ISS-SP-009

学術文献データベースから将来有望な若手を推定する機械学習モデルの開発に向けた基礎的検討

土屋　裕太郎† 　井ノ上　寛人† 　藤田　正典†† 　寺田　隆雄††† 　鉄谷　信二†
† 東京電機大学大学院 　†† 政策研究大学大学院 　††† 千葉商科大学

1. はじめに

技術の発展には有望な若手の獲得や育成が必要であり，そのような人材を効率的に早期発掘する手段が求められている．研究人材に対する評価指標としては，現在，論文の被引用件数や h-index が参考にされているが，これらの指標は論文発表後に評価が蓄積される遅行性を有するため，学生や若手の評価には適さない．これらの背景から，著者らは将来の研究開発を担う若手の評価指標として，論文の共著ネットワークの成長特性に着目している．これまでの成果として，共著ネットワークの媒介中心性の成長性を評価すると，JSPS 特別研究員に採択され得る若手を学術文献データベースから早期に抽出できる可能性が示唆されている[1]．

本稿では，媒介中心性と他の特徴量を組み合わせた機械学習モデルを構築し，学術文献データベースに収録されている情報から JSPS 特別研究員をより正確に推定する上での課題を検討した．

2. 機械学習モデルによる JSPS 特別研究員の推定

JST が提供する学術文献データベース JSTPlus および JMEDPlus に収録されているデータのうち，1974 年から 2019 年までに発行された文献およそ 3,871 万編分のデータを解析の対象とし，研究者 ID ごとに初めて文献情報が収録されてから 3 年分のデータを集計して研究者 ID が持つ特徴量とした．機械学習モデルは，精度の高さとモデルの解釈性を考慮して LightGBM[2]を採用し，モデルの構築に用いる特徴量は，(1)論文数，(2)所属機関数，(3)論文発行国，(4)論文発行年数，(5)媒介中心性，(6)媒介中心性の上昇率の 6 項目とした．正解ラベルは JSPS 特別研究員に採択された経歴がある研究者 ID に付与した．

データはランダムサンプリングにより 8:1:1 の割合で，学習データ，検証データ，テストデータに分割した．学習データはモデルの学習に，検証データはモデルのハイパーパラメータチューニングに，テストデータは最終的なモデルの評価に用いた．なお，JSPS 特別研究員の数は全研究者の数に対して極めて少ないため，データの数に対する正解ラベルの数が等しくなるようにダウンサンプリングを適用した．モデルの評価指標は，適合率(precision)，再現率(recall)，ROC-AUC とした．検証実験の結果を表 1 に示す．

表 1　構築したモデルの評価

	学習	評価	テスト
適合率(precision)	0.80	0.69	0.68
再現率(recall)	0.65	0.60	0.56
ROC-AUC	0.96	0.94	0.94

3. 結果および考察

表 1 より，本稿で構築した機械学習モデルは ROC-AUC が 0.94 と高いことが示された．したがって，このモデルは正解ラベルである確率に基づいて研究者 ID を並べる精度は高いといえるため，JSPS 特別研究員相当に将来有望と推定される若手を降順に並べたい場合などに有用と考えられる．

一方で，このモデルは適合率(precision)に対して再現率(recall)が低いことから，JSPS 特別研究員であるという判定を誤ることは少ないが，JSPS 特別研究員でない者に誤って正解ラベルを付けしてしまう傾向があるといえる．この理由として，学習データに対するラベル付けの精度が十分でなかった可能性が挙げられる．具体的には，学術文献データベースには JSPS 特別研究員であるという情報が収録されていないが，KAKEN データベースを参照すると JSPS 特別研究員である可能性が極めて高い研究者 ID が多数見受けられた．

5. まとめ

本稿では，学術文献データベースに収録されている情報に機械学習アルゴリズムを適用すると，JSPS 特別研究員を推定する上で ROC-AUC が高いモデルを構築できることを示した．今後の課題として，モデルの再現性向上のために，研究者 ID の名寄せ方法について検討することが挙げられる．

本稿では，JST が提供している科学技術文献データベースの利用にあたって，株式会社ジー・サーチから協力を受けた．関係者の皆様に感謝の意を表する．

参考文献

[1] M. Fujita, et al., “Analyzing Promising Researchers using Network Centralities of Co-authorship Networks from Academic Literature,” New Generation Computing, https://doi.org/10.1007/s00354-020-00102-2, 2020.

[2] K. Guolin, et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” Proc. of the 31st International Conference on Neural Information Processing Systems, pp.3149-3157, 2017.