文書分類問題におけるカテゴリに注目した可変長 N グラム法

1X08C016-3　井上大樹
指導教員　後藤正幸

1　研究背景・目的

近年，情報化の進展により，World Wide Web，電子メール，電子書籍など大量のテキスト文書を扱う機会が増加している．これらから効率的に情報を獲得する技術として，高性能な文書自動分類法が注目されている．

文書分類の性能向上において，入力文書のカテゴリの推定には，主に素性選択と分類器の構築の 2 つの要素が重要である [1]．本研究においては，素性選択に注目する．単語の出現順序も考慮した手法として，N 個の単語の組を素性とする単語 N グラム [2] があげられる．しかし，最適な N の値は未知であるとともに，単語や文脈によって最適な N が変化するという問題がある．そこで，N グラムの N の値を事前に定めず，入力文書に対して，事前情報を特徴づける N を文書中の出現系列ごとに決定していく可変長 N グラムが提案された [3]．この可変長 N グラムを文書分類に適用した手法として，相澤による手法 [4] がある．この手法では，カテゴリが既知である訓練文書を事前情報として用い，Ziv-Merhav Crossparsing[3](以下 ZM 法) によって，可変長 N グラムに系列分解する．そして，系列分解された入力文書から，各カテゴリの事後確率を計算し，入力文書を分類している．

この相澤による手法では，可変長 N グラムによる系列分解において，訓練文書全体を事前情報として系列分解を行っている．しかし，カテゴリに所属する文書中に出現する単語列はカテゴリ毎に特徴が異なると考えられる．一般にカテゴリの特徴を捉えた素性を選択することが分類精度向上には重要であるが，相澤による手法は分類するカテゴリの特徴を捉えた素性が選択されないという問題点がある．

そこで，本研究では相澤による手法において，カテゴリを考慮した素性選択をすることにより，分類精度を高める文書分類手法を提案する．提案手法の有効性を示すため，実際の新聞記事の分類問題に適用し，分類精度が向上することを検証する．

2　従来手法

従来手法として，相澤による手法 [4] では ZM 法 [3] により，入力文書を可変長の単語 N グラムに系列分解を行う．そして，可変長 N グラムを素性とし，ナイーブベイズ法により各カテゴリ毎に事後確率を求める．従来手法の概念を図 1 に示す．

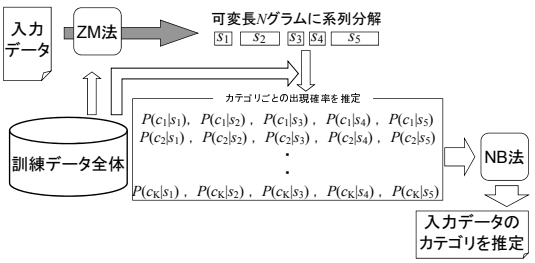

図 1. 従来手法の概念図

2.1　Ziv-Merhav Crossparsing

ZM 法において，訓練文書集合 $\mathcal{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_J\}$ を用いて新たな入力文書 $\boldsymbol{x} = x_1 x_2 \cdots x_M$ を系列分解する．訓練文書を，$\boldsymbol{d}_j = y_{j1} y_{j2} \cdots y_{jR_j}$ とする．ただし，x_m, y_{jr} は，文書中の素性を表し，出現順に並んでいるものとする．本研究において，x_m, y_{jr} は単語を表す．M, R_j は，各文書中の総単語数を表している．q_j は入力文書と訓練文書において，一致する単語列の単語数を表す．

以下に，具体的な ZM 法のアルゴリズムを示す．

［アルゴリズム］

Step1　$m = 1$，$i = 1$ とする．

Step2　$j = 1, 2, \cdots, J$ の全ての訓練文書 d_j に対し，単語列 $y_{j1} y_{j2} \cdots y_{jRj}$ と単語列 $x_m x_{m+1} \cdots$ と比較し，最長一致する単語数を q_j とする．

Step3　$Q = \max_j q_j$ とする．

Step4　$Q > 0$ のとき $\boldsymbol{s}_i = x_m \cdots x_{m+Q-1}$，$m = m + Q$，$i = i + 1$ とする．$m + Q = M$ のとき，アルゴリズムを終了．さもなければ，$m = m + 1$ として Step2 へ．

一般的に 1 単語が素性として，よく用いられるが，本研究において，連続した N 個の単語の組 $\boldsymbol{s}_i$ を素性として用いる．アルゴリズムでは，入力文書 $\boldsymbol{x}$ を先頭から順に，訓練文書の単語列と最長一致した N 個の単語列を素性 $\boldsymbol{s}_i$ とすることにより，可変長 N グラムへの系列分解を行う．

2.2　ナイーブベイズ法

ナイーブベイズ法（以下，NB 法）は，代表的な確率的分類手法である．入力文書 $\boldsymbol{x}$ は ZM 法により，$\boldsymbol{x} = \boldsymbol{S} = \boldsymbol{s}_1 \boldsymbol{s}_2 \cdots \boldsymbol{s}_I$ と I 個に系列分解された単語列を素性として定義する．そして，各カテゴリは $\mathcal{C} = \{c_1, c_2, \cdots, c_K\}$ の K 個のカテゴリが割り当てられる．$\boldsymbol{S}$ が与えられた際の各カテゴリ c_k における事後確率は，各単語列の独立性を仮定した場合，ベイズの定理から (1) 式を用いて算出可能となる．

$$P(c_k|\boldsymbol{S}) = P(c_k) \prod_{i=1}^{I} \frac{P(\boldsymbol{s}_i|c_k)}{P(\boldsymbol{s}_i)} = P(c_k) \prod_{i=1}^{I} \frac{P(c_k|\boldsymbol{s}_i)}{P(c_k)} \quad (1)$$

$P(\boldsymbol{s}_i|c_k)$ の算出には膨大な計算量が必要なため，$P(c_k|\boldsymbol{s}_i)$ を用いる．また，$\boldsymbol{x}$ のカテゴリを予測する際に，訓練文書に含まれない素性が入力文書に含まれることがある．その場合，$P(c_k|\boldsymbol{S}) = 0$ となり，カテゴリが推定できないという問題が発生する．このため，相澤による手法 [4] では，(2) 式を用いて，スムージングを行っている．

$$P(c_k|\boldsymbol{s}_i) = \frac{c_k \text{における} \boldsymbol{s}_i \text{の頻度} + 1}{\text{全訓練文書における} \boldsymbol{s}_i \text{の頻度} + K} \quad (2)$$

(3) 式により，$c_{\hat{k}}$ を入力文書 $\boldsymbol{x}$ のカテゴリと判別する．

$$\hat{k} = \underset{k}{\operatorname{argmax}} \left\{ \log P(c_k) + \sum_{i=1}^{I} \log \frac{P(c_k|\boldsymbol{s}_i)}{P(c_k)} \right\} \quad (3)$$

2.3　従来手法の問題点

従来手法においては，訓練文書全体を事前情報とし $\boldsymbol{S}$ への系列分解が行なわれる．入力文書 $\boldsymbol{x}$ には，訓練文書全体に出現する一般的な単語列が含まれる場合がある．このとき，入力文書 $\boldsymbol{x}$ は所属カテゴリ以外からの単語列と最長一致し，系列分解が行われる可能性がある．正しい所属カテゴリ以外の単語列を素性として確率計算をした場合，所属カテゴリにおいて低い確率が掛け合わせられることになり，正しく分類されない．よって，入力文書 $\boldsymbol{x}$ は訓練文書全体に出現する単語列から系列分解することが問題となる．

3　提案手法

提案手法では ZM 法をカテゴリ毎に適用させることで，カテゴリ毎で可変長の単語 N グラムが選択される．そして，カテゴリ c_k から系列分解した単語列を素性とし，各カテゴリの事後確率が最大となるカテゴリに分類する．提案手法の概念を図 2 に示す．

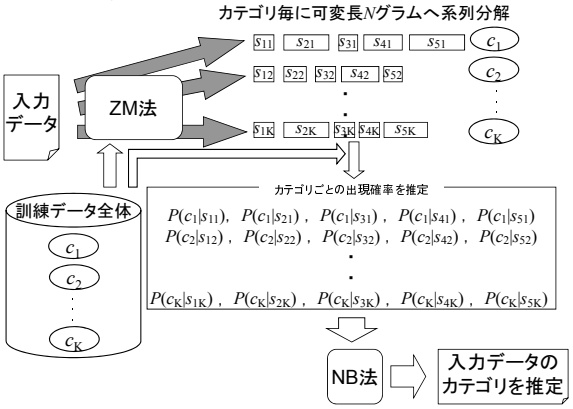

図 2．提案手法の概念図

3.1　カテゴリを考慮した系列分解

提案手法ではカテゴリ毎に ZM 法を適用させることにより，$\boldsymbol{S}$ は各カテゴリで系列分解した素性を選択する．カテゴリ毎に最長一致した単語列が系列分解されることから，各カテゴリの特徴を捉えた素性が選択される．ここで，カテゴリ c_k の訓練文書を $\boldsymbol{d}_{j_k} = y_{j_k1}y_{j_k2}\cdots y_{j_kR_{j_k}}$ とする．R_{j_k} は，文書 d_{j_k} の総単語数を表している．

［アルゴリズム］

Step1　$m=1$，$k=1$，$i=1$ とする．

Step2　カテゴリ k に属する全訓練文書 d_{j_k} に対し，単語列 $y_{j_k1}y_{j_k2}\cdots y_{j_kR_{j_k}}$ と単語列 $x_mx_{m+1}\cdots$ と比較し，最長一致する単語数を q_{j_k} とする．

Step3　$Q=\max_{j_k} q_{j_k}$ とする．

Step4　$Q>0$ のとき $\boldsymbol{s}_{i_k} = x_m\cdots x_{m+Q-1}$，$m=m+Q$，$i=i+1$ とする．$k\leq K-1$ かつ，$m+Q=M$ であれば，Step5 へ．$k=K$ かつ，$m+Q=M$ となるとき，アルゴリズムを終了．さもなければ，$m=m+1$ とし，Step2 へ．

Step5　$k=k+1$ とし，Step2 へ．

カテゴリ k に属する訓練文書を事前情報とした場合，各カテゴリで分解されるパターンは異なるため，$\boldsymbol{S}_k = \boldsymbol{s}_{1_k}\boldsymbol{s}_{2_k}\cdots\boldsymbol{s}_{I_k}$ をカテゴリ c_k において I_k 個に系列分解された単語列を素性として定義する．

3.2　ナイーブベイズ法による分類

提案手法において，$\boldsymbol{S}_k$ は訓練文書のカテゴリ別に ZM 法を用いた場合の素性が選択される．カテゴリ k の訓練文書を事前情報とした $\boldsymbol{S}_k$ 中の素性はカテゴリ k に属する訓練文書中に必ず存在するため，ゼロ頻度問題は発生しない．(4) 式により，$c_{\hat{k}}$ を入力文書 $\boldsymbol{x}$ のカテゴリと判別する．

$$\hat{k} = \underset{k}{\operatorname{argmax}} \left\{ \log P(c_k) + \sum_{i=1}^{I_k} \log \frac{P(c_k|\boldsymbol{s}_{i_k})}{P(c_k)} \right\} \quad (4)$$

4　実験方法

提案手法の有効性を検討するため，新聞記事のデータを用いて文書分類実験を行い，分類精度の評価を行なった．また，実験では新聞記事を形態素解析により分かち書きした語を単位とする．

4.1　実験条件

実験には，毎日新聞 2000 年の 4 カテゴリ（社会・経済・スポーツ・芸能）の記事を使用した．実験に用いた記事は唯一のカテゴリに属し，記事が他のカテゴリに重複することはない．各カテゴリで 550 記事ずつの合計 2200 記事をランダムに選び，訓練文書として各カテゴリ 500 個，テスト文書として各カテゴリ 50 個にランダムに分ける．提案手法との比較のため，単語を素性とした NB 法と，相澤による手法を従来手法とした，３つの手法を適用させる．

4.2　実験結果

単語を素性とした場合の NB 法と，従来手法と提案手法の実験結果を図 3 に示す．また，各カテゴリにおける分類精度と分解された系列数との関係を表 1 に示す．

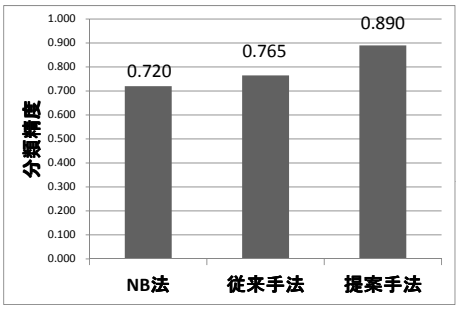

図 3．各手法による分類精度

表 1．提案手法の各カテゴリの分類精度と比率 1 と比率 2

	社会	経済	スポーツ	芸能
提案手法	0.72	0.94	0.92	0.98
比率*	32/50	46/50	40/50	41/50
比率**	3/50	2/50	1/50	1/50

比率*　そのカテゴリに属する入力文書のうち，提案手法によって，最も系列数が少なく分解された文書の割合

比率**　そのカテゴリに属する入力文書のうち，提案手法によって，最も系列数が少なく分解されたものの，別のカテゴリに分類された文書の割合

4.3　考察

表 1 より，提案手法において，正解カテゴリで最も系列数が少なく分解された文書の割合が多いことがわかる．すなわち，正しい所属カテゴリでマッチングしたときに，文書は平均的に長い N が得られるといえる．また，比率 2 より，誤分類した文書の比率も少ないことから，各カテゴリの訓練文書を用いて，入力文書を可変長 N グラムに系列分解したことが分類精度に影響したと考えられる．

5　まとめと今後の課題

本研究ではカテゴリ別に可変長 N グラムへの系列分解を行う素性選択を提案した．その結果，NB 法と従来手法より分類精度を高めることができ提案手法の有効性が示された．

今回，分類手法として代表的な確率的分類手法である NB 法を用いたが，NB 法以外の分類手法を，提案手法の分類手法に適用することが今後の課題である．

参考文献

[1] 鈴木誠,“ カテゴリ間の単語頻度の差分を用いたテキストの自動分類，”日本経営工学会論文誌,Vol.59 No.4,pp. 355–363.2008.

[2]William B. Cavnar and John M. Trenkle,“ N-gram-based text categorization, ”*In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*,pp. 1–38, 1994.

[3]Jacob Ziv and Neri Merhav,“ A Measure of Relative Entropy Between Individual Sequences wit Application to Universal Classification, ”*IEEE Trans. Inform*,Theory,VOL.39,pp. 161–175, 1993.

[4] 相澤彰子,“ 多クラス文書分類問題における Ziv-Merhav Crossparsing の適用と評価，”情報処理学会論文誌,Vol.52 No.10,pp. 2953–2964, 2011.