4M1-3

米国選挙における潜在立候補者の得票を予測する手法の提案

A new approach to predict pre-candidate performance in the US election

箕浦 慶 *1　　松尾 豊 *1
Kei Minoura　Yutaka Matsuo

*1東京大学
The University of Tokyo

In recent years, predicting election results has caught people's attention around the world. This has been especially true in the United States. In 2008, Nate Silver, an American statistician, predicted the US presidential election results with a high degree of accuracy. Since then, large news organizations, such as Washington Post and New York Times, started to predict the US election results. However, standard approaches do not consider the problem of pre-candidate performance. Predicting pre-candidate performance is considered useful for political parties as it enables them to select candidates who are most likely to win the elections.In this study, we predict pre-candidate performance in the US election. We propose a prediction model using historical data from county-level election results and candidates' features.

1. はじめに

近年、世界各国で、メディアを中心として、国政選挙を事前に予測する動きが活発である。特に、世界一の経済・軍事力を誇るアメリカ合衆国の国政選挙の動向は、世界情勢に与える影響が大きく、国内のみならず世界中より高く注目される。選挙結果の予測の有用性は高いと考えられ、 既定立候補者 (選挙に立候補した候補者) の選挙結果を予測する様々な研究が行なわれてきた。特に、政治経済指標、候補者のプロフィールより、大統領選挙の結果を予測する手法が研究されてきたが、国レベルでの投票結果を予測するものが主で、メディアの注目を集めることはなかった。

選挙結果予測が、初めてメディアに注目されるようになったのは、2008 年である。アメリカの統計学者である Nate Silver 氏が、アメリカ合衆国大統領選挙の結果を、大統領選挙に関する複数の世論調査データを用いることで、50 州の内 49 州で結果を的中させた。以来、Washington Post や New York Times などメディアも、選挙結果を独自に予測し、発表するようになった。また、現在、新たに、選挙結果を予測する手法として、Facebook、Twitter などのソーシャルデータや、Google、Yahoo などの検索データなど、ウェブデータを用いて選挙結果を予測する研究が行なわれている。

しかし、上記にあげた研究は、いずれも既定立候補者の選挙結果を予測するものであり、潜在立候補者 (選挙に立候補する可能性がある者) の結果を予測する、つまり、潜在立候補者が立候補すれば どのくらいの結果を残すことができるのかを予測する研究は行われてこなかった。二大政党である共和・民主党にとっては、潜在立候補者の結果を予測できた方が、候補者の選定に役立ち、選挙において国政の運営に有利な立場に立つために、有効であると考えられる。

本研究では、潜在立候補者の結果を予測する手法を提案する。立候補者が決まる前に、潜在立候補者がどのくらい得票できるか予測する必要があり、その時点で取得することの出来るデータを元に予測を行なわなければならない。また、多様な潜在立候補者の得票の予測を行なうことができるよう、候補者のプロフィールを元に、候補者の得票を予測できる手法を提案する。

データセットとして、ウェブから独自に収集した、2004 年、2008 年、2012 年度の大統領選挙と 2004 年度の上院議員選挙における、カウンティーレベルでのアメリカ合衆国選挙結果データと候補者のプロフィールデータを用いて、各カウンティーが、候補者のプロフィールに対し、どのような投票結果を生成するか、予測モデルを作成する。予測モデルを作成するのに、線形回帰とニューラルネットワークを用いる。

本研究の結果、線形回帰を用いた予測モデルの平均絶対誤差が 14.9%であったのに対し、ニューラルネットワークを用いた予測モデルの平均絶対誤差が 30.6%であり、線形回帰の方が、予測モデルの作成に有効であることが分かった。本研究の提案手法が、従来の研究で用いられてきた政治経済指標による予測と同様の予測精度を得ることができることを確認した。また、カウンティーレベルで、潜在立候補者の得票数を予測する場合、カウンティーをクラスターに分けて、クラスターごとに予測モデルを作成すると、予測精度が上がることが分かり、今後、カウンティーレベルでの選挙予測に有効であると思われる知見が得られた。

本論文の構成は以下の通りである。まず、第 2 章では本研究と関連性の高い、アメリカ合衆国選挙の選挙予測に関する先行研究を紹介する。第 3 章では、本研究に用いたデータセットの取得方法や概観について述べる。第 4 章では、線形回帰とニューラルネットワークを用いたコンテンツベースフィルタリングの手法をベースに、協調フィルタリングのクラスターモデルを一部取り入れた提案手法について述べ、第 5 章では、データセットに対して行なった実験の結果と考察を述べ、第 6 章で今後の手法の発展の展望も含めて、結論を述べ、本論文をまとめる。

2. 先行研究

本章では、主にアメリカ合衆国の選挙予測に関する研究を、予測手法の素性となるデータの種類別に紹介する。

2.1 政治経済指標・候補者のプロフィールから選挙結果の予測を行なう研究

政治経済指標・候補者のプロフィールから選挙結果の予測を行なう研究は、主に、大統領選挙の国レベルでの結果の予測

連絡先: 箕浦 慶，東京大学，minoura@weblab.t.u-tokyo.ac.jp

を行なってきた。Kramer 氏が確立した、有権者は、回顧的な投票選択をとり、特に、現大統領による経済政策の効果に影響されるという投票理論 [Kramer 71] を元に、多くの予測モデルが、経済指標を素性として用いている。実質 GDP 成長率、実質 GNP 成長率を素性として用いている研究が多い。実質 GNP 成長率を素性としてモデルに用いている代表的な研究として、Beck 氏らによる研究 [Lewis-Beck 84] がある。他にも、Erikson 氏と Wlezien 氏による研究 [Wlezien 96] では、景気動向指数 (Index of Leading Economic Indicators)、所得成長率を素性としてモデルに用いている。また、予測の精度を上げるために、経済指標とともに、政治指標を、素性として用いるのが一般的である。Abramowitz 氏の研究 [Abramowitz 88] は、現職大統領の支持率、在職期間を素性としてモデルに用いている。政治経済指標を用いて予測モデルを作成する研究が多いが、いつ選挙の結果を予測するのか、または、選挙の結果に影響を与える指標を何と考えるかによって、予測モデルに用いる素性は異なり、Armstrong 氏と Graefe 氏の研究 [Armstrong 11] では、候補者のプロフィールを、選挙の結果に影響を与えるものとして、素性に用いている。

2.2 世論調査から選挙結果の予測を行なう研究

世論調査は、有権者にどちらの候補者に投票するか聞くことで、選挙を予測するための直接的なデータを取得することができる一方、偏りのないデータを取得するのが難しいことでも知られる。調査方法、サンプルの選択方法、質問する言葉などによって、一方の候補者に、実際の結果 よりも、結果が偏ることがある。政治経済指標・候補者のプロフィールから選挙結果の予測を行なう研究でも、ある特定の機関による世論調査を用いた研究があるが、ここで紹介する世論調査データを用いる研究は、複数の機関による世論調査を用いることで、世論調査が持つ偏りを少なくして、高い予測精度を実現している。複数の世論調査を、サンプル数、実施時期、過去の正確さなどの要素から、各世論調査に重みをつけた結果を統合して用いることで、予測の精度を上げている。ただし、世論調査を用いるだけでは、選挙終盤まで、高い精度での予測ができないので、選挙序盤では、従来の政治経済指標・候補者のプロフィールより選挙結果の予測を行なう手法を用 いて結果を予測し、投票日に近づくにつれ、複数の世論調査を統合した素性を、徐々に重みをつ けながら予測に用いるモデルが多い。代表的な研究に、Linzer 氏の研究 [Linzer 13] があり、Nate Silver 氏や Washington Post など各メディアも類似したモデルを用いて予測を行なっている。また、従来の手法が、大統領選挙の国レベルでの予測を主に行なってきたのに対し、州レベルでの世論調査を用いることで、州レベルでの大統領選挙の結果を高い精度で予測している。ここ数年は、大統領選挙だけではなく、連邦議会議員選挙の予測も行なわれている。

2.3 ウェブデータから選挙結果の予測を行なう研究

世論調査の替わりに、有権者の民意を反映させるデータとして、ウェブデータを用いて選挙結果を予測する研究がある。または、近年、候補者や有権者が、選挙にウェブを活用するのが一般的であり、候補者のウェブでの影響力を、社会的影響力として捉え、選挙予測に用いた研究もある。ウェブデータを用いた選挙予測の研究は、ウェブデータを、有権者の民意を反映したデータ として予測に用いる研究と、候補者の社会的影響力を反映したデータとして予測に用いる研究に大きく分けることができる。

前者の研究としては、ソーシャルメディアや検索エンジン上での、有権者の候補者への反応を、 選挙結果の予測に用いた研究がある。ソーシャルデータを用いた研究としては、ドイツ議会選挙の予測ではあるが、Tumasjan 氏らが政党名を含むツイートの数より、各政党の得票数を予測する 研究 [Tumasjan 10] を行った。また、Brendan 氏らの研究 [O'Connor 10] では、アメリカ合衆国大統領選挙の候補者名 や選挙関連語などを含むツイートの感情を分析し、世論調査と相関関係があることを示し、Shi 氏らの研究 [Shi 12] では、候補者名が含まれているツイート・リツイート数を素性として用いて、共和党の大統領予備選挙を予測している。検索データを用いた研究としては、Lui 氏ら [Lui 11] と Chen 氏ら [Chen 12] が、各々、Google トレンドを用いてアメリカ合衆国の選挙予測を行なっている。また、日本衆議院議員選挙ではあるが、政党名や候補者名の Yahoo!での検索量を用いて、Yahoo! Japan が、選挙予測を行なっている。後者の研究としては、ソーシャルメディア上での、候補者の影響力を、選挙結果の予測に用いた研究がある。ニュージーランドの総選挙の予測ではあるが、Camerona 氏らは、Facebook と Twitter での候補者の持つネットワーク (Facebook の友達や、Twitter のフォロワーの数など) の素性を用いて、選挙結果を予測した [Cameron 14]。また、こちらもアメリカ合衆国選挙の予測ではなく、日本の衆議院選挙の予測ではあるが、那須野氏らの研究 [Kaoru 14] で、Twitter における候補者の情報拡散力から 選挙結果を予測したものがある。

2.4 本研究との関連性

先行研究が、既定立候補者の予測であったのに対し、本研究では、潜在立候補者の得票予測を行なう。また、これまで研究されたことのないカウンティーレベルでの予測を行なう。候補者のプロフィールを元に、潜在立候補者の得票を予測できるモデルを作成するが、その際、政治経済指標を用いた従来手法との予測精度の比較についても行なう。今回は、データ収集ができなかったため用いることはできなかったが、ウェブデータから選挙結果の予測を行なう研究で用いられた指標を、本研究の予測モデルに取り入れることも、今後考えることができる。また、本研究では、早い段階での選挙結果予測が求められるため、世論調査は予測モデルに用いないが、カウンティーレベルでの予測モデルを作成することで、世論調査から選挙結果の予測を行なう研究にも、今後寄与することができるのではないかと考えられる。

3. データセット

本章では、本研究に用いたデータセットの取得方法と概観について述べる。潜在立候補者の得票数を予測するために、候補者のプロフィールを元に、カウンティーレベルでの候補者の得票を予測するモデルを作成する。そのため、過去のアメリカ合衆国選挙結果データと候補者のプロフィールデータを取得する必要があった。

ウェブマイニングの手法によって、Washington Post のウェブサイトより 2004 年、2008 年、2012 年度の大統領選挙と 2004 年度の上院議員選挙の候補者のプロフィールデータ (71 名の候補者) を取得した。取得したプロフィールデータより、選挙出馬時の年齢、人種、宗教、出生州、居住州、職業、学歴ポイントを予測モデルの素性として作成した。

また、Washington Post の REST API より 2004 年、2008 年、2012 年度の大統領選挙と 2004 年度上院議員選挙のカウンティーレベルの選挙結果データ (内訳は、2004 年度大統領選挙の結果データ 6,222 件、2008 年度大統領選挙の結果データ 6,222 件、2012 年度大統領選挙の結果データ 6,218 件、2004

年度大統領選挙の結果データ 4,068 件となる) を取得した。

4. 提案手法

本章では、アメリカ合衆国選挙の潜在立候補者の得票を予測するために作成した予測モデルと、その実験結果について述べる。

4.1 問題の定式化

アメリカ合衆国選挙で、潜在立候補者が立候補すれば、どのくらい得票を獲得するかを予測することが目的である。潜在立候補者の得票を予測するための予測モデルに用いる素性として、以下の条件が求められる。

- 大統領選挙と連邦議員選挙に出馬する候補者を選択するための共和・民主党による予備選挙は、早いもので、選挙が実施される年度の 1 月に実施されるため、その時点に、取得可能である
- 潜在立候補者のそれぞれの特性を反映するもので、潜在立候補者ごとに異なる多様な予測結果を生み出すことができる

以上の条件を満たす素性として、潜在立候補者のプロフィール素性がある。大統領選挙と連邦議員選挙の共和・民主党による予備選挙が行なわれる以前に取得可能であり、また潜在立候補者ごとに、異なるプロフィールを持つので、プロフィール素性を予測モデルに組み入れると、潜在立候補者それぞれの特性を反映した異なる予測結果を取得できる。

また、より多様な潜在立候補者の予測に対応できるモデルを作成したい。そこで、推薦システムでユーザがアイテムを評価しているように、カウンティーが候補者に投票 (評価) していると捉え、コンテンツベースフィルタリングの手法をベースに、協調フィルタリングのクラスターモデルを一部取り入れた予測モデルを作成する。

推薦システムにおけるカウンティーと候補者の関係を、図 1 に表した。例えば、California 州の Orange County で、2012

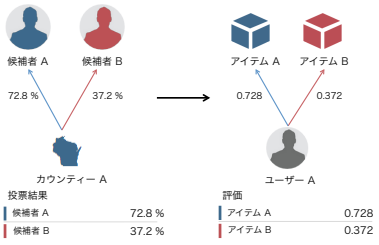

図 1: カウンティーと候補者の関係

年に実施された大統領選挙への投票では、共和党の大統領候補者である Mitt Romney 氏が 54.2%の票を、民主党の大統領候補者である Barack Obama 氏が 45.8%の票を獲得した。この投票結果を、Orange County が、0 から 1 の評価基準の中 (1 が最大基準である) で、Mitt Romney 氏に対し 0.542 の評価を、Barack Obama 氏に対し、0.458 の評価を行ったと本研究では捉える。

4.2 予測モデル

推薦システムに用いられている手法を用いて、ユーザがまだ評価したことのないアイテムや新しい アイテムに対してどのような評価をするか予測するのと同様に、カウンティーが潜在候補者に対してどのような投票を行なうのか予測する。推薦システ ムの問題として、本研究を捉えることで、取得した選挙結果を、図 2 のようにカウンティーと候補者の評価値行列で表すことができる。各候補者はプロフィール素性を持ってい

	カウンティー A (州A)	カウンティー B (州A)	カウンティー C (州B)	カウンティー D (州C)	カウンティー E (州D)	カウンティー F (州E)	カウンティー G (州F)	カウンティー H (州G)	カウンティー I (州H)
大統領候補 A	0.422	0.319	0.242	0.653	0.511	0.433	0.281	0.754	0.434
大統領候補 B	0.584	0.690	0.763	0.453	0.492	0.575	0.720	0.253	0.572
上院議員候補 A	0.467	0.332							
上院議員候補 B	0.542	0.674							
上院議員候補 C			0.736						
上院議員候補 D			0.277						
上院議員候補 E				0.473					
上院議員候補 F				0.539					
上院議員候補 G					0.566				
上院議員候補 H					0.648				
上院議員候補 I						0.432			
上院議員候補 J						0.571			
上院議員候補 K							0.717		
上院議員候補 L							0.298		

図 2: カウンティーと候補者の評価値行列

るため、評価値行列を元に、アイテム (候補者) の特徴を用いて評価値を予測するコンテンツベースフィルタリングの手法を用いることができる。コンテンツベースフィルタリングでは、推薦を機械学習の問題として捉え、評価値行列の列ごとに、入力値であるプロフィール素性 X に対して、評価値 y を出力する予測モデル

$$f : X \rightarrow y$$

を作成して、アイテムの推薦に用いる手法がある。本研究では、この手法を用いる。また、より多様な潜在候補者予測に対応したモデルを作成するために、過去に類似した投票行動を行なったカウンティーの過去の投票をモデルに反映させることのできるクラスターモデルの協調フィルタリングの手法を取り入れた。K-means 法によって、各カウンティーを K 個のクラスターに分類し、評価値行列のデータを、クラスター別に統合し、クラスターごとに予測モデルを作成する。

5. 実験

評価値行列の列ごとに作成する予測モデルの関数として、代表的な線形モデルである線形回帰 (従来の選挙結果の予測に用いられてきた) と、代表的な非線形モデルであるニューラルネットワークを用いて実験を行なう。また、その際、クラスタリングによるデータの統合が、実際に予測精度の向上に寄与するかどうかを確認するために、クラスタリングを行なう場合と行なわない場合の両方で、予測モデルを作成して実験を行なう。また、本研究の手法が有用であるかを確認するため、従来手法で用いられている経済社会指標を素性として用いた予測モデルでも実験を行ない、本研究の手法との予測精度を比較する。予測モデルの評価方法であるが、データセットに対する過学習を防ぐために、K=10 の K-fold 交差検定を行なってモデルの評価を行なう。 評価方法の尺度に、予測値と実測値にどれだけの誤差があるのか、平均絶対誤差を用いる。最終的に評価するための誤差は、10 分割交差検定を 10 回行なって得られた各検定での平均絶対誤差の平均をとる。

5.1 実験結果

表 1 の実験結果を得た。

表 1: 実験結果の概要

予測モデル	用いた素性	クラスタ	誤差
線形回帰	X_{age}, ... , X_{party}	なし	38.0%
線形回帰	X_{age}, ... , X_{party}	あり	14.9%
ニューラル	X_{age}, ... , X_{party}	なし	41.7%
ニューラル	X_{age}, ... , X_{party}	あり	30.6%
線形回帰 (従来手法)	X_{gdp}, $X_{approval}$	あり	14.5%

5.2 考察

本研究で提案したプロフィール素性を用いた線形回帰とニューラルネットワークによる予測モデルであるが、どちらの予測モデルにおいても、クラスタリングが予測精度を向上するために有効な 手法であることが分かる。線形回帰を用いたモ デルの方が、ニューラルネットワークを用いた予測モデルよりも有用であることが分かる。ニューラルネットワークで高い予測精度が得られ なかったのには、学習データの数が少なかったため、過学習が起きたと考えられる。また、本研究で提案した予測モデルは、従来手法とほとんど変わらない予測精度を得ており、プロフィールのみを素性として用いているので、候補者を選定するのに適している。

ただし、作成した予測モデルの平均絶対誤差がいずれも 14% を超えており、アメリカ合衆国での候補者の擁立に有用な情報になるには、もう少し予測精度を上げる必要がある。 最初に、予測モデルの精度を上げるために、データセットをより拡充する必要がある。より、多くの学習データが有れば、よりミクロなカウンティーのクラスタリングが可能になり、予測精度が向上すると考えられる。ただし、連邦議員選挙の候補者のプロフィールデータを取得するのは比較的難しい。複数のウェブサイトへのウェブマイニングとテキストマイニングの手法を確立することが必要である。また、本研究では取得することができず用いることはなかったが、ソーシャルメディア上での影響力や、経済力を候補者の社会的な影響力を表す指標として素性に加えることができると考えられる。また、データセットの拡充以外にも、推薦システムの手法を発展させることが考えられ、人口構成や経済力などカウンティーの特徴も考慮した推薦システムの手法を用いることで、より予測の精度が上がるかもしれない。

6. まとめ

本研究では、ウェブより独自に収集したアメリカ合衆国選挙のカウンティーレベルでの選挙結果と、候補者のプロフィールを用いて、潜在立候補者の得票を予測する手法の提案を行なった。先行研究は、既存立候補者の選挙の結果を予測したものであり、潜在立候補者の選挙の結果を予測した研究はなかった。また、カウンティーレベルでの選挙結果の予測に関しても研究されておらず、候補者のプロフィールを用いた選挙結果の予測に関してもほとんど研究されてこなかった。

ユーザーのアイテムへの評価を予測する推薦システムを参考に、カウンティーの潜在立候補者への投票 (評価) を予測するモデルを作成した。カウンティーをクラスタリングすることで、他カウンティーの候補者への評価も予測モデルの作成に反映させることで、モデルの精度を上がるのに有効であることを示した。また、従来の研究で用いられてきた政治経済指標による予測と同様の精度が得られることを確認した。

本研究によって、今後のカウンティーレベルでの選挙結果予測の研究に有用となると考えられる推薦システムの手法を用いた予測モデル、また潜在立候補者を選定するのに適した予測モデルを提案することが出来たと考えられる。

参考文献

[Abramowitz 88] Abramowitz, A. I.: An improved model for predicting presidential election outcomes, *PS: Political Science & Politics*, Vol. 21, No. 04, pp. 843–847 (1988)

[Armstrong 11] Armstrong, J. S. and Graefe, A.: Predicting elections from biographical information about candidates: A test of the index method, *Journal of Business Research*, Vol. 64, No. 7, pp. 699–706 (2011)

[Cameron 14] Cameron, M. P., Barrett, P., and Stewardson, B.: Can Social Media Predict Election Results? Evidence from New Zealand, *Journal of Political Marketing*, No. just-accepted (2014)

[Chen 12] Chen, Y., Zhang, F., and Yue, Y.: Predicting US president election result based on Google Insights (2012)

[Kaoru 14] Kaoru, N. and Yutaka, M.: Twitter における候補者の情報拡散に着目した国政選挙当選者予測, 人工知能学会全国大会論文集, Vol. 28, pp. 1–4 (2014)

[Kramer 71] Kramer, G. H.: Short-term fluctuations in US voting behavior, 1896–1964, *American Political Science Review*, Vol. 65, No. 01, pp. 131–143 (1971)

[Lewis-Beck 84] Lewis-Beck, M. S. and Rice, T. W.: Forecasting presidential elections: A comparison of naive models, *Political Behavior*, Vol. 6, No. 1, pp. 9–21 (1984)

[Linzer 13] Linzer, D. A.: Dynamic Bayesian forecasting of presidential elections in the States, *Journal of the American Statistical Association*, Vol. 108, No. 501, pp. 124–134 (2013)

[Lui 11] Lui, C., Metaxas, P. T., and Mustafaraj, E.: On the predictability of the US elections through search volume activity (2011)

[O'Connor 10] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A.: From tweets to polls: Linking text sentiment to public opinion time series., *ICWSM*, Vol. 11, pp. 122–129 (2010)

[Shi 12] Shi, L., Agarwal, N., Agrawal, A., Garg, R., and Spoelstra, J.: Predicting US primary elections with Twitter, *URL: http://snap. stanford. edu/social2012/papers/shi. pdf* (2012)

[Tumasjan 10] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment., *ICWSM*, Vol. 10, pp. 178–185 (2010)

[Wlezien 96] Wlezien, C. and Erikson, R. S.: Temporal horizons and presidential election forecasts, *American Politics Research*, Vol. 24, No. 4, pp. 492–505 (1996)