

木アンサンブルモデルを利用した グラフ分類回帰問題に対する 効率的アルゴリズムの検討

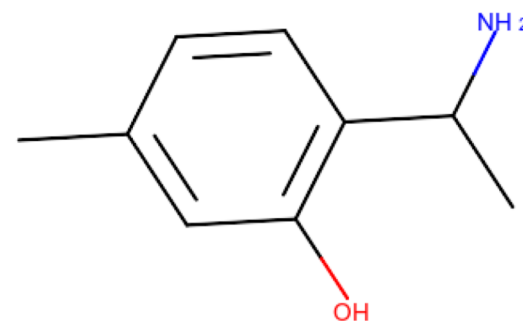
北海道大学大学院 情報科学研究科 情報理工学専攻

修士2年 白川 稜

背景

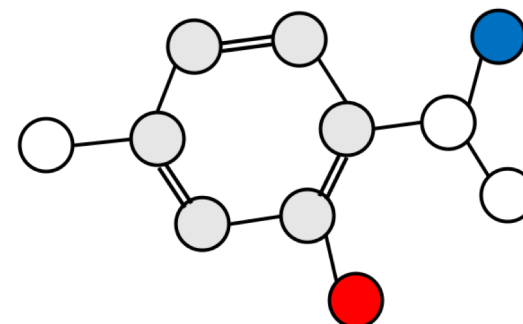
グラフは広く用いられる重要なデータ構造

- 低分子化合物の構造式
- RNA二次構造
- 自然言語処理における構文木



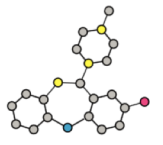
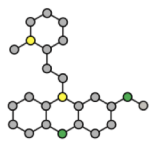
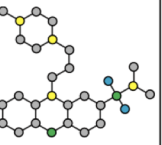
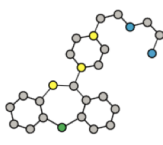
グラフデータからの教師付き学習

- 創薬の分野
- 生命科学や物質化学の分野



グラフ分類回帰問題

input: グラフデータ

G_1	G_2	G_3		G_n
			...	



予測器
 f



output: グラフの性質

y_1	y_2	y_3		y_n
0.1	0.7	1.2	...	0.9

一般的に、特徴量として
部分グラフの有無を利用

問題点

グラフサイズに対して
部分グラフの総数は指数関的に増加

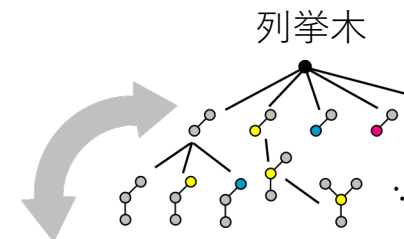
既存手法 (2step approach)

① 特徴ベクトル作成
(部分グラフパターン探索)

注： グラフサイズ
頻出度等の制約あり



② 任意の学習モデル



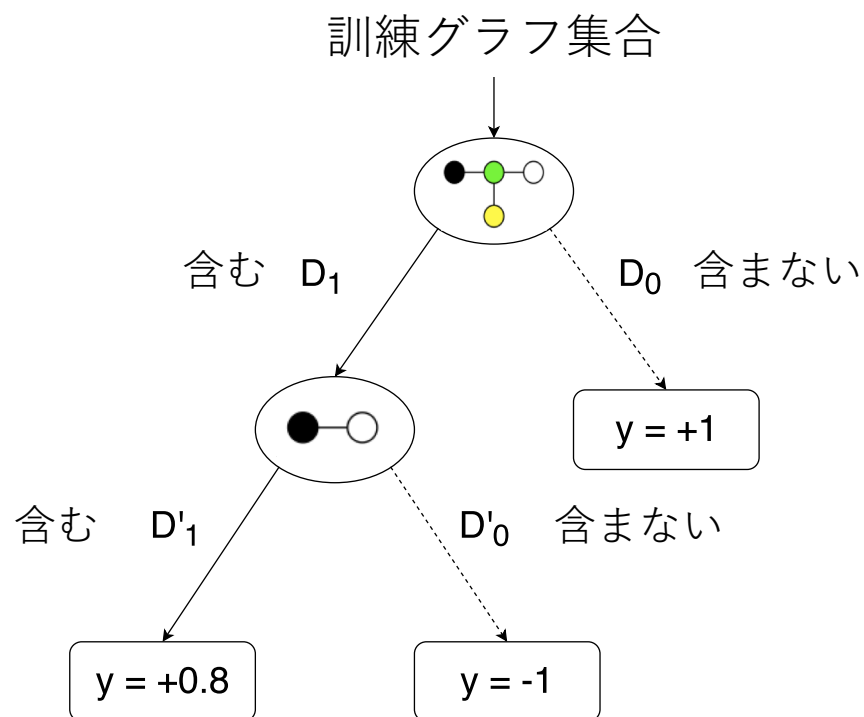
y	G									...
0.1		1	1	1	1	1	1	1	1	...
0.7		1	1	1	0	1	1	1	1	...
0.9		1	1	1	0	1	1	1	1	...
⋮	⋮									...
1.2		1	1	1	0	1	1	1	1	...

問題点：重要な特徴を見落とす恐れ

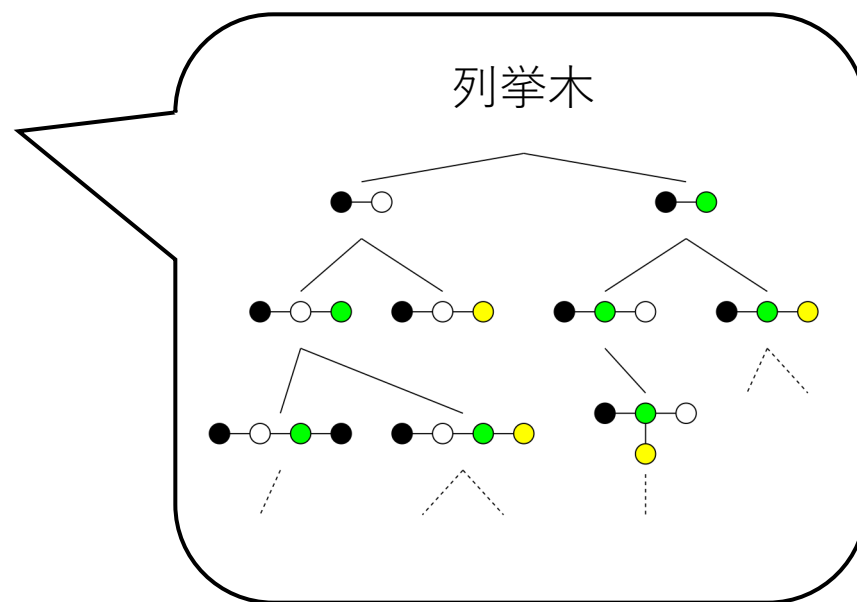
提案手法

モデルの学習と部分グラフ探索・選択を同時に行う
(陽に特徴ベクトルを作成しない)

回帰木

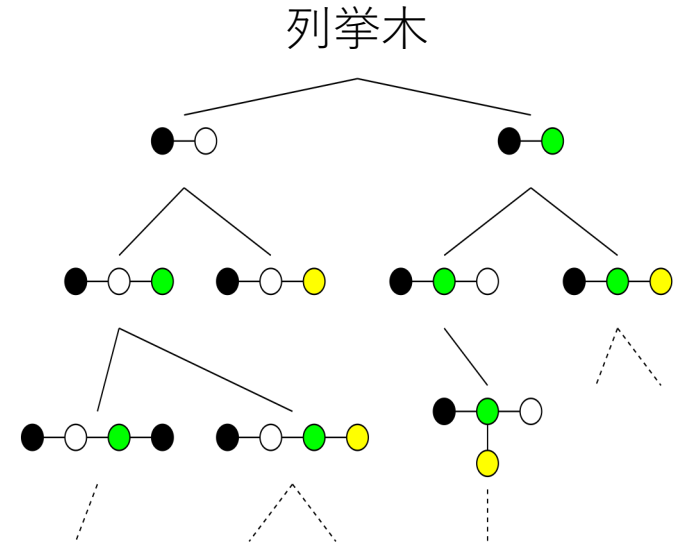


特徴探索



特徴探索

回帰木の分割において最も不純度が低くなるような部分グラフを探索する



枚举木の性質：子孫が親の拡大グラフ



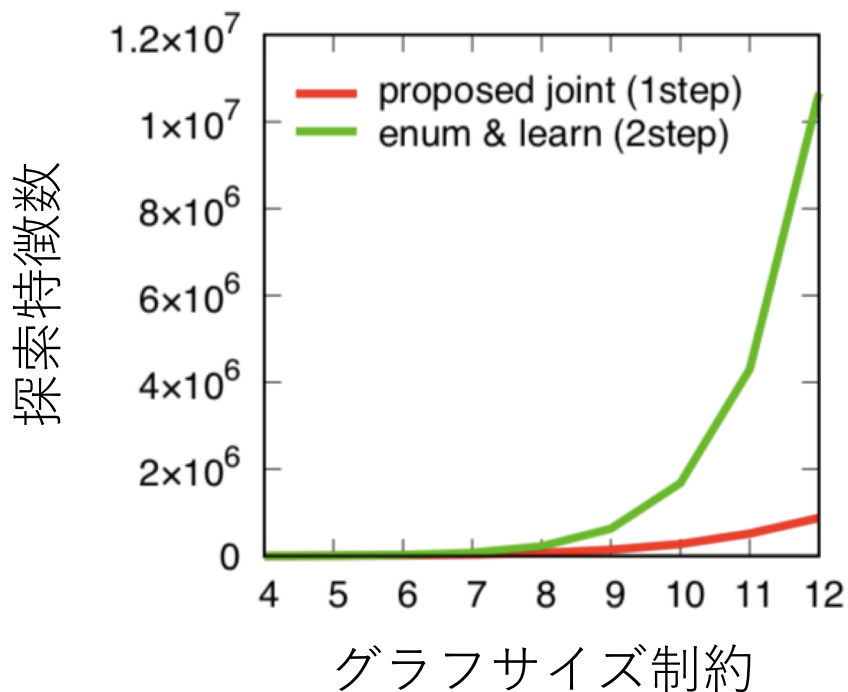
子孫の探索により得られる不純度の下限值が計算可能



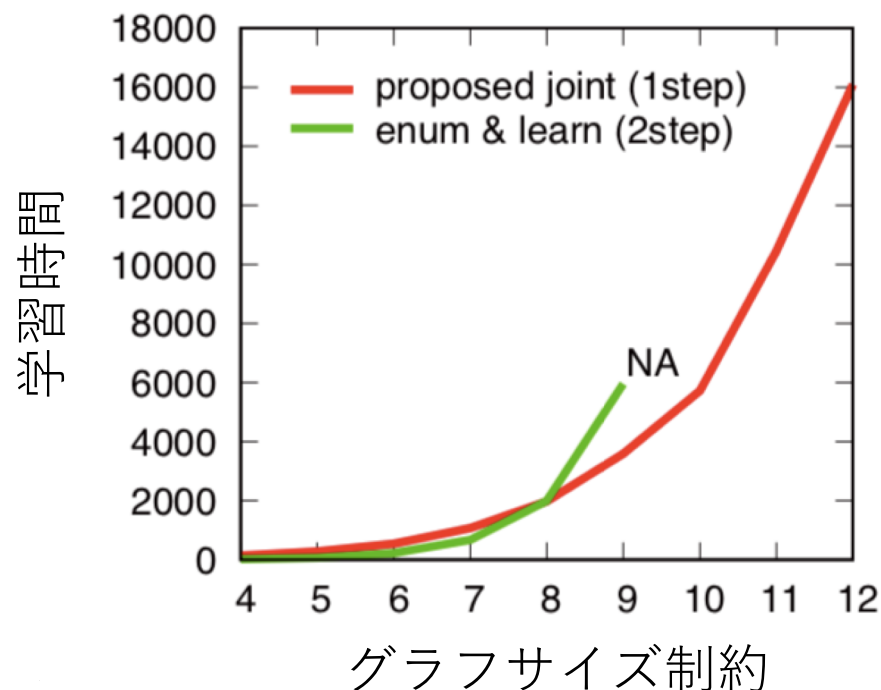
下限値を利用した探索の枝刈り
(現在の最良値 < 子孫での下限値)

結果

緑：既存 赤：提案



$\frac{1}{10}$ 以上の探索コスト削減



スケール不可能な問題を解決

発表：

The 14th International Conference on Mining and Learning with Graphs (MLG 2018), KDD'18 Workshop

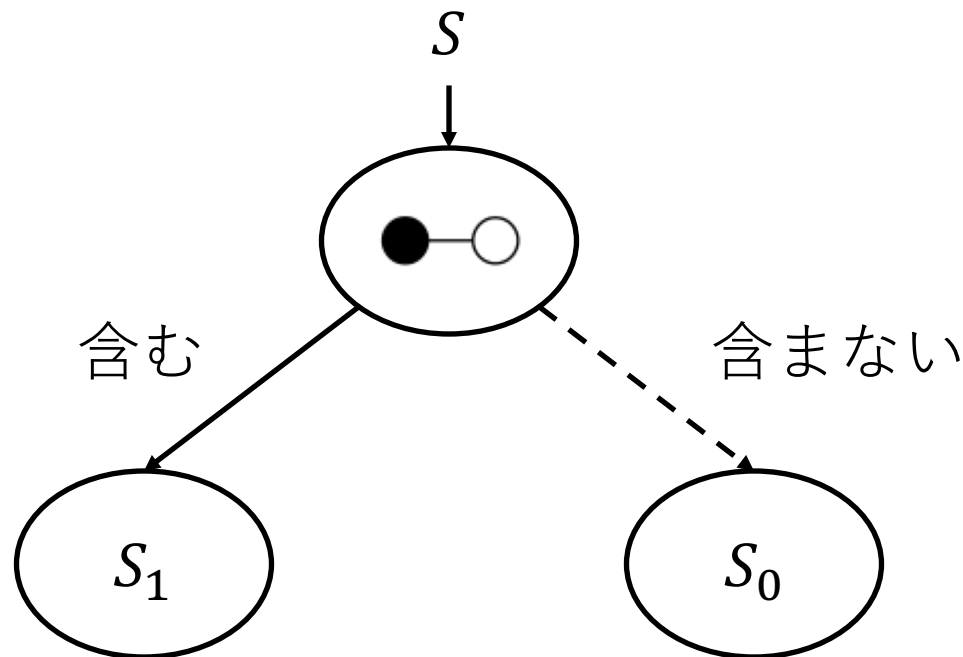
まとめ

- 適応的な特徴探索に基づく回帰木学習アルゴリズムを提案
- 既存手法と比べ大幅な時間・空間コストの削減を達成

取り組みたいテーマ

- 自然言語処理、対話システム
- 機械学習、データマイニング、データ構造

不純度（二乗誤差和）



$$\text{不純度} = \sum_i^{S_0} (y_i - \bar{S}_0)^2 + \sum_i^{S_1} (y_i - \bar{S}_1)^2$$

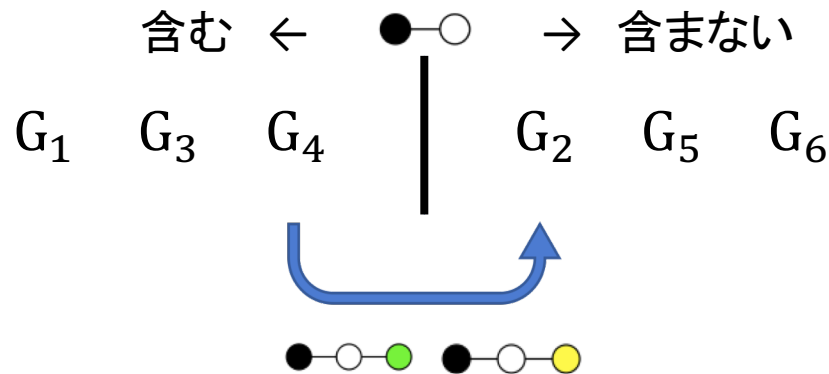
y_i : ラベル, \bar{S}_0, \bar{S}_1 : ラベル平均

下限値の計算

探索木の特徴：子孫 (x') は親 (x) の拡大グラフ

$$G_i \not\supseteq x \Rightarrow G_i \not\supseteq x', x' \supseteq x$$

含むグラフが含まない側に移る方向性しかない



任意のグラフの組み合わせを含まない側へ移したときの
不純度を全て計算すれば下限値が求まる

下限値の計算

組み合わせの数は膨大

不純度が二乗誤差和であることを利用する



探索中の部分グラフパターンを含むグラフ集合において
ラベルの値で昇順、降順ソートし一つずつ移動



線形時間で下限値の計算が可能