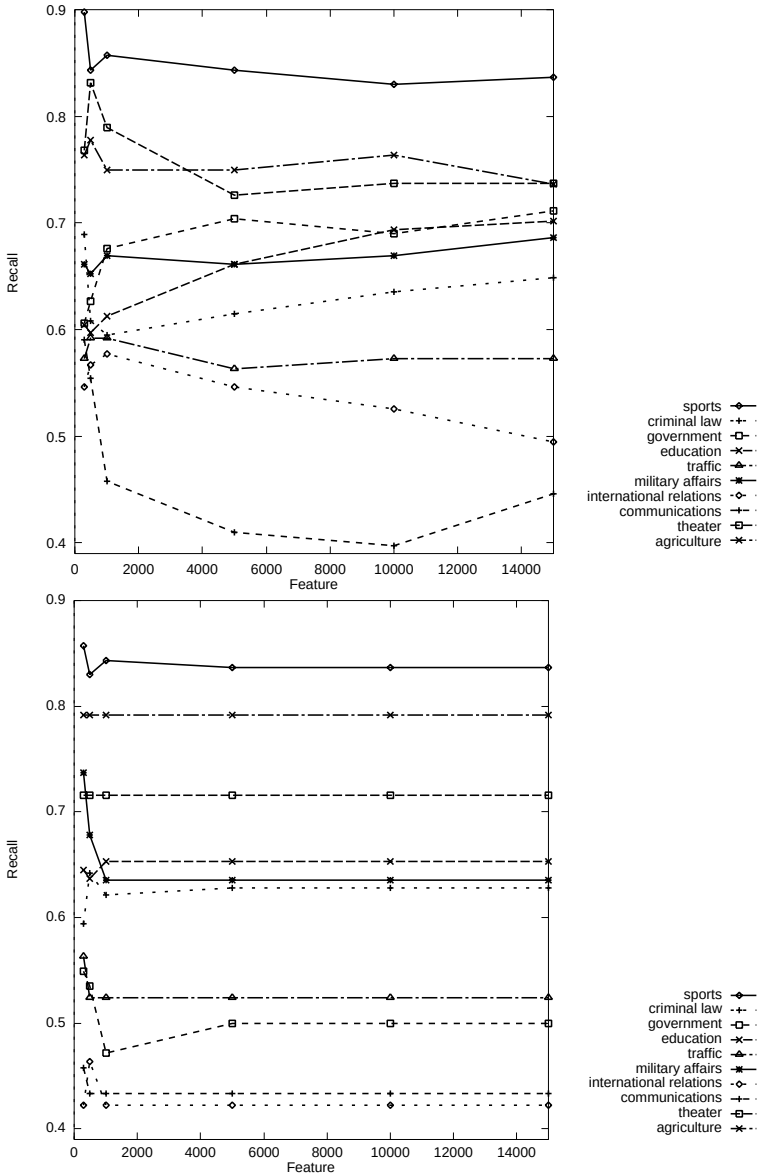

図 2 相互情報量フィルタリング使用時の再現率〔SVM（上），C4.5（下）〕
Fig. 2 Recall with MI features on SVMs (top) and on C4.5 (bottom).

精度に近い精度が得られるほど大きな役割を担っている．固有名詞，サ変名詞，動詞は半数以上のカテゴリで精度向上に役立っている．未定義語は 3 カテゴリで寄与がある．これは，カテゴリによっては未定義語中に重要なキーワードが含まれていることによると考えられる．このような品詞ごとの分類への効果は従来の学習手法では過学習による精度の低下に隠され分からなかったことで興味深い．

4. 考　　察

相互情報量フィルタリングと品詞フィルタリングに注目し SVM を使ったテキスト分類における属性選択の影響について述べた．その結果，1) カテゴリごとに最適な属性集合が異なるが，2) 平均の最高精度は SVM にすべての属性が与えられたときに得られる，ことが分かった．

また，個々のフィルタリング手法に関しては，

相互情報量フィルタリング　最適な属性数はカテゴリ

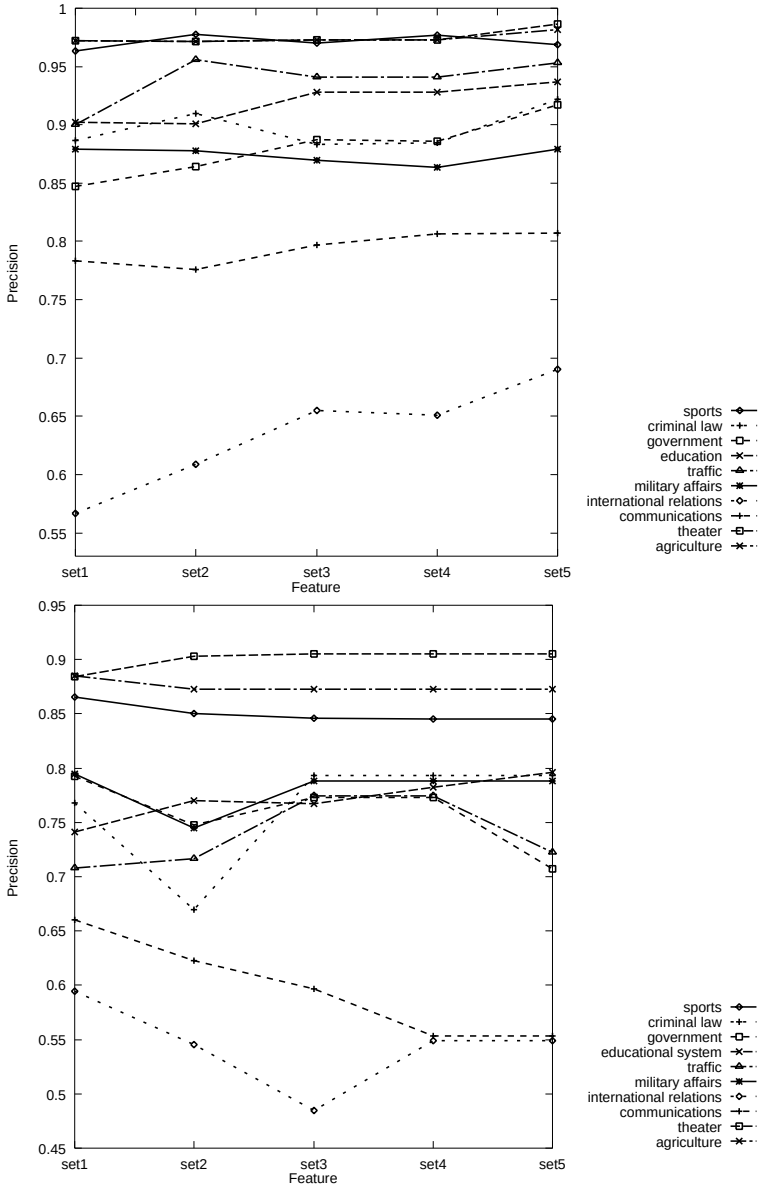

図 3 品詞フィルタリング使用時の適合率〔SVM（上），C4.5（下）〕
Fig. 3 Precision with POS features on SVMs (top) and on C4.5 (bottom).

ごとに大きく異なり先験的に決定するのは難しい．平均での最高精度はすべての単語を使ったときに得られる．

品詞フィルタリング 最適な属性集合はカテゴリによって異なる．平均での最高精度は普通名詞，固有名詞，サ変名詞，未定義語，動詞の 5 品詞の単語を使ったときに得られる．各品詞が精度向上に寄与するがカテゴリごとに影響は異なる．

ということが分かった．2 つのフィルタリングを比較すると表 4，表 6 を比較して分かるように品詞フィルタリングがつねに相互情報量フィルタリングの性能を上回っている．ここで相互情報量フィルタリングでは助詞，接続詞など全品詞を含んでいるが，品詞フィルタリングでは普通名詞，固有名詞，サ変名詞，未定義語，動詞の 5 品詞だけを選択していることに注意したい．つまりテキスト分類において SVM は高い汎化能力を持つが無関係な品詞の検出まではできないことが分かる．

最後に，kernel 関数について簡単に述べる．画像認識のような問題[1)]においては kernel 関数はきわめて重

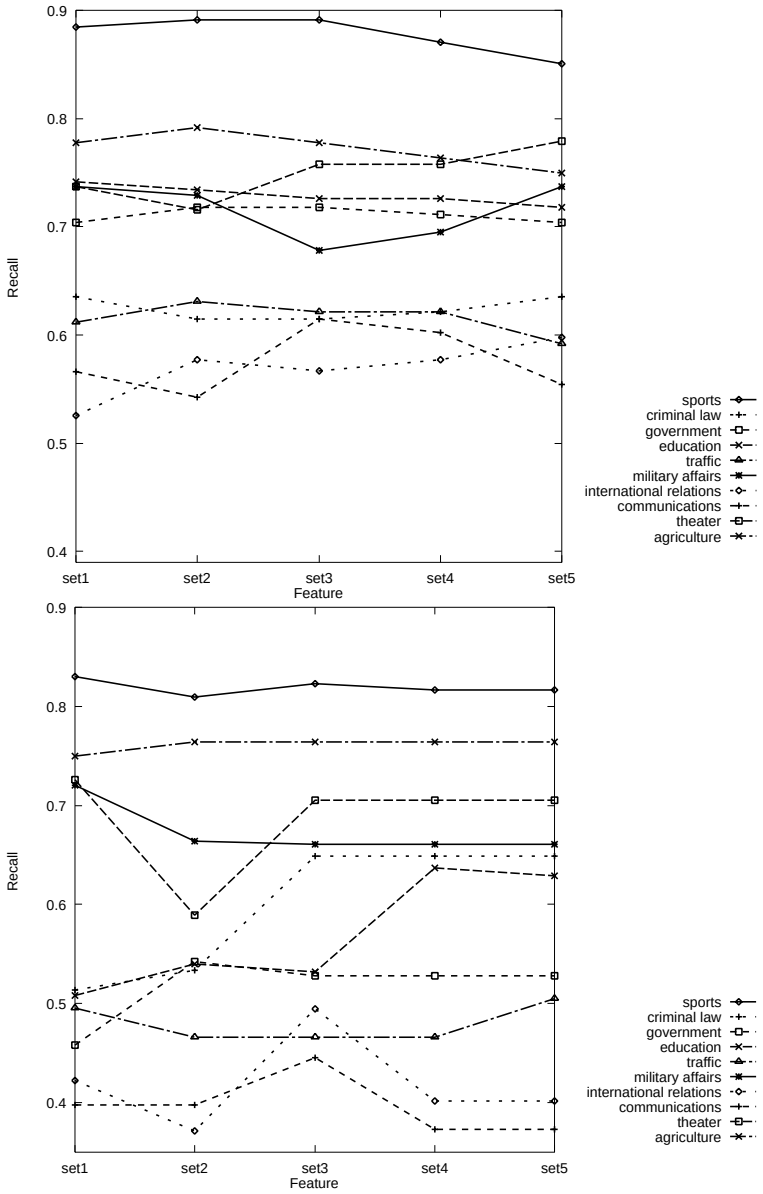

図 4 品詞フィルタリング使用時の再現率〔SVM（上），C4.5（下）〕
Fig. 4 Recall with POS features on SVMs (top) and on C4.5 (bottom).

要な役割を果たしている．我々のテキスト分類の実験でも kernel 関数に多項式関数を選び次元を変化させてみたが，精度に明確な違いは表れなかった．

5. 結　　論

本論文では，SVM を用いたテキスト分類における属性選択の効果について述べた．実験の結果，SVM は大規模な単語属性を適切に扱えるが無関係な品詞を除外する能力には限界があることが明らかになった．これは，品詞選択のような簡単なフィルタリング手法で得られた単語すべてを使うという，単純で実際的な手法が有効であることを示している．SVM などの large margin classifier による手法は今後複雑な自然言語処理の問題を扱う際にますます重要な役割を果たすものと思われる[4]．

謝辞 毎日新聞 94 年版の使用に関して，記事データの研究利用許諾をいただいた毎日新聞社に感謝いたします．

参 考 文 献

1) Cortes, C. and Vapnik, V.: Support Vector Networks, *Machine Learning*, Vol.20, pp.273–297 (1995).
2) Cover, T. and Thomas, J.: *Elements of Information Theory*, John Wiley & Sons (1991).
3) Dumais, S., Platt, J., Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization, *Proc. 7th International Conference on Information and Knowledge Management* (1998).
4) Haruno, M., Shirai, S. and Ooyama, Y.: Using Decision Trees to Construct a Practical Parser, *Machine Learning*, Vol.34, pp.131–149 (1999).
5) Joachims, T.: Text Categorization with Support Vector Machines, *Proc. European Conference on Machine Learning* (*ECML*) (1998).
6) Lewis, D. and Ringuette, M.: A Comparison of Two Learning Algorithms for Text Categorization, *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp.81–93 (1994).
7) Matsumoto, Y., Kitauchi, A., Yamashita, T., Imaichi, O. and Imamura, T.: Japanese Morphological Analysis System Chasen Manual, NAIST Technical Report NAIST-IS-TR97007 (1997).
8) Osuna, E., Freund, R. and Girosi, F.: Training Support Vector Machines: An Application to Face Detection, *Proc. Computer Vision and Pattern Recognition '97*, pp.130–136 (1998).
9) Quinlan, J.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).
10) Salton, G. and Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol.24, No.5, pp.513–523 (1988).
11) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag, New York (1995).
12) Yang, Y.: Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval, *Proc. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.13–22 (1994).
13) 情報科学技術協会：国際十進分類法，日本語中間版第 3 版，丸善 (1994).
14) 豊浦 潤, 徳永健伸, 井佐原均, 岡 隆一：RWC における分類コード付きテキストデータベースの開発，電子情報通信学会研究報告，NLC96-13, pp.27–32 (1996).

(平成 11 年 6 月 10 日受付)
(平成 12 年 2 月 4 日採録)

平 博順（正会員）

1994 年東京大学理学部卒業．1996 年同大学院修士課程修了．同年日本電信電話（株）入社．同社コミュニケーション科学基礎研究所研究員．機械学習による自然言語処理の研究に従事．

春野 雅彦（正会員）

1991 年京都大学工学部電気工学第二学科卒業．1993 年同大学院修士課程修了．1998 年奈良先端科学大学院大学博士後期課程修了．博士（工学）．1993 年日本電信電話（株）入社．1997 年まで同社コミュニケーション科学研究所研究員．1997 年より ATR 人間情報通信研究所研究員．計算論的神経科学，機械学習，自然言語処理に興味を持つ．Society for Neuroscience，ACL 各会員．