5J-08

機械学習を用いたグラフアルゴリズムの実行時間予測に関する一検討

深澤祐輔 [†1]　小松一彦 [†2]　佐藤雅之 [†3]　小林広明 [†3]

東北大学工学部機械知能・航空工学科 [†1]　東北大学サイバーサイエンスセンター [†2]
東北大学情報科学研究科 [†3]

1　はじめに

近年，グラフデータはSNSやオンラインショッピングなど，社会においてさまざまな分野で応用されており，グラフデータを高速に処理し，サービスを提供する需要が高まっている．グラフデータとはノードと呼ばれる頂点同士を，エッジと呼ばれる辺によって連結した構造を持つネットワーク状のデータのことである．例えばSNSの場合，アカウントをノードで表し，二つのアカウントにフォロ関係があれば，ノード同士をエッジによって結ぶことで関係を表す．しかしながら，グラフデータはそれぞれ異なるサイズやトポロジを持つため，解析に用いるグラフアルゴリズムによっては，解析時間が膨大になってしまう場合がある．そのため解析前に，グラフデータに適したグラフアルゴリズムを把握する必要がある．

そこで本研究では，グラフデータの特徴量を説明変数とした機械学習によって，グラフアルゴリズムの実行時間の予測をする．事前に複数のグラフデータをグラフアルゴリズムによって解析し，得られた実行時間を目的変数，グラフデータの特徴量を説明変数とした学習データを作成する．学習データを用いた回帰モデルによって，グラフアルゴリズムの実行時間の予測を行う．これにより適切なグラフアルゴリズムを選択し解析をすることができる．

2　グラフ処理フレームワーク

本検討では，複数のグラフアルゴリズムを同一の環境で実行できるグラフ処理フレームワーク Vector Graph Library (VGL)[1] を用いて，グラフデータの解析をおこない，実行結果を機械学習に用いる．VGLには，14種類のグラフアルゴリズムが実装されており，例えばある始点から終点までの最短経路を求めるアルゴリズムとして広く用いられるDijkstra法や，Webページの重要度を決定するためのアルゴリズムであるページランクなどがある．解析するグラフデータはWebサイト [2] からダウンロード可能であり，Mapデータや SNSユーザデータといった様々な種類のグラフデータが存在する．VGLは汎用CPUだけでなく，ベクトルエンジン (VE) やGPUといったアクセラレータにおいても実行可能で，プロセッサごとの実行結果の評価が可能である．

3　機械学習を用いた実行時間予測

本提案ではグラフアルゴリズムの実行時間を事前に予測するために，グラフデータの特徴量を学習データとして回帰モデルに入力する．学習データの説明変数は，目的変数を予測する上で指標となる特徴量であることが必要となる．本提案では，グラフデータのサイズを示すエッジ数とノード数，最大次数，平均次数，グラフデータの種類，エッジ情報である有向無向，そしてエッジ数を最大次数で除した数値を説明変数として設定し，学習データを作成する．

予測を行う回帰モデルとして，線形回帰，木構造モデルであるランダムフォレストとXGBoostの計3種類を用いて，グラフアルゴリズムの実行時間を予測する．線形回帰モデルは

$$y(\omega, x) = \omega_0 + \omega_1 x_1 + \ldots + \omega_p x_p$$

のような回帰式を用いて，説明変数から目的変数の値を予測する．ランダムフォレストとXGBoostはともに複数の決定木を作成し予測をおこなう回帰モデルであり，ランダムフォレストはバギング，XGBoostはブースティングという手法を用いて予測を行う．

A Study on An Execution-Time Prediction Method for Graph Algorithms Using Machine Learning

Yusuke Fukasawa[†1], Kazuhiko Komatsu[†2], Masayuki Sato[†3], Hiroaki Kobayashi[†3]
Department of Mechanical and Aerospace Engineering, Tohoku University[†1]
Cyberscience Center, Tohoku University[†2]
Graduate School of Information Sciences, Tohoku University[†3]

4 評価

4.1 実験条件

本提案では,Webサイトの被リンクのデータや，Mapデータを含む計1,306個のグラフデータを用いる．全グラフデータを14種のグラフアルゴリズムによって解析し，それぞれの実行時間を計測する．実行時間は1種類のグラフデータあたり600秒を制限時間として計測する．実行環境は，Intel Xeon Gold 6126 × 2 ソケットを用いる．

解析によって得られた実行時間と，Webサイトから取得したグラフデータの特徴量から学習データを作成する．それを回帰モデルに入力し，14種類のグラフアルゴリズムの実行時間をそれぞれ予測する．機械学習では，得られた1,306個分の学習データを8：2に分割し，それぞれ訓練データ，検証データとして用いる．回帰モデルはともにscikit-learn ver 0.24.2およびxgboost 1.7.1を用いて実装する．線形回帰と木構造モデルとで予測を行い，予測精度について比較を行う．

回帰モデルによる実行時間予測の精度を評価する指標として，決定係数を用いる．決定係数は，予測値が実際の目的変数の値とどれだけ一致しているかを示しており，

$$\text{決定係数} = 1 - \frac{\sum_{i=1}^{n}(x_i - p_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

から導出できる．最大で1をとり，1に近い数値であるほど高精度で予測ができていることを表す．

4.2 結果と考察

図1に，回帰モデルの決定係数と，比較対象として，エッジ数のみを特徴量とした最小二乗法の決定係数を示す．決定係数が負の数になったグラフアルゴリズムは，値を0として示し，回帰モデルごとの決定係数の平均をaverageとして示す．図1から，prやhitsといったグラフアルゴリズムは，いずれの回帰モデルにおいても高い精度で予測できていることがわかる．さらにランダムフォレストとXGBoostを用いた場合，bfsやmfにおける決定係数が高く，最小二乗法よりも予測精度が向上している．

prやhitsが多くの予測手法において決定係数が高い理由を考察する．グラフデータのエッジ数と実行時間とで相関係数を求めると，prは0.97，hitsは0.98と高い値を示す．そのため高い予測精度が得られたと考えられる．さらに木構造モデルがbfsやmf, tc-purdomsにおいて最小二乗法以上の精度となったのは，複数の特徴量を学習に用いたことで，より多くの情報を用いた学習が可能となったためである．

また，木構造モデルの予測が線形回帰に比べて精度が高いことがわかる．ランダムフォレストとXGBoostは，ともに決定木を複数作成しバギング，もしくはブースティングを行う手法である．決定木を用いることで，実行時間を予測するのに有効な特徴量が抽出され，高精度の予測が可能になる．一方で，線形回帰の場合，全ての特徴量を説明変数として回帰をおこなうため，正確な予測ができなかったと考えられる．

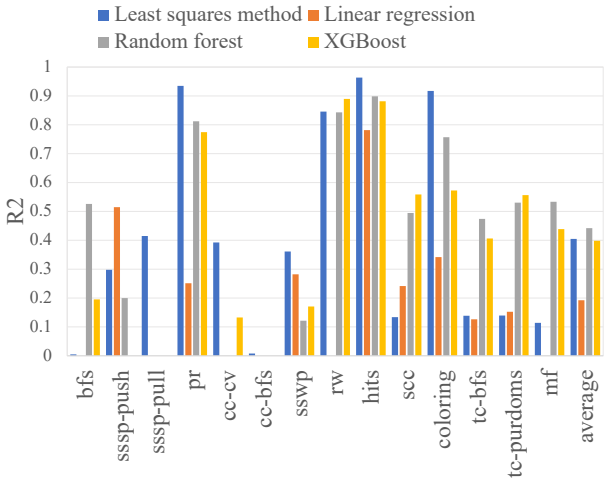

図 1: 各回帰モデルにおける決定係数

5 おわりに

本稿では，機械学習を用いたグラフアルゴリズムの実行時間予測手法を提案した．提案手法では，複数のグラフデータの特徴量を説明変数，事前に取得した実行時間を目的変数とした学習データを回帰モデルに入力することで実行時間の予測を行う．実験により，最小二乗法では予測が不十分であったグラフアルゴリズムでも，木構造系の回帰モデルを用いることで高い決定係数となる予測が可能となり，適切なアルゴリズムの選択ができることがわかった．今後は予測精度の低いグラフアルゴリズムの精度向上や，他の計算機を用いた結果を用いた予測等に取り組む．

謝辞

本研究の一部は，文部科学省「次世代領域研究開発」(高性能汎用計算機高度利用事業費補助金)「量子アニーリングアシスト型次世代スーパーコンピューティング基盤の開発」，文部科学省「次世代計算基盤に係る調査研究」新計算原理調査研究，科研費基盤A #19H01095，科研費基盤C #20K11838および JSPS-二国間交流事業 JPJSBP120214801 の補助を受けて実施している．

参考文献

[1] Ilya Afanasyev,VGL:Vector Graph Library https://vgl.parallel.ru/index.html
[2] Networks, http://konect.cc/networks/