政治的テキストの文法
—機械学習のための政治的テキストデータの構造—

重村壮平 [1]　宋財泫 [2]

[1] 神戸大学法学研究科博士課程後期課程
[2] 神戸大学法学研究科博士課程後期課程、日本学術振興会特別研究員 (DC)

概要

- 背景
  - ビッグデータの入手コストの低下 ⇒ ビッグデータに対する関心の上昇
  - ビッグデータ = 莫大なデータ量 ⇒ 高い分析コスト
  - 機械学習：分析コストを下げる可能性 ⇔ 機械学習を用いた分析方法は発展途上 [1]
- 目的
  - 政治的テキスト (選挙公約) の効率的な機械的コーディングのためのデータ構造を提示
- 結論
  - 選挙公約の情報損失を最小限に止めるデータの構造を提示
    ⇒ 選挙公約の高精度かつ機械的なコーディングが可能

先行研究

- 機械的コーディング
  - コーディングの過程

1. トレーニングセット（全データの一部）
2. 機械学習アルゴリズム
3. 分類器の生成
4. テストセット（全データの残部）→ 分類器 → 5. カテゴリを予測

  - ヒューマン・コーディングを代替 ⇒ 時間的・金銭的コスト↓
- 選挙公約の機械的コーディング [2]
  - 分類器の精度：「人間によるコーディング」と「機械によるコーディング」の一致率
  - 約 8 割の一致率 ⇒ 機械によるコーディングの有用性を提示
- 選挙公約の構造 [3]
  - 例：「子ども手当を実現する」(候補者 ID が 120，1 番目の公約)

$$M_{120,1} = 43\text{v}3\text{w} = \{\ \text{対象} = \text{生活者}(43),\ \text{内容} = \text{少子高齢化対応}(\text{v}3),\ \text{方向} = \text{改革}(\text{w})\}$$

  - 「対象」「内容」「方向」の三次元構造 ⇒ 公約を特定のカテゴリにコーディング可能
  - 誤ったコーディングによる情報損失が深刻
    （例）「方向」を誤ってコーディング ⇒3 分の 1 の情報を損失

ID	対象	内容	方向
1	生活者	少子高齢化対応	改革

⇒

ID	対象	内容	方向
1	生活者	少子高齢化対応	維持

選挙公約の文法（Grammar of Manifesto; GoM）

- 構造主義 [4]
  - 対象を複数の要素に分解し，分解された要素の組み合わせや，要素間の関係に注目
  - 意味を有する最小単位（=公約素）の集合として選挙公約を捉える（マクロ ⇒ ミクロ）
- GoM のデータ
  - 選挙公約を公約素で表現しデータ化

$$\text{例: } M_{120,1} = \{\ \text{対象} = \{\text{国民・民意} = 0, \text{市民} = 0, \text{生活者} = 1, \text{地域公約} = 0, \text{有権者} = 0, \dots\},\ \text{内容} = \{\text{保育} = 1, \text{福祉} = 1, \text{交通} = 0, \dots\},\ \text{方向} = \{\text{維持} = 0, \text{推進} = 1, \dots\}\}$$

  - 二層構造
    - 第一層：「対象」「内容」「方向」の三次元ベクトルで構成
    - 第二層：第一層内の各要素もベクトルで構成（例：「市民」「有権者」...etc）
  - 誤ったコーディングによる情報損失の最小化
    （例）一つのカテゴリ（「推進」）を誤ってコーディング ⇒ 情報の損失は僅少

ID	対象		内容		...	方向	
	生活者	市民	保育	福祉	...	推進	維持
1	0	1	1	1	...	1	0

⇒

ID	対象		内容		...	方向	
	生活者	市民	保育	福祉	...	推進	維持
1	0	1	1	1	...	0	0

  - 各要素をダミー変数で表現 ⇒ 単純なデータの扱いに秀でる機械学習と好相性

データ & 方法

- データ
  - 2009 年衆議院議員選挙に出馬した候補者の選挙公約
- 機械学習アルゴリズム
  - 決定木（Decision Tree; DT）
  - ランダム・フォレスト（Random Forest; RF）
  - ニューラル・ネットワーク（Neural Network; NN）
- コーディング結果の評価方法
  - 的中率
  - κ 統計量：0.6 以上で一致度が高いと評価

結果

- 既存のデータ構造より的中率↑
- アンサンブル分類器を用いることで κ 統計量↑

項目	DT 的中率	DT κ	RF 的中率	RF κ	NN 的中率	NN κ
対象						
40: その他 (対象なし)	0.767	0.504	0.844	0.659	0.841	0.663
41: 国民、民意	0.955	0.497	0.966	0.529	0.981	0.811
42: 市民	0.999	-	0.999	0.400	0.999	-
43: 生活者	0.998	-	0.988	-	0.998	-
44: 地域公約	0.869	0.420	0.903	0.492	0.908	0.634
46: 庶民	0.999	-	0.999	-	0.999	-
47: 消費者	0.999	-	0.999	-	0.999	-
48: 住民	0.996	0.314	0.998	-	0.998	-
50: 被災者	1.000	-	1.000	-	1.000	-
51: 高齢者	0.971	0.682	0.980	0.745	0.988	0.875
52: 女性	0.991	0.432	0.993	0.448	0.995	0.731
53: 子ども・青少年	0.932	0.713	0.953	0.779	0.965	0.857
54: 青少年 (有権者)	0.988	0.044	0.991	-	0.997	0.783
55: 社会人	1.000	-	1.000	-	1.000	-
56: 障害者	0.992	0.515	0.994	0.610	0.997	0.843
57: 低所得者	0.995	0.331	0.997	0.374	0.995	0.346
58: 外国人	0.999	-	0.999	-	0.999	-
59: 被爆者	1.000	-	1.000	1.000	1.000	-
61: 労働者	0.998	0.285	0.999	-	0.999	-
62: 勤労者	0.981	0.531	0.986	0.568	0.987	0.688
63: パート	0.993	0.835	0.995	0.861	0.998	0.940
64: 働く女性	0.996	0.249	0.996	-	0.996	-
65: 福祉従事者	0.995	0.468	0.996	0.598	0.997	0.755
66: 中小企業	0.985	0.738	0.990	0.809	0.992	0.865
67: 農漁業	0.966	0.583	0.977	0.695	0.983	0.804
68: 大企業	0.992	0.712	0.996	0.855	0.998	0.912
70: 商店街	0.997	0.284	0.997	-	0.997	-
71: 戦争被害者	1.000	-	1.000	-	1.000	-
72: 社会的弱者	0.993	0.497	0.997	0.665	0.997	0.735
73: ベンチャー企業	0.999	0.500	0.999	0.500	0.999	-
99: その他 (対象あり)	0.992	0.702	0.996	0.818	0.995	0.802
内容						
a: 内閣	0.913	0.565	0.941	0.652	0.938	0.723
b: 自治	0.928	0.477	0.948	0.514	0.981	0.811
c: 安保・外交	0.969	0.737	0.973	0.739	0.970	0.771
f: 大蔵	0.948	0.629	0.964	0.695	0.965	0.770
g: 文科	0.950	0.647	0.961	0.674	0.973	0.823
h: 厚生	0.897	0.701	0.936	0.805	0.931	0.813
i: 労働	0.939	0.612	0.960	0.689	0.972	0.834
j: 農水	0.962	0.684	0.972	0.740	0.974	0.795
l: 通産	0.940	0.473	0.951	0.440	0.958	0.659
m: 運輸	0.977	0.415	0.983	0.411	0.989	0.711
n: 郵政	0.995	0.754	0.995	0.774	0.997	0.870
o: 建設	0.960	0.614	0.973	0.689	0.971	0.720
q: 環境	0.975	0.529	0.979	0.474	0.980	0.646
r: 政治	0.918	0.676	0.947	0.775	0.943	0.787
v: その他	0.880	0.433	0.912	0.458	0.905	0.601
k: 構造改革	0.996	0.141	0.997	0.181	0.997	0.153
方向						
t: 維持	0.773	0.546	0.865	0.730	0.848	0.697
w: 転換	0.782	0.556	0.864	0.723	0.843	0.678
z: その他	0.996	-	0.996	-	0.996	-
x: 業績	0.960	0.114	0.969	-	0.957	0.306

考察

- 機械学習による高精度なコーディングに寄与する GoM 構造
- 各項目が独立しているため，データのメンテナンスが簡単
- 本研究で生成した分類器で異なる選挙の公約データをコーディング可能 ∴ 機械学習の導入 ⇒ ビッグデータ分析の敷居↓
- 今まで “見えなかった” 政治的現象を観察できる可能性↑[1]

参考文献

[1] Grimmer, J. & Stewart, B. 2013. “Text as Data.” *Political Analysis* 21(3): 267–297.
[2] 上神貴佳・佐藤哲也. 2009.「政党や政治家の政策的な立場を推定する–コンピュータによる自動コーディングの試み」『選挙研究』25(1): 61–73.
[3] 品田裕. 1998.「< 資料 > 選挙公約政策データについて」『神戸法學雑誌』48(2): 541–72.
[4] Saussure, Ferdinand de. 1916. Le Cours de linguistique générale. Lausanne.