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研究

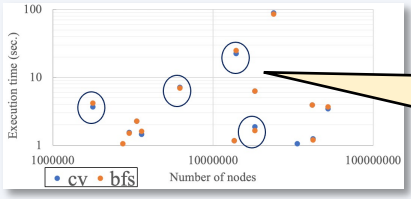
情報基礎科学専攻　小林・佐藤研究室　修士1年　深澤 祐輔

背景・課題

- グラフデータの高速処理によるサービス提供の需要の増大

- グラフアルゴリズム
  - グラフデータを解析するアルゴリズム
  - 探索、最短経路、フローネットワークなど目的に応じて使い分けがされる

2種類のアルゴリズムにおける解析実行時間の比較



先に解析し終えるアルゴリズムはデータごとに異なる

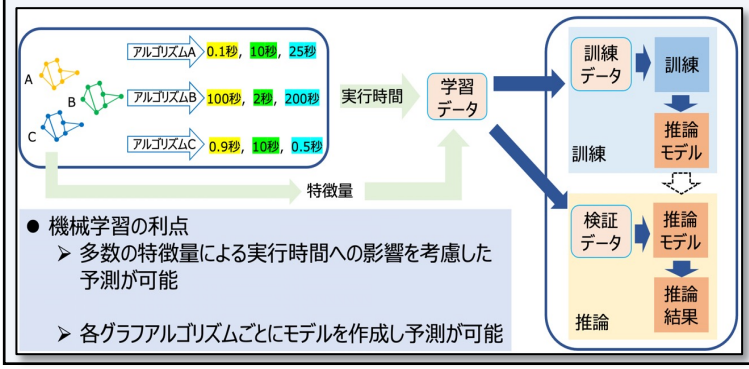
- 解析時間を短縮する適切なアルゴリズムの選択が必要

目的・提案手法

目的：解析実行前において適切なアルゴリズムを選択
提案手法：グラフデータの特徴量を用いて、機械学習によってアルゴリズムの実行時間を予測

提案手法の概要

1. 複数のグラフアルゴリズムを用いて実行時間を取得
2. グラフデータの特徴量と実行時間を学習データとし回帰モデルに入力



- 機械学習の利点
  - 多数の特徴量による実行時間への影響を考慮した予測が可能
  - 各グラフアルゴリズムごとにモデルを作成し予測が可能

学習に用いた特徴量

目的変数	説明変数	
実行時間	ノード数	ノード数/平均次数
	エッジ数	Wエッジ数/最大次数
	LCC サイズ	Wエッジ数/平均次数
	ノード数/最大次数	エッジ数/LCC サイズ
	エッジ数/最大次数	ノード数/平均次数

複数ある説明変数の中から、予測において重要な特徴量のみを抽出することが可能

機械学習モデル

線形回帰モデル
- Elastic Net

$$J(\theta) = MSE(\theta) + r\alpha\sum_{i=1}^{n}|\theta_i| + \frac{1-r}{2}\alpha\sum_{i=1}^{n}\theta_i^2$$

木構造系回帰モデル
- ランダムフォレスト
- XGBoost

評価

実行環境

- Intel Xeon Gold 6126
- グラフ処理フレームワーク「Vector Graph Library」
- グラフアルゴリズム 14種類
- グラフデータ数 1,306種類

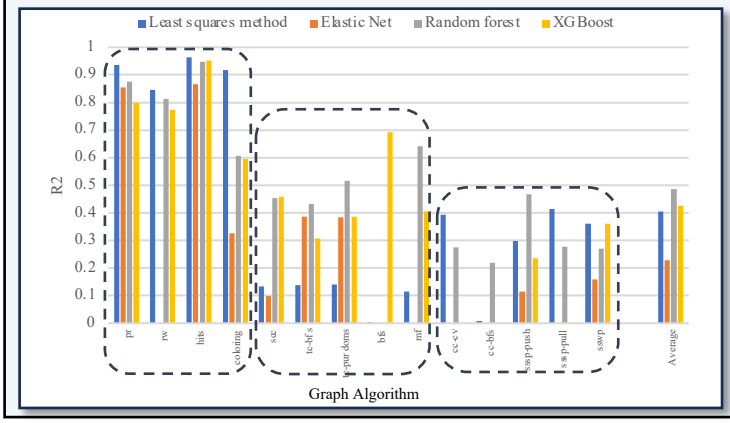
評価指標

- 決定係数(R2) データに対する推定されてた回帰式の当てはまり具合 最大で1

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_i)^2}$$

機械学習による予測結果

各回帰モデルにおけるアルゴリズムの実行時間予測の精度



結果から３種類のグループに分けられる
- 最小二乗法の精度が最も高い
  - pr, rw, hits, coloring,
- 木構造系アルゴリズムの予測精度が最も高い
  - scc, tc-bfs, tc-purdoms, bfs, mf, sssp-push
- いずれのアルゴリズムも予測精度が低い
  - cc-cv, cc-bfs, sssp-pull, sswp

Averageより木構造系回帰モデルが最小二乗法よりも予測精度が向上
- 複数の特徴量から予測に有効な特徴量を抽出し予測するため
- 線形回帰モデルは曲線的な実行時間の推移を正確に予測できない

結論

- 木構造回帰モデルを用い、最小二乗法以上の精度で予測可能
  - グラフアルゴリズムごとに実行時間を予測し最も早いアルゴリズムを選択可能

今後の研究計画

- 予測精度の更なる向上
  - ハイパーパラメータの調整
  - グラフアルゴリズムごとに特徴量を選択・追加

(URL) https://www.cal.is.tohoku.ac.jp/
(E-mail) yusuke.fukasawa.q1@dc.tohoku.ac.jp