E-019

# 品詞分布を用いた日本語文書のジャンル分類
# Classification Using POS Distribution in Japanese Document

白井 匡人† 島田 諭‡ 三浦 孝夫†
Masato Shirai　Satoshi Shimada　Takao Miura

## 1. はじめに

近年，インターネットの発達から大量の文書を入手することが可能になり，静的な文書集合に対する分類だけでなく，次々に新たな文書が到着するという文書ストリームに対する分類手法の重要性が高まっている．

ストリームデータは新しいデータが逐次到着することからデータ量が膨大となる．このことから文書ストリームの分類には限られた特徴量で高速に分類する手法が必要となる．ジャンル分類では通常語の頻度が用いられ，ストリームデータの分類に対しても語の頻度を特徴量とした分類法が提案されている[2][3]．語の頻度を用いた場合，系列長の長い文書では語彙数が数万から数十万となり，数万次元の単語分布の比較が必要となる．また，データ数が増加していくことから必要なメモリ量が膨大となる．さらに，ストリームデータでは文書集合が動的に変化するため，ジャンルの特徴となる語が変わり，特徴語を選択することが困難になる．

しかし，特徴量として品詞分布を用いることで，ジャンル分類に必要な次元数が語彙数から品詞数へと大幅に削減でき語の扱いが容易になる．これによりストリームデータのように各ジャンルの特徴が逐次変化する場合でもパラメータの更新が容易に行える．単語分布は変化しても用いる品詞は同じであることから，品詞分布の変化としてジャンルの特徴の変化を捉えることができる．

本研究では，品詞分布を用いて様々なジャンルの文書が次々に到着すると仮定した文書ストリームに対する分類法を提案する．日本語文書の品詞分布には法則性があることが知られており[1]，各ジャンルの品詞の割合の理論値は品詞分布の決定要素である名詞の割合によって求めることが可能である．この品詞分布の法則性と各ジャンルの名詞の割合の事前分布にガウス分布を用いることでジャンル分類が行えることが示されている[4]．各ジャンルのガウス分布のパラメータと多項式係数ベクトルを更新することにより文書ストリームの分類を行う．

第 2 章では提案手法について述べ，第 3 章では実験により有効性を示し，第 4 章で結論とする．

## 2. 提案手法

### 2.1 分類方法

ジャンル分類では各ジャンルの名詞の割合の事前分布であるガウス分布と，品詞分布の線形近似の係数である多項式係数ベクトルを用いて分類を行う．文書の名詞の割合は事前分布であるガウス分布によって決まり，ガウス事前分布のパラメータがジャンルごとの名詞分布の特徴を表す．動詞，形容詞類，接続詞類の割合は名詞の割合と多項式係数ベクトルによって求まる．各品詞の多項式係数ベクトル$(w_{i0},w_{i1})$は以下の式より求まる．

$$w_{i0}^{(j)}=\frac{n\sum_{k=1}^{n}x_k y_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{n\sum_{k=1}^{n}x_k^2-(\sum_{k=1}^{n}x_k)^2},\ w_{i1}^{(j)}=\frac{\sum_{k=1}^{n}x_k^2\sum_{k=1}^{n}y_k-\sum_{k=1}^{n}x_k y_k\sum_{k=1}^{n}x_k}{n\sum_{k=1}^{n}x_k^2-(\sum_{k=1}^{n}x_k)^2}$$

ジャンルごとに学習したパラメータより，動詞，形容詞類，接続詞類の品詞の割合の観測値と理論値の誤差を求め，ガウス分布を仮定した名詞の出現確率の逆数をかけることによりジャンル分類を行う．各ジャンルの名詞の出現確率はガウス事前分布のパラメータである$\mu_j$と$\sigma_j$により以下の式で求まる．

$$N_j=\frac{1}{(2\pi\sigma_j^2)}\exp(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2)$$

各ジャンルの動詞，形容詞類，接続詞類の理論値はテスト文書$\{d_1,d_2,\cdots,d_n\}$の名詞の割合$\{x_1,x2,\cdots,x_n\}$と多項式係数ベクトルによって求まる．ここで j はジャンルであり，i は品詞である．テスト文書の品詞の割合を動詞$\{v_1,v_2,\cdots,v_n\}$，形容詞類$\{a_1,a_2,\cdots,a_n\}$接続詞類$\{c_1,c_2,\cdots,c_n\}$としたとき，理論値との誤差は以下の式より求まる．

$$V_j=(v-(w_{v1}^j x+w_{v0}^j))^2+(a-(w_{a1}^j x+w_{a0}^j))^2+(c-(w_{c1}^j x+w_{c0}^j))^2$$

品詞分布の理論値と観測値の誤差と名詞の出現確率の逆数によって求まる値が最も低いジャンルを分類結果とする．以下の式に基づいて分類を行う．

$$Ans=\arg\min_j V(j)\times N(j)$$

### 2.2 文書ストリームの分類

文書ストリームの分類を行うため，ジャンルの特徴の変化に応じてガウス分布のパラメータと多項式ベクトルを更新する．パラメータの更新は各ジャンルに用意したサンプル文書 $T\{t_1,t_2,\cdots,t_n\}$を用いて，F 検定による回帰直線の当てはまりを求め，更新後の F 値が高くなる場合にパラメータの更新を行う．サンプル文書にはエラー保証を行い，各品詞の割合は理論値から誤差 $\varepsilon$ 以内とする．理論値を K としたき，サンプル文書の割合は

$$K-\varepsilon\le k_t\le K+\varepsilon$$

となる．これによりジャンルの特徴と異なる文書がサンプル文書となることを防ぐ．

提案手法ではテスト文書の品詞分布が得られれば分類が行えるため，到着した文書を一定の長さの区間で区切り，先頭の区間のみを分類に用いる．品詞分布を求める際，低頻度の単語がノイズとなり，品詞の割合の誤差が大きくなることから頻度 1 の単語を除いて品詞の割合を求める．こ

† 法政大学大学院 工学研究科，HOSEI University
‡ 法政大学マイクロ・ナノテクノロジー研究センター，HOSEI University

の品詞の割合より，分類手法を用いて所属するジャンルを求める．次にサンプル文書を用いて F 検定を行い，更新前のパラメータと更新後のパラメータの比較を行う．F 検定では回帰による要因効果と誤差による変動要因により回帰直線の当てはまりの良さを求める．ここで回帰による要因効果についての平方和 $S_F$，平均平方 $V_F$ と誤差による変動要因についての平方和 $S_C$，平方平均 $V_C$ 以下の式で定義する．

$$S_F^{(P)} = \sum_{i=1}^{n} (\bar{Y} - Y_i)^2 \quad , \quad V_F^{(P)} = \frac{S_F}{m-1}$$

$$S_C^{(P)} = \sum_{i=1}^{n} (y_i - Y_i)^2, \quad V_C^{(P)} = \frac{S_C}{n-m}$$

この時,

$$F_P = \frac{V_F}{V_C}$$

となる．ここで自由度は分母に対して(全標本数－群数)であり，分子に対して(群数－1)である．更新後の F 値が高く，サンプル文書のうち一定数以上が新しい理論値から ε 以内に収まっていればパラメータの更新を行う．また，更新されたとき理論値から ε 以内に収まっていないサンプル文書は除去する．分類されたテスト文書の品詞の割合が誤差以内に収まっていればサンプル文書に加え，最も古いサンプル文書は取り除く．

## 3. 実験

本章では，文書ストリームの分類を行い提案手法の有効性を示す．比較手法としてパラメータを更新しない手法を用いる．

### 3.1 実験準備

実験には 4 つのコーパスを用いる．各コーパスは，5 人の著者の小説を 20 作品ずつ計 100 作品，朝日新聞の 2007 年度を 1 月 1 日から 100 日分，日本語話し言葉コーパス(Disc3)を収録順に先頭から 100 文書，NTCIR-3 より 98 年度公開特許公報全文データを 100 文書である．特許データは発明の詳細な説明のみを抽出して用いる．また，これらのうち各 50 文書を学習データとして用いる．実験データには茶筅による形態素解析を行う．

提案手法のパラメータの値は，区間の長さ 2000 単語，最大サンプル文書数 10，更新時の最低サンプル文書数 5，理論値からの誤差 ε =0.1 とする．

### 3.2 評価方法

実験の評価には f 尺度を用いる．f 尺度は再現率と適合率の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率と，正と予測したデータのうち，実際に正であるものの割合である適合率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \qquad P_i = \frac{a_i}{a_i + b_i}$$

$a_i$ は推定結果が正である数，$c_i$ は正であるが負と推定された数，$b_i$ は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である f 尺度を次のように定義する．

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \; f = Average_i \, f_i$$

### 3.3 実験結果

実験結果を表 1 に示す．表 1 より提案手法の f 値は 0.93 であり，比較手法の f 値は 0.898 である．

| | 提案手法 | 比較手法 |
|---|---|---|
| 再現率 | 0.93 | 0.895 |
| 適合率 | 0.936 | 0.901 |
| f 値 | 0.933 | 0.898 |

表 1　分類結果

## 4. 考察

表 1 より，提案手法と比較手法の f 値を比較すると提案手法の f 値が高くなっている．また，表 2，表 3 より各ジャンルの多項式係数ベクトルが更新されていることが確認できる．これらの結果から，パラメータを更新することで精度が向上していると考えられる．

| | 小説 | 新聞 | 特許 | 話し言葉 |
|---|---|---|---|---|
| $w_{vo}$ | -0.570 | -0.785 | -0.784 | 0.050 |
| $w_{v1}$ | 0.619 | 0.797 | 0.782 | 0.264 |
| $W_{a0}$ | -0.328 | -0.200 | -0.124 | -0.345 |
| $W_{a1}$ | 0.298 | 0.187 | 0.126 | 0.277 |
| $W_{c0}$ | -0.102 | -0.014 | -0.092 | -0.705 |
| $W_{c1}$ | 0.082 | 0.016 | 0.092 | 0.459 |

表 2　多項式係数ベクトル(更新前)

| | 小説 | 新聞 | 特許 | 話し言葉 |
|---|---|---|---|---|
| $w_{vo}$ | -0.563 | -0.782 | -0.767 | 0.085 |
| $W_{v1}$ | 0.616 | 0.795 | 0.765 | 0.258 |
| $W_{a0}$ | -0.347 | -0.192 | -0.145 | -0.316 |
| $W_{a1}$ | 0.308 | 0.180 | 0.143 | 0.269 |
| $W_{c0}$ | -0.090 | -0.026 | -0.088 | -0.769 |
| $W_{c1}$ | 0.076 | 0.025 | 0.092 | 0.474 |

表 3　多項式係数ベクトル(分類終了時)

## 5. おわりに

本稿では，品詞分布による文書ストリームの分類法を提案した．ジャンルごとのガウス分布のパラメータと多項式係数ベクトルを更新することにより高い精度で分類が行えることを示した．

### 参考文献

[1] 樺島 忠夫, “類別した品詞の比率に見られる規則性, 国語国文, Vol.250, pp.385-487 (1955).
[2] Chai, K.M., Ng, H.T. and Chieu, H.L. “Bayesian Online Classifiers for Text Classification and Filtering”, Proc.ACM SIGIR,(2002)
[3] Gurmeet Singh Manku, G.S. and Motwani, R. “Approximate Frequency Counts over Data Stream” , Proc. 28th international conference on Very Large Data Bases (VLDB), (2002)
[4] Shirai, M. Miura, T.“Document Classification Using POS Distribution”,ADBIS,(2012).