機械学習を用いた科研費テキストデータに基づく審査区分の推定と応用

国立大学法人 福島大学 Fukushima University

RA協議会第７回年次大会　福島大学　荻 多加之

概要

論文の評価あるいは研究費申請において、各研究が属する学術分野を的確に分類することは重要である。これまで研究者本人あるいは当該分野の専門家によって分類が行われているが、大量の研究課題を客観的に分類することは非常に困難である。

近年、自然言語処理や機械学習の進歩により研究分野の自動分類が可能になりつつある（Goh et al. 2020）。そこで本研究では機械学習を用い科研費の申請書概要データの自動分類を試みた。その結果、各申請書の審査区分を80%を超える精度で分類できることを確認した。

今後は学術変革領域研究や他の研究費などへの適用も試みる。

方法

分類に用いたデータ

科学研究費助成事業データベースからダウンロードした2018年度～2020年度の３年間の基盤研究（S）のデータを用いた。

2018年度については「研究実績の概要」の項目、2019年度、2020年度は「研究開始時の研究の概要」の項目を抽出。合計241課題から英語で書かれた課題（3件）、未記入（5件）を除いた233課題を用いた。

Task1では大区分A,C,Gの合計64課題、Task2では大区分C,E,G,Iの合計87課題の分類を試みた。

大区分	課題数	Task1	Task2
A	18	○	
B	45		
C	25	○	○
D	35		
E	19		○
F	12		
G	21	○	○
H	10		
I	22		○
J	16		
K	10		
合計	233	64	87

単語の抽出、及びターム文書行列への変換

オープンソースの形態素解析エンジンであるMeCabのR言語上での実装であるRMeCabを用い以下のステップで行った。(MacBookAir Early2014, 1.7GHz Core i7, 8GB)

1. ダウンロードした科研費データcsvファイルの前処理（課題名・審査区分・概要の抽出）
2. 分類に用いる課題を選択
3. 各課題概要から名詞のみを抽出
4. 課題ごとに各名詞の出現頻度をカウントし、ターム文書行列を作成
5. tf-idf（term frequency - inverse document frequency）により、各単語の重要度を評価できる指標に変換
6. 各文書ベクトルの大きさが1となるように正規化

tf-idf

$$tf = \frac{\text{概要Aにおける単語Xの出現頻度}}{\text{概要Aにおける全単語数}}$$

$$idf = log\left(\frac{\text{全概要数}}{\text{単語Xを含む概要数}}\right)$$

tf-idfについて
こちらのYouTube動画が参考になりました。

教師なし学習による分類

各文書ベクトル間のユークリッド距離を各課題間の類似度の指標として分類を行う

1. 「用いた大区分数」をクラスターの数としてk-meansクラスタリングを適用
2. MDS（多次元尺度構成法）を用いターム数次元空間を２次元平面に削減し、分類結果を可視化

教師あり学習による分類

SVM（サポートベクトルマシン）を用いて分類し、精度を検証する

1. データを大区分のバランスを保ちつつランダムに８分割し、１つをテストデータして、残りをトレーニングデータとする。選択するテストデータを全８パターン試し、最後に検証結果を平均する（層化８分割交差検証）
2. トレーニングデータを用いてSVMにおけるハイパーパラメータ（c,γ）をグリッドサーチにより最適化
3. 決定したハイパーパラメータを用い分類器を作成
4. 作成した分類器に対してテストデータを適用し分類を行う

R言語によるSVM
こちらのサイトを参考にしました。

SVMの理論と実装
こちらのYouTube動画が参考になりました。

分類精度の検証

- **正確率**（Accuracy）　：真陽性数の和／全体の数
- **適合率**（Precision）　：真陽性／（真陽性＋偽陽性）
- **再現率**（Recall）　：真陽性／（真陽性＋偽陰性）
- **F値**　（F1 score）　：適合率と再現率の調和平均

適合率
ある分野Aに分類された概要の中で、実際にその分野の研究である概要の割合
例：検査で陽性になった人の中で、実際にその病気に罹患している人の割合

Task1 Aの例		実際の区分 A	C	G
推定区分	A	真陽	偽陽性	
	C	偽陰性		
	G			

再現率
ある分野Aの概要の中で、分類によって正解の分野Aに分類された概要の割合
例：実際にその病気に罹患している人の中で、検査で陽性になった人の割合

Reference
Goh, Y.C., Cai, X.Q., Theseira, W. et al. Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics* **125**, 1197–1212 (2020). https://doi.org/10.1007/s11192-020-03614-2

結果

ターム文書行列

Task1の6課題11単語の例（正規化前）
各概要内では頻出するが、他の概要ではあまり登場しない（レアな）単語は高い値を示す。

	A2	A3	C7	C8	G1	G2
的	2.4	2.4	2.4	7.3	0	4.9
政策	4.7	9.4	0	0	0	0
財政	21.0	0	0	0	0	0
ナッジ	0	21.0	0	0	0	0
半導体	0	0	21.0	8.4	0	0
磁性	0	0	26.5	0	0	0
ナノ	0	0	0	12.6	0	0
細胞	0	0	0	0	3.3	9.9
RNA	0	0	0	0	24.0	0
顕微鏡	0	0	0	0	0	5.0
進行	0	0	0	0	0	0

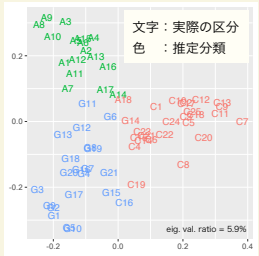
教師なし学習（k-means）

文書ベクトル間の距離（類似度）を指標にk-meansクラスタリングを適用し、MDSにより次元圧縮した平面にプロットした。4区分の場合のマッチの精度は60%程度であった。

・Task1（A, C, G）

		実際の区分 A	C	G	適合率
推定分類	A	17	0	0	100%
	C	1	24	1	92%
	G	0	1	20	95%
再現率		94%	96%	95%	

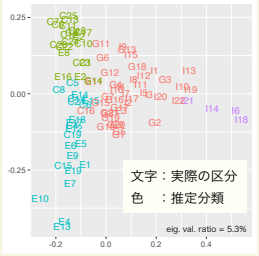
正確率：95%　　F値：95%



・Task2（C, E, G, I）

		実際の区分 C	E	G	I	適合率
推定分類	A	16	3	1	0	80%
	C	8	14	0	0	64%
	E	1	1	20	18	50%
	G	0	0	0	4	100%
再現率		64%	78%	95%	18%	

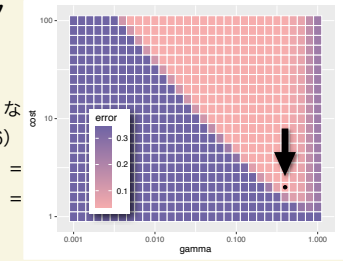
正確率：62%　　F値：59%



教師あり学習（SVM）

・ハイパーパラメータの最適化

Task1のfold1の例
エラーが最も小さくなる（error = 0.036）組み合わせ（cost = 2.00, gamma = 0.40）を採用



教師あり学習（SVM）

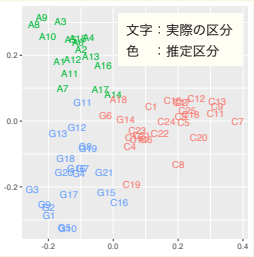
作成した判別器にテストデータを適用し区分を推定した。80%を超える精度で分野の推定が可能であった。

・Task1（A, C, G）

		実際の区分 A	C	G	適合率
推定区分	A	17	0	0	100%
	C	1	24	2	92%
	G	0	1	19	93%
再現率		94%	96%	91%	

正確率：94%　　F値：94%

計算時間：2時間40分

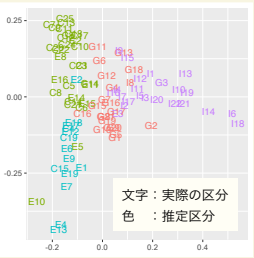


・Task2（C, E, G, I）

		実際の区分 C	E	G	I	適合率
推定区分	C	21	7	1	0	72%
	E	3	11	0	0	79%
	G	1	0	19	1	90%
	I	0	1	1	21	91%
再現率		84%	58%	90%	95%	

正確率：83%　　F値：82%

計算時間：5時間20分



考察

- 4つの大区分という限られた条件ではあるが、80%を超える精度で分野の推定が可能であった。
- 機械学習を用いた研究分野の自動分類は、学術変革領域研究や、JST・財団といった科研費と異なる区分の研究費を含めた研究力分析を、大量かつ迅速に行う際に有用であると考える。
- 機械学習を用いた研究申請書テキストによる分類は、教師データや計算資源が確保できれば、研究指向や起業意識の分類などにも活用できる可能性がある。
- どの区分で応募する方が採択の確率が高いかなど個別の案件については、機械学習の結果は参考程度とし、最終的には人が判断する方が望ましい。

今後の展開
BM25やDCNNなども試したい。

謝辞

本研究はCode for Research Administration (C4RA)の取り組みの一環である**R言語勉強会**の成果です。久保琢也様（横国大）及びご参加のメンバーに感謝申し上げます。