単語の意味を考慮した共起ベクトルによるテキスト分類

尾脇　拓朗†　　福元　伸也†

† 鹿児島大学大学院理工学研究科　〒 890–0065 鹿児島市郡元 1-21-40
E-mail: †{sc108012,fukumoto}@ibe.kagoshima-u.ac.jp

あらまし　本研究では，シソーラスを利用し単語の分類的意味属性に基づいて，テキストの共起行列を生成し識別する手法を提案する．出現単語のみによる共起行列では，意味的に近い単語の共起頻度が別々にカウントされたり，また，共起ベクトルの次元数が増大するなどの問題があった．そこで，分類語彙表を利用して，意味的に近い単語を分類語に当てはめた共起行列を作成し，ランダムフォレストを用いて識別する．単語の属性を反映させることにより，次元数の増大を抑え，意味的に近い単語の共起ベクトルが，より的確なベクトルとして表現されることが期待できる．実験では，ニュース記事に含まれる単語を用いて，提案手法の有効性について検証する．

キーワード　テキスト分類，共起行列，ランダムフォレスト

1. はじめに

近年，インターネットの普及により，膨大な量の文書データを扱う必要性が増大している．情報検索において，さまざまな分野におけるテキストデータを内容的に同じような質を持ったグループに分けるために，テキスト分類に関する研究が長年行われてきた．そこで，大量の文書データを自動的に効率よく分類できる識別法が必要とされている [1], [2]．テキスト分類では，テキストに含まれる文章を構成する語の重みを用いることにより，文書をベクトルとして表現し，文書ベクトル間の類似度を，何らかの類似尺度を利用して定義したのち文書分類を行う．

本研究では，文書ベクトルを構成するため，文書中に現れる共起単語と単語の意味属性を用いた共起行列を用い，つぎに，文書識別アルゴリズムにアンサンブル学習法の 1 つであるランダムフォレストを利用して，文書分類を行う手法を提案する．通常，文書内には，似たような意味を持つ複数の単語が存在する．ある単語に隣接して単語が現れることを共起と言う．単語同士の共起頻度を利用した共起行列を用いて文書分類を行うと，本来似ている意味の単語が，距離の離れたベクトルとして表現されるため，分類精度の低下が生じる可能性が高くなる [3].

本研究では，語を意味により分類したシソーラスである分類語彙表を用いて，単語の意味を考慮した共起行列を作成し，ランダムフォレストを用いて分類する手法について述べる．実験では，ウェブ上のニュース記事を用いて分類を行い，その有効性について検証する．

本稿は，まず，2 章で関連研究について述べる．次に 3 章で文書識別アルゴリズムとして用いるランダムフォレストについて述べ，4 章で提案手法の共起行列の作成法と判別予測の方法について説明する．そして，5 章で実験とその結果を述べ，最後に 6 章でまとめと今後の課題について述べる．

2. 関連研究

文書分類の研究において，先行研究では，語の同一文書内の共起関係に基づいた手法が多くなされている．単語の係り受け関係を用いて文書分類する研究や Web ページの語の共起関係を用いたものなどがある．また，多くの研究が分類精度の向上を目指しており，文書中に含まれる語の重要度を決める方法に注目している．Wang らは，語の重要度の決定に PageRank アルゴリズムを用いることが分類に有効であることを示した [4]．単語の持つ潜在的意味解析による方法 [5] では，単語の共起ベクトルの作成のために，単語同士の共起頻度を利用して，共起行列を作成していた．そのため，共起行列の行と列が単語によって表現されるため，ある単語 A と B との共起頻度が別々にカウントされてしまう．従って，単語 A と B との共起ベクトルが適切でなくなり，A と B は，意味的には似た単語であるのに，共起ベクトル間の距離は離れてしまうという現象が起きる．別所らは，単語同士の共起頻度ではなく，単語とコーパスにおける単語に付随する意味属性との共起頻度を取る手法を提案した [6]．意味属性を利用する手法では，共起ベクトルの質が向上し，高い検索精度が得られることを示した．本研究では，意味属性を利用した共起行列を入力として与えて，アンサンブル学習の 1 つであるランダムフォレストを用いて文書分類を行う方法を提案する．

3. ランダムフォレスト

ランダムフォレストは，複数の木（tree）によって構成される機械学習アルゴリズムである [7]．ここでの木は，決定木のことで，それぞれの決定木の性能はあまり高くなく，それらを複数組み合わせることにより，高い予測精度を持つ学習器となる．ランダムフォレストでは，決定木として二分決定木が主に用いられ，個々の決定木がアンサンブル学習における弱学習器となる．ランダムフォレストのアル

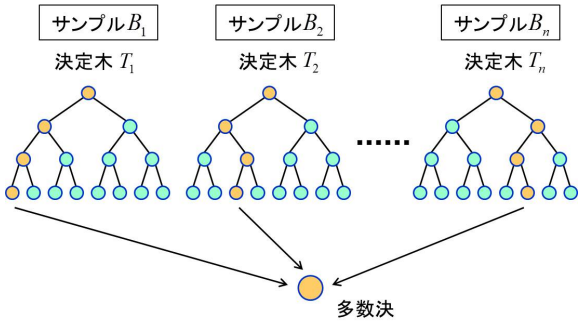

図 1　ランダムフォレスト

ゴリズムを以下に示す [8]. (図 1)

（1） 与えられたデータセットから n 個のブートストラップ・サンプル $B_1, B_2, \cdots, B_n$ を作成する．ただし，構築したモデルを評価するために 1/3 のデータを除いてサンプリングする．除いたデータを OOB（out-of-bag）データと呼ぶ．

（2） $B_k(k = 1, 2, \cdots, n)$ における M 個の変数の中から m 個の変数をランダムサンプリングする．M は，データセットの中の変数の数を表し，m は，$m = \sqrt{M}$ が多く用いられる．

（3） ブートストラップ・サンプル B_k の m 個の変数を用いて，未剪定の最大の決定木 T_k を生成する．

（4） n 個のブートストラップ・サンプル B_k の決定木 T_k について，OOB データを用いてテストを行い，推測誤差を求める．

（5） その結果を統合し，新たに分類器を構築する．分類の問題では多数決をとる．

本研究では，上記アルゴリズムを用いて文書分類を行う．

4. 提 案 手 法

4.1 共起行列の作成

共起行列の作成において，単語同士の共起頻度を利用した共起行列の生成では，

似た意味の単語であるのに，共起ベクトル間の距離は離れてしまう問題があった．図 2 は，単語同士の共起頻度を表した共起行列を示している．例えば,「本棚」と「書棚」に対する共起単語が,「小説」と「雑誌」で，共起頻度が図のようになっているとすると，それぞれの単語における共起ベクトルは離れたものになってしまう．そこで,「小説」,「雑誌」といった単語を用いるのではなく，シソーラスの分類項目に関連付けて，その意味属性である「図書」で共起頻度をカウントする．すると，各単語が意味の似ている語にまとまるため,「本棚」と「書棚」の共起ベクトルの距離は近くなる．

分類項目
意味属性 → 図書

	---	小説	雑誌	---
---	---	---	---	---
本棚	---	3	80	---
書棚	---	73	5	---
---	---	---	---	---

図 2　共 起 行 列

提案手法では，意味の似ている語をまとめると共起ベクトルの距離は近くなるという点を踏まえ，単語同士の共起頻度を用いるのではなく，単語に付随する意味属性を利用する．単語の意味属性には，単語を意味によって分類整理したシソーラスである分類語彙表を利用した [9]．分類語彙表を構成する項目は，図 3 のようになっており，共起行列に用いる意味属性には，その中の「分類項目」を用いた．共起行列の 1 列目（行の先頭）には，形態素解析の結果得られた単語のうち，名詞のみを取り出し入力し，数字の部分は，1 文中に共起する頻度をカウントした数が入った行列となっている．また，1 行目には，意味属性として分類項目の語が入る．

4.2 判 別 予 測

ランダムフォレストは，決定木の変数にすべての変数を使うのではなく，ランダムサンプリングしたサブセットを用いるため，高次の問題に対し有効ではないかと考え，学習器にランダムフォレストを用いることとした．ランダムフォレストでは，得られた共起行列をデータセットとして，ブートストラップ・サンプルを作成する．次に，ブートストラップ・サンプルを用いて，未剪定の最大の決定木を生成する．そして，OOB データを用いてテストを行い，全ての結果から多数決を取り，予測結果とする．ランダムフォレストでは，乱数に依存しない結果を得るため，決定木の数を十分に多くしておく必要がある．本手法においても，十分多くの決定木を与えるものとする．

提案手法におけるテキストの解析・抽出，シソーラスによる変換・共起行列作成，および学習・結果の出力までの処理の流れを図 4 に示す．

5. 実　　験

5.1 記事の分類

実験では，ニュース記事の分類を行った．毎日新聞のサイト [10] では，ニュース記事が，カテゴリーごとに分かれている．その中の，政治，経済，社会，スポーツ，エンター

レコード ID 番号／見出し番号／レコード種別／
類／部門／中項目／分類項目／分類番号／
段落番号／小段落番号／語番号／見出し／
見出し本体／読み／逆読み

図 3　分類語彙表の項目

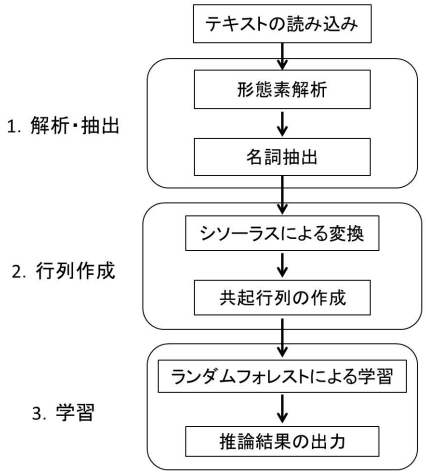

図 4　処理の流れ

テイメントの 5 つのカテゴリーから，それぞれ 20 個，計 100 個の記事を記事データとして用いた．つぎに，選んだ 100 個の記事データを形態素解析器 Igo [11] を用いて形態素解析し，名詞の単語のみを抽出した．Igo は，Java で実装された形態素解析器で，辞書のフォーマットや解析結果は，MeCab と互換があるものとなっている．形態素解析の結果，4,194 個の名詞が抽出された．

抽出された単語に対する意味属性を分類語彙表との対応から選び出し，共起行列を作成する．分類語彙表に無かった単語については取り除く．共起行列に残った単語は，3,083 個となった．共起行列の列数に相当する分類項目の数は，455 個となった．得られた共起行列をデータセットとし，ブートストラップ・サンプルを作成して，ランダムフォレストを行った．ランダムフォレストの決定木の数は，500 個とした．

5.2　学習器による比較

提案手法では，識別のための学習器として，ランダムフォレスト（RF）を用いている．学習器により，識別率に違いが出るのかを調べるため，次の 2 つの学習器を用いた．

1 つは，カーネル法による分類器であるサポートベクターマシン（SVM:Support Vector Machine）[12]，2 つ目は，アンサンブル学習の 1 つであるバギング（Bagging）[13] による方法で，それらの学習器を用いた分類結果を表 1 に示す．表中の値は，識別の結果，正しく分類された割合 (%) を示しており，また，表中のラベル名は，記事のカテゴリーを意味し，それぞれ，(POL : 政治)，(ECO : 経済)，(SOC : 社会)，(ENT : エンターテイメント)，(SPO : スポーツ) を示している．

図 5 は，3 つの手法における識別率を示しており，エラーバーは，95% の信頼区間である．提案手法のランダムフォレストによる手法は，他の手法より識別率が高く，有意差が見られた．

表 1　識 別 結 果

	POL	ECO	SOC	ENT	SPO	Ave
提案手法 (RF)	91.8	96.7	85.0	96.8	86.1	91.3
SVM	73.1	92.9	70.8	92.9	66.8	79.3
Bagging	84.7	89.0	73.3	94.7	74.1	83.2

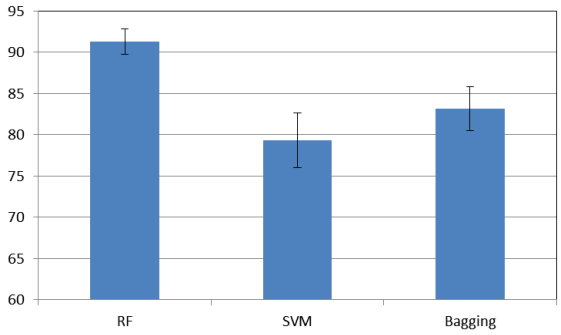

図 5　各手法の識別率

5.3　次元圧縮の違いによる比較

次に，単語に対する意味属性を分類語彙表を用いて，次元圧縮する提案手法と，特異値分解 $D = U\Sigma V^T$ を用いて [14]，次元圧縮を行った場合において，識別率を比較した．ただし，D は，文書行列で，ここでは，単語同士の共起頻度を表した共起行列を使用した．得られた共起行列の学習器には，ランダムフォレストを用いた．特異値分解による次元圧縮において，提案手法と同じ次元に圧縮するため，3,083 次元を 455 次元に圧縮した．提案手法における識別率は，91.3%(Ave) で，特異値分解を用いた場合は，89.4%(Ave) とわずかながら改善がみられた．

5.4　分類項目の重要度

提案手法において，分類語彙表を用いて抽出された分類項目が，単語の意味属性としてランダムフォレストで，どの程度重要であるのかを調べた．その結果を図 6 に示す．図の左側は，決定木における正解率を基準とした平均値を示しており，また，図の右側は，計算の過程で算出された Gini 係数の平均値を大きい順に並べて示している．Gini 係数は，分岐の前と後での誤差の改善度合いを表し，判別における変数の重要度として用いられ，図の上位に表された分類項目ほど，重要度の高い語ということになる．

表 2 は，図 6 において上位に挙がった分類項目の語と多く共起した記事中の単語を表している．演劇・映画では，俳優，脚本，舞台，ミュージアムなど，スポーツでは，選手，プロ，練習，優勝などといった，それぞれの分類項目の語に関連の強い単語が挙がっていることがわかる．表に出ている単語から，木の分岐に重要となる単語が挙がっていることを確かめることができる．

6.　お わ り に

本研究では，共起行列の作成において，単語に付随する意味属性を利用するため，シソーラスである分類語彙表を利用して，その共起行列をデータセットとして，ランダムフォレストを用いて文書識別を行う手法を提案した．実験において，ニュース記事のカテゴリ分別を行い，学習器

表 2　上位の分類項目と共起した単語

分類項目	記事中に現れた単語
演劇・映画	俳優，脚本，舞台，ミュージアム
スポーツ	選手，プロ，練習，優勝
創作・著述	映画，俳優，デビュー，物語
経済・収支	緩和，金利，上昇，銀行
興行	映画，舞台，動員，放送

による識別率に対する影響を調べた．その結果，ランダムフォレストを用いた提案手法において，SVM やバギングを用いた手法と比較して，高い識別率が得られた．また，分類語彙表を用いて次元圧縮を図った提案手法と特異値分解で次元圧縮を行った場合の共起行列を用いて識別率を調べたところ，提案手法で改善が見られた．また，重要度の高い分類項目と共起した単語の関係からは，木の分岐に重要となる単語が選ばれていることを確かめることができた．

今後の課題として，今回，共起行列を作成するために分類語彙表を利用したが，生成した共起ベクトルのコーパス依存も考えられるため，他のコーパスを利用した場合の結果の違いなどを調べていきたい．

7. 謝　　辞

本研究の一部は，JSPS 科研費 (24500120) の助成を受けて実施された．

文　　献

[1] R. M. Samer Hassan and C. Banea: “Random-walk term weighting for improved text classification”, Proc. of the First Workshop on Graph Based Methods for Natural Language Processing (2006).
[2] F. Sebastiani: “Machine learning in automated text categorization”, Proc. ACM Computing Surveys, **34**, 1, pp. 1–47 (2002).
[3] 那須川哲哉, 河野浩之, 有村博紀：“テキストマイニング基盤技術”, 人工知能誌, **16**, 2, pp. 201–211 (2001).
[4] W. Wang, D. B. Do and X. Lin: “Term graph model for text classification”, Springer-Verlag Berlin Heidelberg 2005, pp. 19–30 (2005).
[5] H. Schutze: “Automatic word sense discrimination”, Computational Linguistics, **24**, 1, pp. 97–124 (1998).
[6] 別所克人, 内山俊郎, 内山 匡, 片岡良治, 奥 雅博：“単語・意味属性間共起に基づくコーパス概念ベースの生成方式”, 人工知能誌, **49**, 12, pp. 3997–4006 (2008).
[7] L. Breiman: “Random forests”, Machine Learning, **45**, pp. 5–32 (2001).
[8] 金明哲：“統計的テキスト解析”, ESTRELA, 182 (2009).
[9] 国立国語研究所：“分類語彙表－増補改訂版”, 大日本図書刊 (2004).
[10] 毎日新聞：http://mainichi.jp/.
[11] Igo: http://igo.sourceforge.jp/.
[12] C. Cortes and V. Vapnik: “Support-vector networks”, Machine Learning, **20**, pp. 273–297 (1995).
[13] L. Breiman: “Bagging predictors”, Machine Learning, **24**, pp. 123–140 (1996).
[14] P. Dewilde and E. Deprettere: “Singular value decomposition: An introduction”, SVD and Signal Processing, pp. 3–41 (1988).

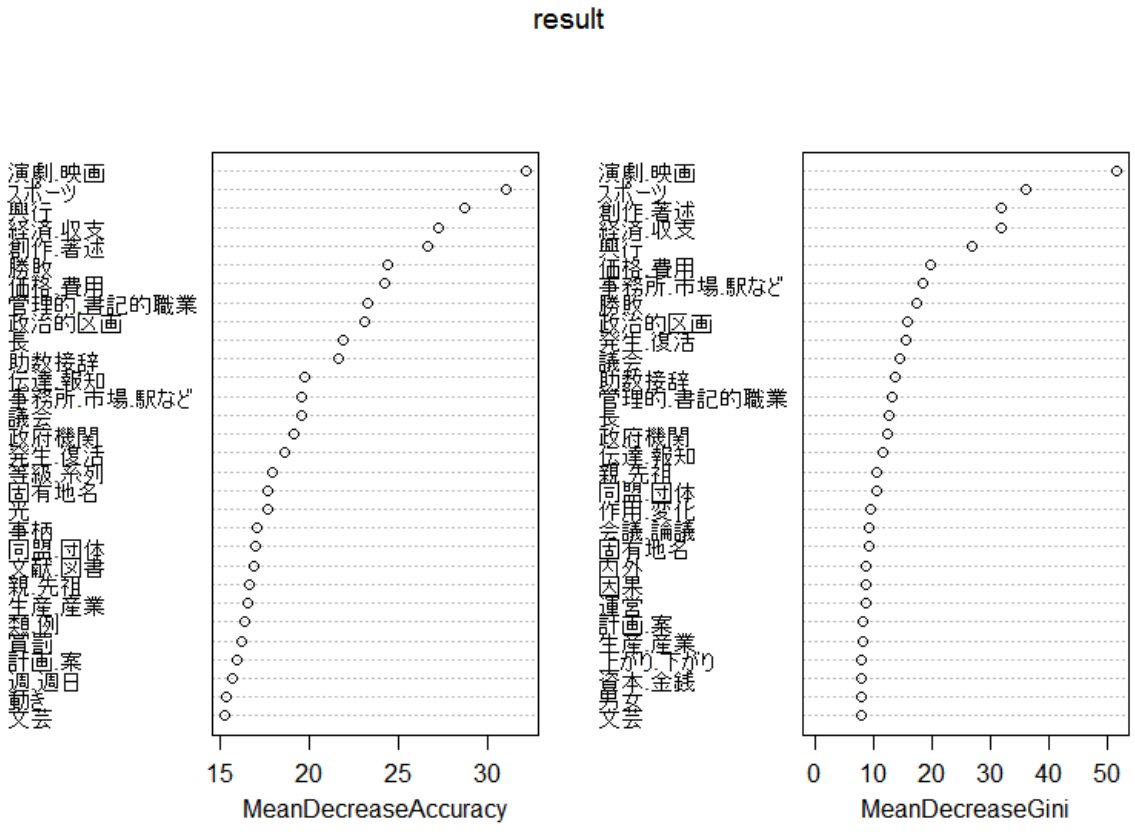

図 6　変数（分類項目）の重要度