D-029

# 相互情報量に基づく特徴選択を用いた文書自動分類
# Text Classification Using Feature Selection Based on Mutual Information

津田 裕一 † 山岸 英貴 † 石田 崇 † 平澤 茂一 †
Yuichi Tsuda Hidetaka Yamagishi Takashi Ishida Shigeichi Hirasawa

## 1. はじめに

近年，膨大かつ多種多様な電子化文書が氾濫しているため，文書の効率整理や，不要な情報のフィルタリング技術へのニーズが高まり，文書の自動分類の研究が盛んに行われている [1].

文書分類とは与えられた文書を既存のカテゴリに自動的に割り当てる技術である．文書分類において，分類に用いる効果的な単語を選択する処理は重要である．これを特徴選択といい，この指標として一般に相互情報量が用いられる．相互情報量による特徴選択は，カテゴリに偏って出現する単語を選択するが，選択された単語中には，分類に用いる単語として適切でない場合がある [2].

そこで本研究では，一つのカテゴリに限定して出現する単語をより優先的に選択することにより，分類性能を改善する手法を提案する．また，本手法を新聞記事データ [5] に適用し，その有効性を示す.

## 2. 従来の文書自動分類方法

### 2.1 ベクトル空間モデルに基づく文書自動分類手法

文書自動分類方法には様々な手法が提案されているが，今回は最も代表的な手法であるベクトル空間モデルに基づく手法 [4] を考える．ベクトル空間モデルに基づく文書自動分類手法では，文書を形態素解析し，単語の切り出しを行い, 特徴選択を行う. 本論文では 2.1 で説明する相互情報量を用いて特徴選択を行っている. また，文書自動分類手法は学習フェーズと分類フェーズがあり，前者は 2.1.1，2.1.2 で説明し，後者は 2.1.3 で説明する.

### 2.2 相互情報量に基づく特徴選択

単語を $x_i \in \mathcal{T}$ $(1 \leq i \leq I)$，カテゴリを $c_k \in \mathcal{C}$ $(1 \leq k \leq K)$ とする. 単語 $x_i$ とカテゴリ間の相互情報量 $MI(x_i, C)$ を以下のように定義する.

**定義 1 (相互情報量)**

$$MI(x_i, C) = \sum_{k=1}^{K} P(x_i, c_k) \log \frac{P(x_i, c_k)}{P(x_i)P(c_k)} \tag{1}$$

$P(x_i, c_k)$ : 全文書中で単語 $x_i$ を含み，かつカテゴリ $c_k$ に属する文書の割合
$P(x_i)$ : 全文書中で単語 $x_i$ を含む文書の割合
$P(c_k)$ : 全文書中であるカテゴリ $c_k$ に属する文書の割合 □

相互情報量が大きな値を取る単語は，カテゴリ間でその単語の出現文書数に偏りがあり，かつ出現文書数の多い単語といえる．相互情報量を降順に並べ，その上位の単語を分類に用いる単語集合を $\mathcal{T}_s$ $(\mathcal{T}_s \subset \mathcal{T})$ とする.

#### 2.2.1 カテゴリの特徴ベクトルの作成

特徴選択により選択した単語を $y_j \in \mathcal{T}_s$, $\mathcal{T}_s = \{x_{i_1}, x_{i_2}, \ldots, x_{i_J}\} = \{y_1, y_2, \ldots, y_J\}$ $(J < I)$ としたとき，カテゴリ $c_k$ の特徴ベクトル $\boldsymbol{g}_{\boldsymbol{c_k}}$ は以下のように定義される.

**定義 2 (カテゴリの特徴ベクトル)**

$$\boldsymbol{g}_{\boldsymbol{c_k}} = \left( \frac{N_{y_1 c_k}}{M_{c_k}}, \frac{N_{y_2 c_k}}{M_{c_k}}, \frac{N_{y_3 c_k}}{M_{c_k}}, \ldots, \frac{N_{y_J c_k}}{M_{c_k}} \right) \tag{2}$$

$N_{y_j c_k}$ : カテゴリ $c_k$ における単語 $y_j$ の出現文書数
$M_{c_k}$ : カテゴリ $c_k$ の文書数 □

†早稲田大学大学院理工学研究科経営システム工学専攻

#### 2.2.2 新規文書の分類

新規文書についても各文書の特徴ベクトルを作成し，カテゴリ $c_k$ の特徴ベクトル $\boldsymbol{g}_{\boldsymbol{c_k}}$ をもとに，以下の手順で新規文書を分類する.

1. 1) 新規文書について形態素解析を行い，各単語について $Z_{y_j}$ を以下のようにする.

$$Z_{y_j} = \begin{cases} 1 & (\text{新規文書に単語 } y_j \text{が出現する}) \\ 0 & (\text{新規文書に単語 } y_j \text{が出現しない}) \end{cases} \tag{3}$$

2. 2) $Z_{y_j}$ から新規文書の特徴ベクトル $\boldsymbol{u}$ を構成する.

$$\boldsymbol{u} = (Z_{y_1}, Z_{y_2}, Z_{y_3}, \ldots, Z_{y_J}) \tag{4}$$

3. 3) 全てのカテゴリ $c_k$ に対して，$\boldsymbol{u}$ と各特徴ベクトル $\boldsymbol{g}_{\boldsymbol{c_k}}$ との内積値 $S_{c_k} = < \boldsymbol{u} \cdot \boldsymbol{g}_{c_k} >$ を求める.
4. 4) $S_{c_k}$ の最も大きいカテゴリに新規文書を分類する.

### 2.3 従来手法の問題点

相互情報量に基づく特徴選択では，単語の出現頻度の偏りの大きいものが抽出される傾向がある. ただし，必ずしも上位の単語がカテゴリを特徴づける単語とは限らない. なぜならば，出現頻度の多い単語では複数のカテゴリに同程度の頻度で出現する場合でも，出現頻度の偏りが大きいので抽出されてしまう. また，一方で，下位の単語でもカテゴリを特徴づけているものが存在すると考えられる. これは，全体的な出現頻度が比較的少なくても 1 つのカテゴリに集中して出現するものがあると考えられるためである.

表 1: 単語出現数の例

| 選択順位 | 単語 | 経済 | 家庭 | 芸能 | スポーツ |
|---|---|---|---|---|---|
| 1 | 観客 | 3 | 2 | 71 | 77 |
| 1000 | 野球 | 2 | 1 | 3 | 24 |

表 1 は相互情報量により選択された順位を示した例である．表 1 を見ると「観客」の順位は高いが，芸能とスポーツ両カテゴリにおいて出現する単語数が同等であり，特定のカテゴリを特徴付けている単語といえない. 一方「野球」は特定のカテゴリを特徴付けている単語といえるが，順位が低いため分類に用いる単語として選択されない可能性がある.

## 3. 提案手法

本研究では，上で述べた問題を解決するために，分類率向上を目的とした，新たな単語の抽出法を提案する. これにより，カテゴリ間の距離を大きくすることができ, 分類精度を向上することが期待できる. なお提案手法では，カテゴリベクトルの作成法，新規文書の分類法は 2.1 節と同様である.

### 3.1 分類多岐語

相互情報量による特徴選択の目的は，あるカテゴリに特徴的に出現する単語を選択することである．しかし，選択された単語の中には，特定の複数のカテゴリにおいて頻出する単語が含まれ，こうした単語は分類誤りの原因となる可能性がある [3]. このような単語を，以下に定義する $V(y_j)$ 値を用いて，分類多岐語 $y_{j*}$ と定義する.

**定義 3 (分類多岐語)**

$V(y_j)$ 値を

$$V(y_j) = \left(\frac{1}{\sum_k n_{y_j c_k}}\right)^2 \left(n^{(1)}_{y_j c_l} - n^{(2)}_{y_j c_m}\right)^2 \quad (5)$$

とする．ただし，

$n^{(1)}_{y_j c_l}$ : ある単語 $y_j$ について最も出現単語数が高いカテゴリ $c_l$ での出現単語数

$n^{(2)}_{y_j c_m}$ : ある単語 $y_j$ について 2 番目に出現単語数が高いカテゴリ $c_m$ での出現単語数

$\sum_k n_{y_j c_k}$ : ある単語 $y_j$ の総出現単語数 □

このとき $V(y_j)$ が 閾値$\varepsilon$ 以下の単語を分類多岐語 $y_j*$ と定義する．

### 3.2 提案アルゴリズム

step1[相互情報量の計算]

単語を相互情報量の降順に並べる．

step2[分類多岐語の削除]

すべての単語 $I$ 個から $V(y_j) < \varepsilon$ となる分類多岐語 $y_j*$ を削除する．

step3[単語の選択]

相互情報量の上位単語から $J$ 個を用いる単語として選択する．

## 4. 評価実験

### 4.1 実験方法

実験データは毎日新聞の 1 年分の記事データ [5] を利用した．今回の実験では用いるカテゴリ数を 9 カテゴリとし，学習文書を 9000 件，テスト文書を 4500 件, 単語として名詞だけを用い，全単語数 $I$ は 42721 であった．また，今回の実験では予備実験を行い，選択する単語数 $J$ を 4000 単語とした．なお，削除多岐語数 0 は従来手法と同等である. 提案手法の閾値 $\varepsilon$ は 0.002 刻みで 0.001 から 0.019 まで変化させ (このときの削除多岐語数は 312 〜 1390 に対応している)，正解率，$F$ 値の変化を調べた．実験結果を図 1 と表 2 に示す.

### 4.2 評価方法

評価方法は以下の評価基準を用いる．

$$\text{正解率} = \frac{\text{正分類数}}{\text{全テスト文書数}} \qquad F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

### 4.3 実験結果

図 1 は閾値を変化させた時の精度を表したものである．削除多岐語数 0 は従来手法の精度を表している．表 2 は図 1 において最も高い精度を取った時の正解率・平均 $F$ 値の比較である．

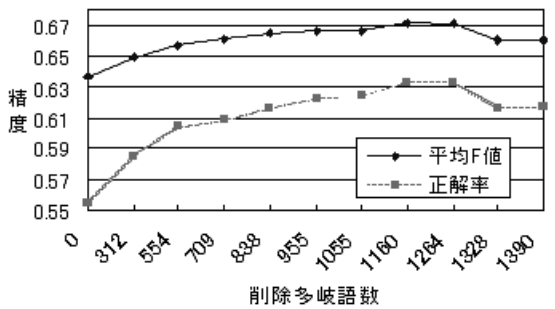


図 1: 削除多岐語数による正解率・平均 $F$ 値の変化

表 2: 正解率・平均 $F$ 値の比較

| | 従来手法 | 提案手法 (削除多岐語数 1264) |
|---|---|---|
| 正解率 | 0.556 | 0.634 |
| 平均 $F$ 値 | 0.637 | 0.672 |

### 4.4 考察

1) 表 2 より正解率・平均 $F$ 値が向上し，提案手法の有効性が示された．従来手法と提案手法で最も有効性が見られたとき (削除多岐語数 1264 語) で各カテゴリの特徴ベクトル同士の類似度を見ると, 表 3 のように，提案手法は従来手法に比べて小さい. これは，提案手法により各カテゴリ間の距離が大きくなるような特徴ベクトルを構成できたことを示しており，また精度も上がっているので，提案手法での選択単語によるカテゴリ重心ベクトルが真のカテゴリ特徴ベクトルに近づいたと思われる．

表 3: 従来手法と提案手法のカテゴリ特徴ベクトル間の類似度

| | 従来手法 | 提案手法 |
|---|---|---|
| 最小値 | 0.466 | 0.189 |
| 最大値 | 1.549 | 0.708 |
| 平均値 | 0.782 | 0.343 |

2) 表 4 に従来・提案手法における正解・不正解文書数を示す. また，従来手法では誤分類されたが, 提案手法で正しく分類できた 361 文書について，提案手法において 1 番目に高いカテゴリとの類似度と, 2 番目に高いカテゴリとの類似度の差の最小値 (0.000877)，最大値 (3.77)，平均値 (0.33) を調べた. 提案手法では，従来手法で誤分類された文書に対し，単に正しく分類されるだけでなく，平均値が最大値に近いので，最小値に対しこの差が大きく，誤分類の可能性が小さいことがわかる. これは，提案手法での選択単語が分類に効果的な単語であったことを示している．

表 4: 従来・提案手法における正解・不正解文書数

| | | 従来不正解 | 従来正解 |
|---|---|---|---|
| 提案 (削除多岐語数 1264) | 不正解 | 1396 | 53 |
| | 正解 | 361 | 2144 |

3) 図 1 より，削除する多岐語数を増やすと，1264 語までは正解率・平均 $F$ 値が向上するが，それ以上では下がっている．これは，分類多岐語にも分類に有効な単語があるにも関わらず，削除されてしまったからであると考えられる．

4) 閾値が小さい場合は精度の低下は見られなかったが，閾値の設定方法に関しては今後検討が必要である．

5) 提案手法では，分類多岐語数分だけを $V(y_j)$ 値を求める計算量が増えることになるが，$V(y_j)$ ひとつの計算量は非常に小さく，全分類多岐語の計算時間は実用面には問題はないと思われる．

## 5. まとめと今後の課題

本研究では，相互情報量を用いてカテゴリを特徴付ける単語を抽出し分類多岐語を削除し，分類率を向上させる手法を提案した．今後は本手法を他の分類器にも適用し，その有効性を検討していきたい．

## 参考文献

[1] 徳永健伸,「情報検索と言語処理」, 財団法人東京大学出版会, 1999.

[2] 相澤彰子, “ 低頻度語の利用によるテキスト分類性能の改善と評価 ”, 情報処理学会論文誌, Vol.44, No.7, pp.1720-1730, 2003.

[3] 荒木淳，中村文隆，中山雅哉，”多義性を軽減した素性セットによるテキスト分類方式 ”, 自然言語処理, 85-92, 2003.3.7

[4] 呉勇, 山田祥, 岸本陽次郎, “ 名詞頻度を使った分類用辞書の構築と評価 ”, 電子情報通信学会論文誌, No.2 , Vol.J84-D-I, pp.213-221, 2001.

[5] 毎日新聞社, CD-毎日新聞'94, 日外アソシエーツ, 1995.