文書分類モデルの統計的性質に関する一考察

* 後藤正幸 (武蔵工業大学 環境情報学部)
平澤茂一 (早稲田大学 理工学部)
俵 信彦 (武蔵工業大学 工学部)

1.はじめに

近年, インターネットの普及により膨大なテキストデータからの知識発見を扱うテキストマイニングの技法が注目されている[1],[2]. 本研究では, テキストマイニングが取り扱う問題の中でも, 特に文書分類の問題を取り上げ, 伝統的な統計的仮説検定と漸近論の理論的枠組みから得られる性質について分析を行う. とくに, 文書のクラス数が2である場合には, 文書分類の問題は仮説検定とほぼ同様の枠組みで捉えることができる. 本稿では, 文書クラスを特定するために, 形態素解析後の単語の出現分布としてある確率モデルのクラスを仮定し, そこから得られる性質を検討することで, 文書分類の本質的な挙動について明らかにする.

2.文書モデル

本節では, 形態素解析により, 各文書 について単語への切り出しが行われた後, 情報検索やテキスト分類の問題を取り扱い易い問題に落とし込んだモデルであるベクトル空間モデルについて述べる.

2.1 ベクトル空間と文書－単語行列

分析対象である文書集合を $\Delta=\{d_1,d_2,\cdots,d_D\}$とする. Δ内の全ての文書について, 文書内に含まれる単語を抽出する. この単語抽出には, 通常, 文書の分類や検索のために有効となる単語(有効語)を選定して抽出する. すなわち, 助詞や句読点など, 文書の内容にあまり関係なく出現する語は分類や検索には意味をなさないため除外する. 通常は, 有効語として名詞や動詞の語幹の中から全文書中での頻度を考慮して選定される.

全文書から抽出された有効語の集合を $\Sigma=\{w_1,w_2,\cdots,w_W\}$とすれば, 各文書の特徴ベクトルを各特長語の出現頻度に応じて, W次元ベクトルで表現することができる.

すなわち, 文書集合Δから得られる全有効語によってベクトル空間が構成され, 文書 d_i を次式で表現することができる.

$$d_i=(v_{i1},v_{i2},\cdots,v_{iW})^T \tag{1}$$

ただし, T は転置を表す. ここで, この文書ベクトルを集めた行列

$$A=(d_1,d_2,\cdots,d_D)^T \tag{2}$$

を文書－単語行列(document word matrix)と呼ぶ.

2.2 TF-IDF Measure と文書間の類似度判定

いま, f_{w_j} を全ての文書中の単語 w_j の頻度, f_{d_i} を文書 d_i 内の全単語の総頻度, F を全文書中の全単語の総頻度とする. すなわち,

$$F=\sum_{w_j}\sum_{d_i} f_{ij}=\sum_{d_i} f_{d_i}=\sum_{w_j} f_{w_j} \tag{3}$$

の関係があるとする. (1)式の要素である v_{ij} として, 相対頻度を考え, $v_{ij}=f_{ij}/F$ とする方法や, 文書の長さによる影響を解消するために $v_{ij}=f_{ij}/f_{d_j}$ とする方法もある.

通常, 全ての文書にまんべんなく表れる単語は, 文書の特徴を規定するためにはあまり意味がない. むしろ, 少数の文書において集中的に表れる単語は分類や検索に有効である. そこで, 各単語の出現頻度だけでなく, 全文章中でその単語が現れる割合を考慮した特長量の算出が必要であり, そのための方法が TF-IDF measure である. TF は Term Frequency の略であり, 文字通り単語の出現頻度を表す. 一方, IDF は Inverse Document Frequency の略であり, 全文書中の単語の出現割合の減少関数を表す. ここでは, TF を文書 d_i における単語 w_i の相対頻度とし,

$$tf(d_i,w_j)=f_{ij}/F \tag{4}$$

とおく. IDF は単語 w_i を含む文書の数を $df(w_j)$ とすると,

$$idf(w_j)=\log D/df(w_j) \tag{5}$$

のような関数で定義される. このとき, 文書 d_i における単語 w_j の特徴量 v_{ij} は,

$$v_{ij}=tf(d_i,w_j)\cdot idf(w_j) \tag{6}$$

で与えられる.

各文書の特徴量がベクトル表現されれば, 文書 d_i と文書 d_k の類似度は, これらの距離を使って測ることができる. この距離には, ユークリッド距離や内積を用いることも可能であるが, 文書ベクトル d_i と文書ベクトル d_k の余弦をとって類似度とする方法が一般的である.

$$sim(d_i,d_k)=\frac{d_i{}^T d_k}{|d_i||d_k|} \tag{7}$$

以上のように, 文書分類問題では独特の特徴量算出と距離計算が行われ, 実際にこれらの

経験的に導かれた式による分類性能が良い．その理由を解明することは課題であるが，本稿では仮説検定の枠組みから文書分類問題の特性について述べる．

3.統計的仮説検定モデルによる考察

本研究では，文書分類の基本的性質を調べるため，最もシンプルと考えられる確率モデルを仮定し，その挙動について述べる．

いま，各文書は2つのクラスC_1とC_2から生起するものとする．一般的な仮定として，2 つのクラスC_1とC_2において文書と単語の出現確率が異なると考えることができる．本稿ではさらに，各文書と単語の出現確率は独立であり，確率$p^t{}_j$をクラスC_tから出現する文書の第 j 単語w_jの出現確率とする．

この時，一般性を失うことなく，$j=1,2,\cdots,p$については$p^1{}_j > p^2{}_j$，$j=p+1,p+2,\cdots,p+q$については$p^1{}_j < p^2{}_j$，$j=p+q+1,p+q+2,\cdots,W$については$p^1{}_j = p^2{}_j$であると仮定する．すなわち，初めからp個の単語はクラスC_1の文書で生起し易く、次のq個の単語はクラスC_2の文書で生起し易く，さらにそれ以降の$W-p-q$個の単語は両クラスで生起確率が同じであり，識別するための情報を与えない単語である．ここで，p^tからみたp^uの KL 情報量$L(p^t;p^u)$を

$$L(p^t;p^u)=\sum_{j=1}^{W} p^t{}_j \log\frac{p^t{}_j}{p^u{}_j} \tag{8}$$

と定義する。いま，ネイマン-ピアソンの定理より，2 つのクラスへの判定領域を

$$\Pi_K=\left\{\hat{q}:\sum_{j=1}^{W}\hat{p}_j\log\frac{p^1{}_j}{p^2{}_j}\geq K\right\} \tag{9}$$

$$\Pi^C{}_K=\left\{\hat{q}:\sum_{j=1}^{W}\hat{p}_j\log\frac{p^1{}_j}{p^2{}_j}< K\right\} \tag{10}$$

とかく．判別しようとしている文書の単語頻度分布が$\hat{q}\in\Pi_K$であればクラスC_1に分類し、$\hat{q}\in\Pi^C{}_K$であればクラスC_2に分類する．また，αをクラスC_1から出現した文書をクラスC_2に分類してしまう誤り（タイプIの誤り）の確率，βをクラスC_2から出現した文書をクラスC_1に分類してしまう誤り（タイプ II の誤り）の確率とする．

さらに，両クラスを識別するために情報を与えない単語を全て削除できた場合の両クラスの確率分布モデルを

$$S^t=\sum_{j=1}^{p+q}p^t{}_j \tag{11}$$

として，$\tilde{p}^t{}_j = p^t{}_j / S^t$とおき，

$$L^*(p^t;p^u)=\sum_{j=1}^{p+q}\tilde{p}^t{}_j\log\frac{\tilde{p}^t{}_j}{\tilde{p}^u{}_j} \tag{12}$$

とする．このとき，仮説検定の結果として知られる Stein の補題と Sanov の定理[3]のアナロジーとして以下の定理が得られる．

【定理1】 $\beta\in(0,1)$を固定する．α^*を，タイプ II の誤りがβを超えない条件で，あらゆる決定ルールの中で最小化したタイプ I の誤り率とする．このとき文書長を十分長くし$f_{d_i}\to\infty$とすると，

$$\alpha^*\to\exp\{-SL^*(q^1;q^2)\}$$

となる．ただし，$S=S^1=S^2$である． □

【定理2】 文書がクラスC_1から出現するとき，

$$\Pr\{\hat{q}\in\Pi_K\}\leq\exp\{-SL^*(q^*;p^1)\}$$

で与えられる．ただし，q^*は次式で与えられる．

$$L^*(q^*;p^1)=\min_{\hat{q}\in\Pi_K}L^*(\hat{q};p^1) \qquad \square$$

これらの定理が意味するところを考察してみよう．定義より$0<S\leq 1$であり，これは文書分類に意味をなす単語の出現確率を表す．文書分類問題では通常，多くの単語が自動的に切り出され，文書-単語ベクトルが自動生成されるため，分類に不要な単語も多く含まれることが考えられる．その不要な単語の確率が$1-S$であるとき，文書の判別誤りの指数部にSが現れるので，その分だけ分類精度が劣化する．TF-IDF measure は，有効語に大きな重みを持たせる事で，このような不要語を排除しようとする方法であり，このような方法が有効に働くのは上で示したような誤り率の劣化によって説明できる．

4．まとめ

本稿では，基本的な文書分類モデルの解析結果について述べた．ここでは，基本的なモデルについて考察したが，文書の確率自体が 1 点ではなく，分布で与えられる場合についても漸近正規性の議論を用いて分析することが可能である．

参考文献

[1] 金明哲，村上征勝，永田昌明，大津起夫，山西健司： 言語と心理の統計，岩波書店，(2003)

[2] 後藤正幸：“自然言語情報の分析手法と経営学的諸問題への応用”，武蔵工業大学環境情報学部紀要，to appear, (2006)

[3] Richarde E. Blahut: Information Theory, Addison-Wesley Publishing Co., (1987)