

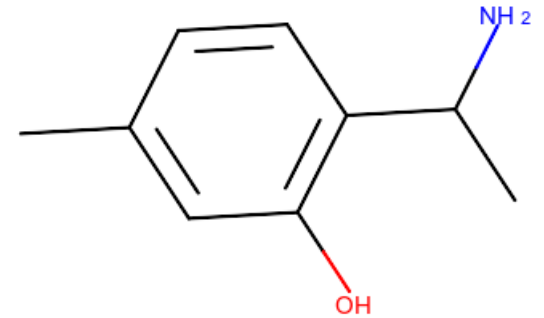
Subgraph-Feature Search for Learning Classifiers and Regressors under Fixed Budget Constraint

情報認識学研究室 白川 稜

背景

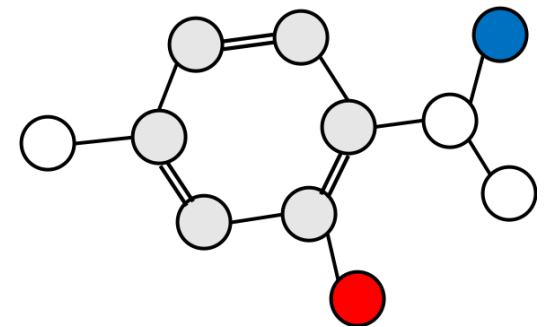
グラフは広く用いられる重要なデータ構造

- 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



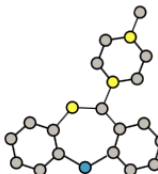
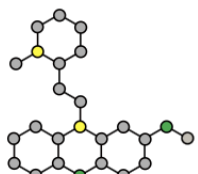
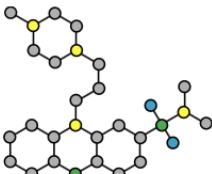
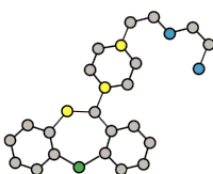
グラフデータからの教師付き学習

- 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

グラフ分類・回帰問題

特徴量

部分グラフの有無

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	...
\vdots	\vdots									...

問題点

グラフサイズに対して部分グラフの総数は組合せ爆発

既存研究

- 2-step approach [Wale et al., 2007]
- Simultaneous approach [Saigo et al., 2009][Shirakawa et al., 2018]

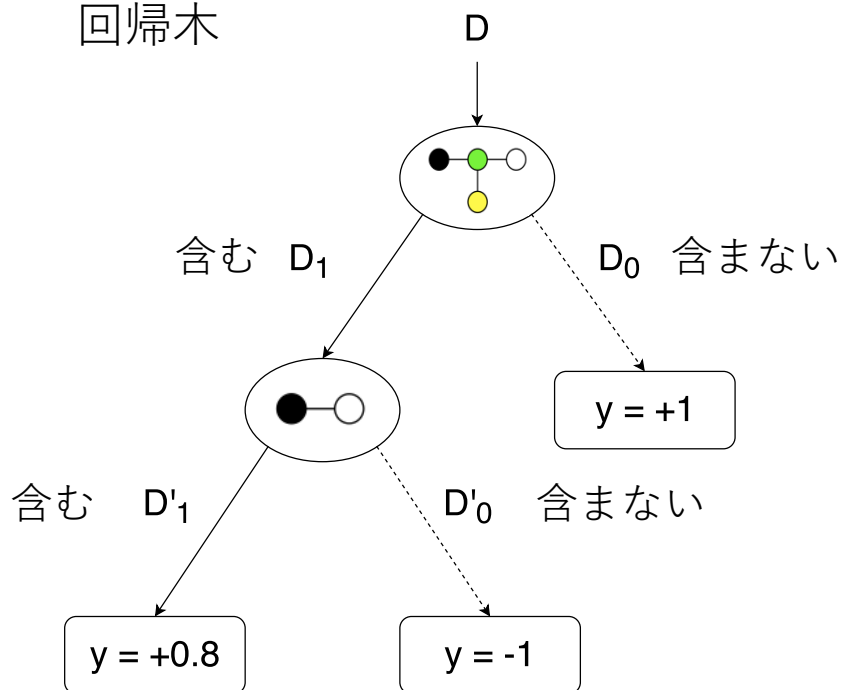
Simultaneous approach

モデルの学習と部分グラフ探索・選択を同時に行う

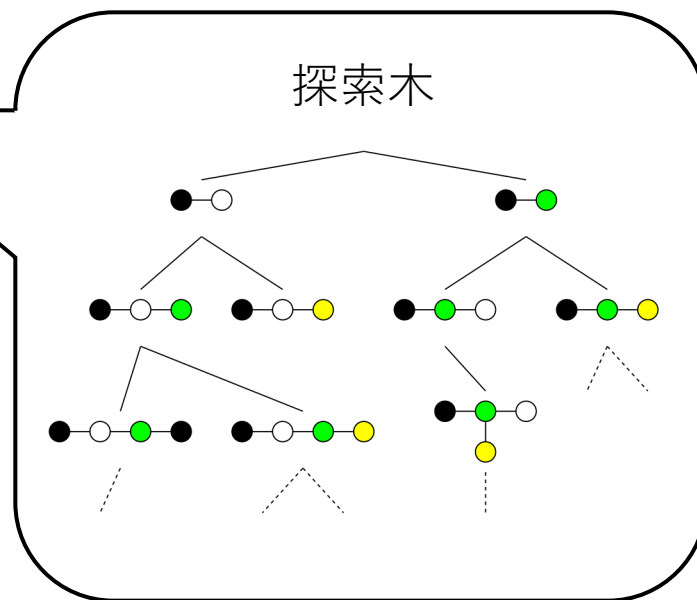
モデル (Gradient Tree Boosting)

特徴探索

回帰木



探索木



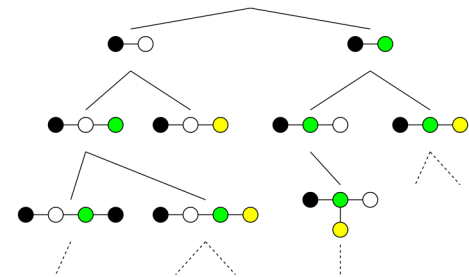
従来手法

分割後の不純度（二乗誤差和：TSS）が最小になる部分グラフ特徴（ g ）を深さ優先的に探索

$$\operatorname{argmin}_g \operatorname{TSS}(D_0(g)) + \operatorname{TSS}(D_1(g))$$

$$D_0(g): \{(G_i, y_i) \in D \mid G \not\supseteq g\}$$

$$D_1(g): \{(G_i, y_i) \in D \mid G \supseteq g\}$$



探索木の包含関係を利用すると子孫ノードでの分割における不純度の下限值（Bound）が計算可能

➡ Boundの値を用いた枝刈りが可能

従来手法

枝刈りを利用した深さ優先に基づく厳密な特徴探索

問題点

- ・ 問題のスケールによって枝刈りだけでは不十分
➡ 特徴探索にかかるコストが大きい

改善点

- ・ 厳密探索 ➡ 事前に探索コストを設定
- ・ 深さ優先探索 ➡ 不純度およびBoundの値を利用した効率的な探索方針の考案

提案手法

目的

低コストで高精度なモデルの構築

➡ 制限されたコスト内において
より良い特徴を探索する探索方針の考案

手法

- モンテカルロ木探索（MCTS）を応用
- 各特徴の不純度の値を報酬に設定し、
事前の探索知識を利用した効率的な探索方針の提案

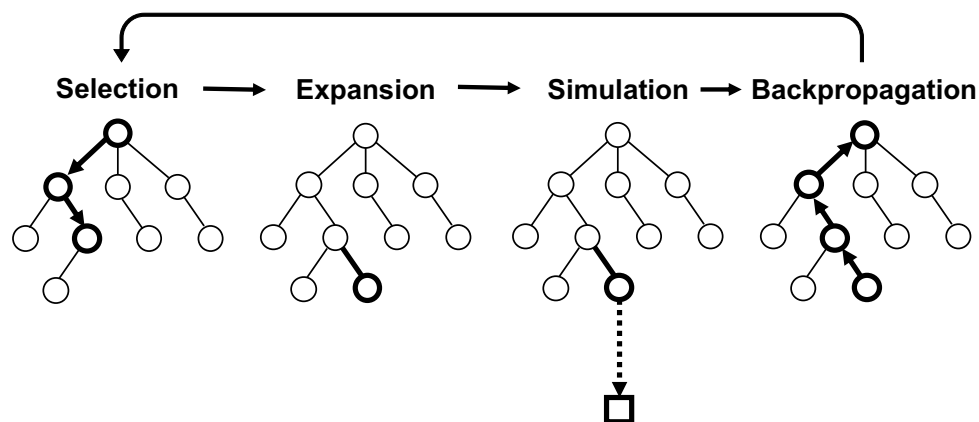
提案手法

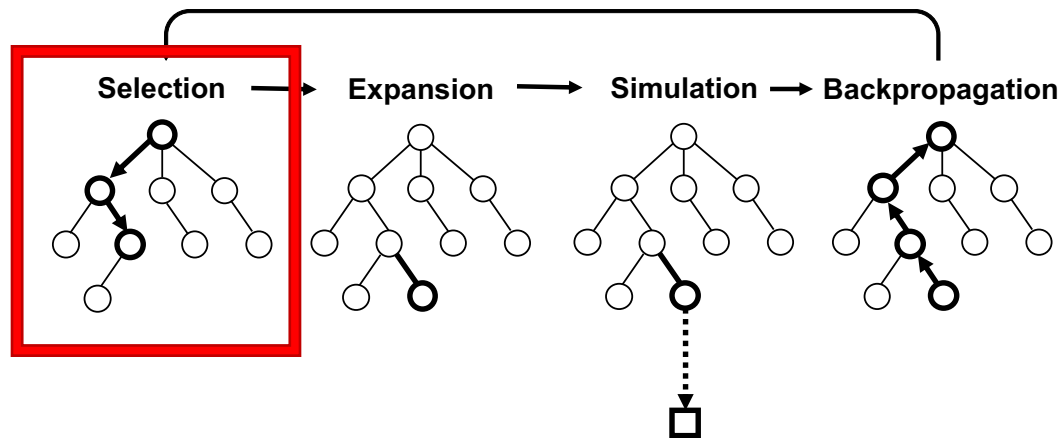
モンテカルロ木探索（MCTS）の一つである
UCTアルゴリズム[Levente et al., 2006]をグラフ探索に適用
UCB（Upper Confidence Bound）の値をもとに探索

手法

以下の操作を反復

1. Selection
2. Expansion
3. Simulation
4. Backpropagation



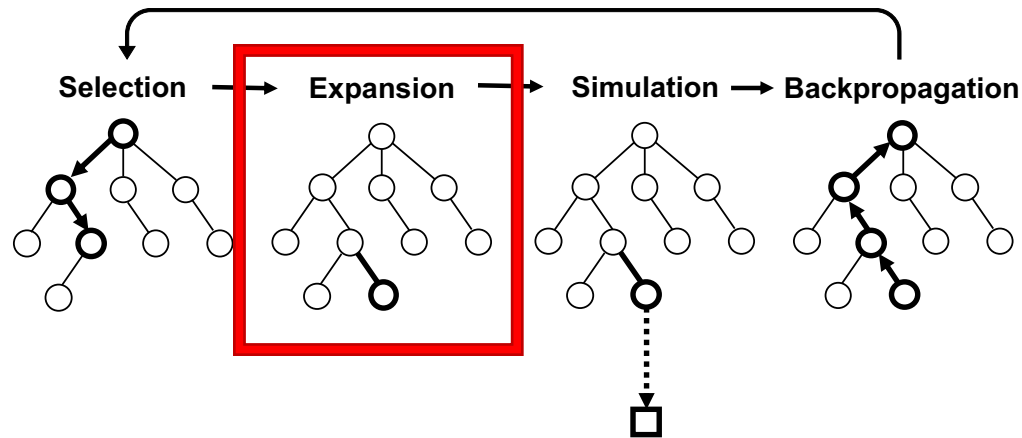


- Selection

根ノードを始点にUCBの値に基づき
探索済みノードの末端までノードを選択する

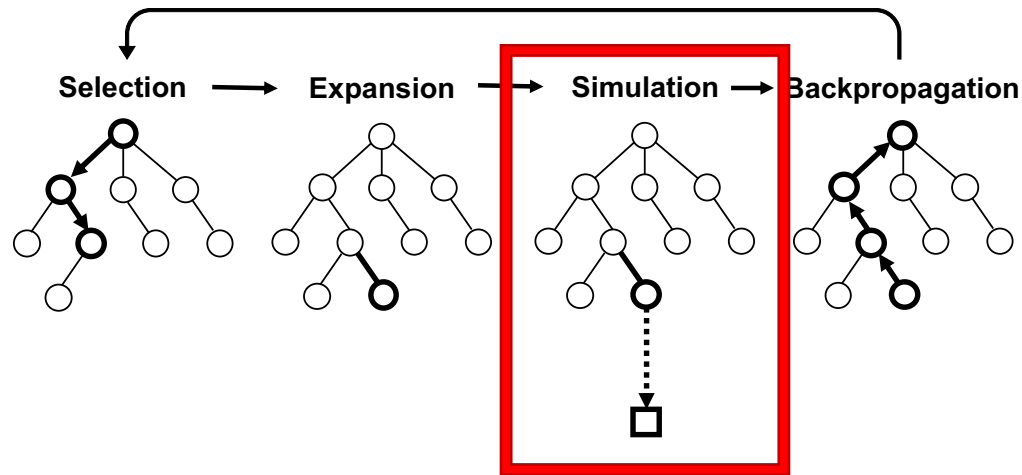
$$UCB = \bar{X}_i + C \times \sqrt{\frac{\ln n}{2n_i}}$$

i: 子ノード番号, \bar{X}_i : 報酬平均, C: 探索強度パラメータ,
n: 親ノード選択回数, n_i : 子ノードi選択回数



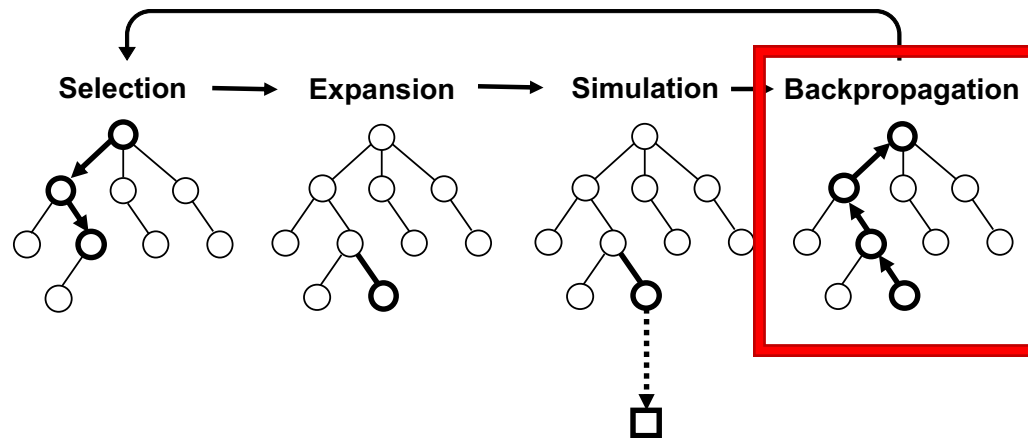
- Expansion

Selectionによって選択された末端ノードからランダムに子ノードを一つ展開し探索済み集合に追加する



- Simulation

Expansionによって展開されたノードから
モンテカルロシミュレーションによりパスを降下



- Backpropagation

Simulationによって選択されたノードの報酬を計算

$$\text{報酬} = - \frac{\text{TSS}(D_0(g)) + \text{TSS}(D_1(g))}{\text{TSS}(D_0(g) \cup D_1(g))}$$

報酬をSelectionで選択したパスに逆伝搬

実験準備

実データセット

Dataset	CPDB	Mutag	AIDS(CA _{vs} CM)	CAS
# data	684	188	1503	4337
# ($y = +1, -1$)	(341, 343)	(125, 63)	(422, 1081)	(2401, 1936)
# nodes	25.2	26.3	59.0	30.3
# edges	25.6	28.1	61.6	31.3
# of nodes and edges are average.				

人工データセット

実データセット (CAS) より100個のグラフを
ランダムサンプリング → ランダムラベル(y)付与
上記のデータセットをそれぞれ100セット作成

artificial1 : $y \in (+1, -1)$

artificial2 : $y \in [+1, -1]$

実験 1

2つの人工データセットに対してコスト制約無しに特徴探索を1回行う

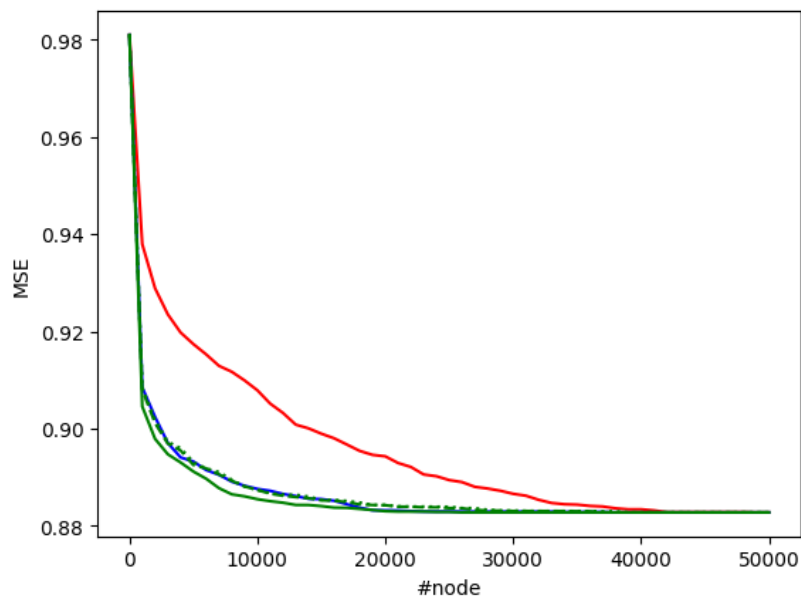
特徴探索における暫定解更新の様子を各手法で比較する

- 深さ優先探索（従来手法1）
- 最良優先探索（従来手法2）
- モンテカルロ木探索（提案手法）

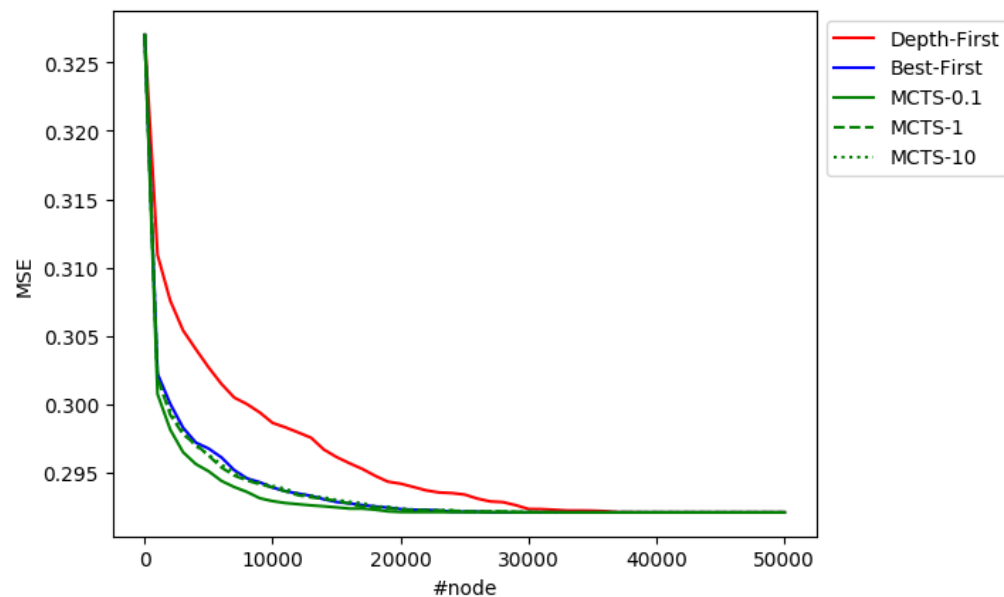
※モンテカルロ木探索の探索強度パラメータ（0.1, 1, 10）

実験 1

artificial1, $y \in \{+1, -1\}$



artificial2, $y \in \{+1, -1\}$



- 提案手法がより早くに良い特徴を発見
- 探索の後半はあまり重要ではない

実験 2

実データセットに対してコスト制約を設けた上で
アンサンブルモデルの学習を行う

学習パラメータ

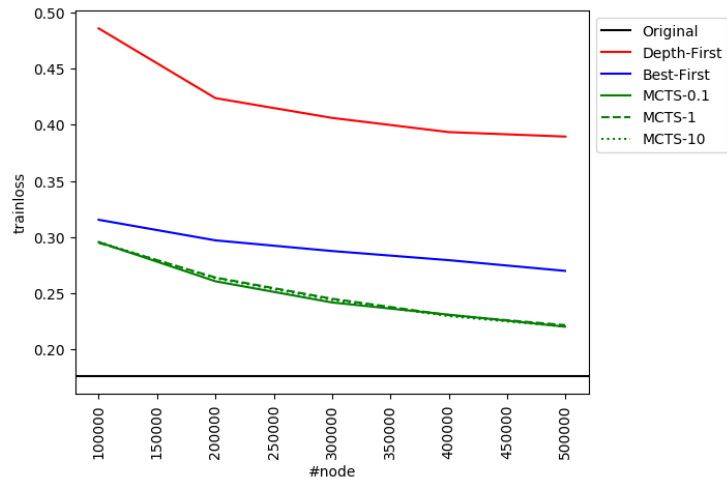
- 木の本数：100
- 木の深さ：1
- ステップ幅：1

コスト制約(一回の特徴探索にかけるノード数)

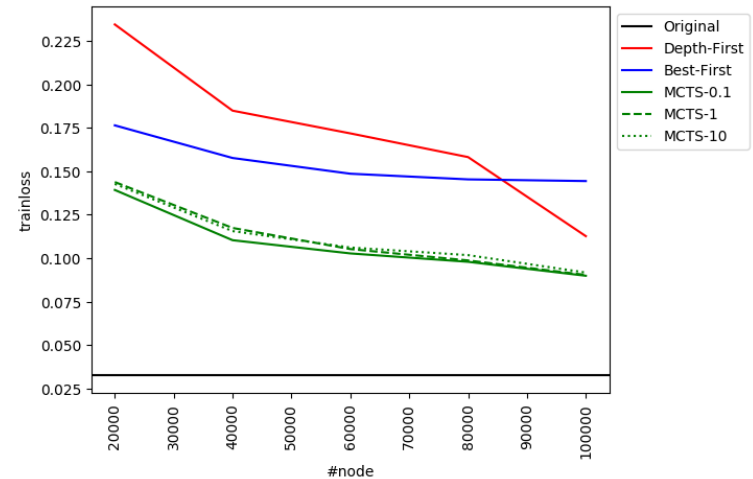
- CPDB： (1000, 2000, 3000, 4000, 5000)
- Mutag： (100, 200, 300, 400, 500)
- AIDS： (1000, 2000, 3000, 4000, 5000)
- CAS： (5000, 10000, 15000, 20000, 25000)

実験 2 (Training Loss)

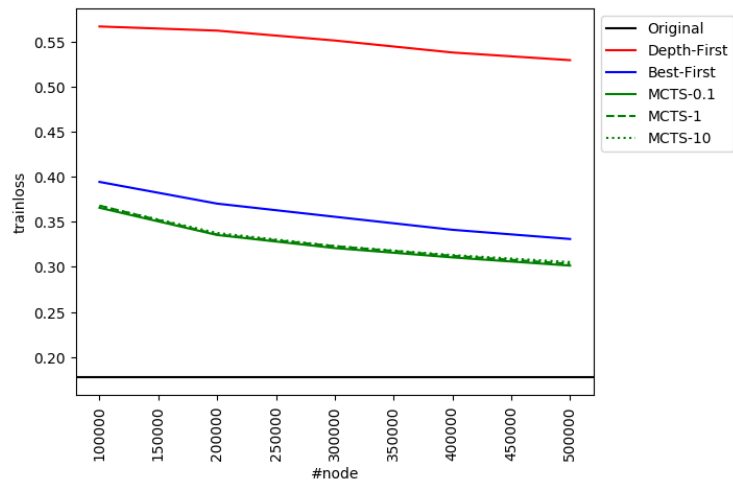
CPDB



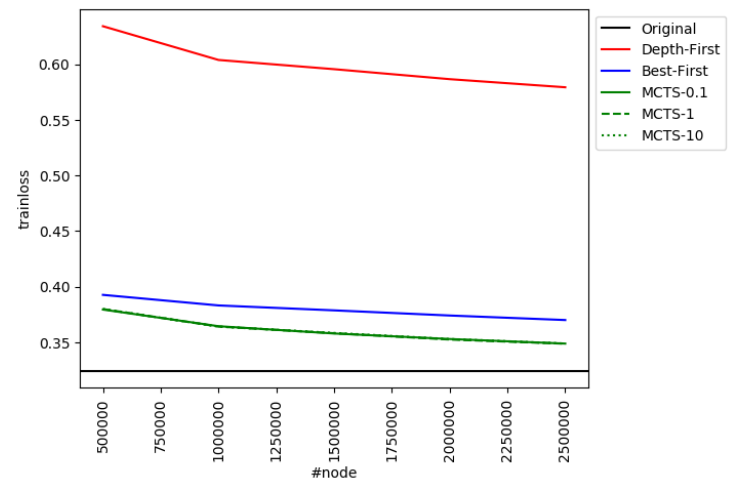
Mutag



AIDS

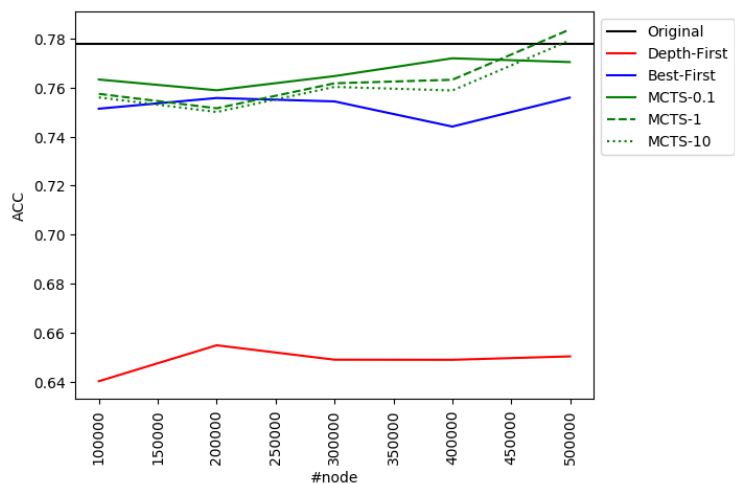


CAS

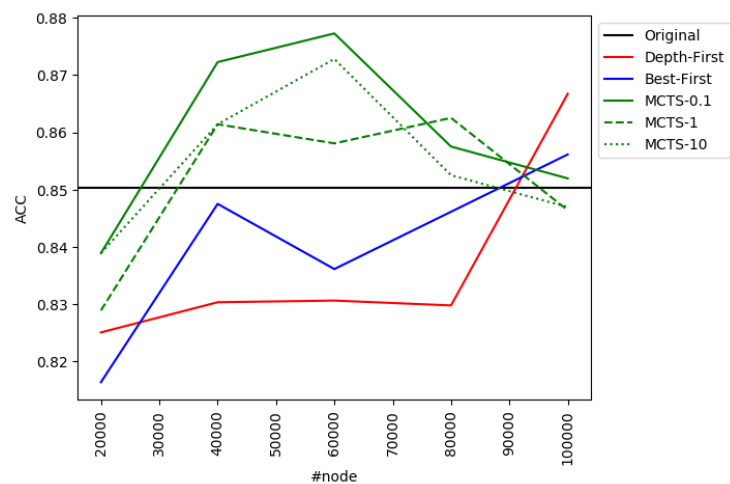


実験 2 (ACC)

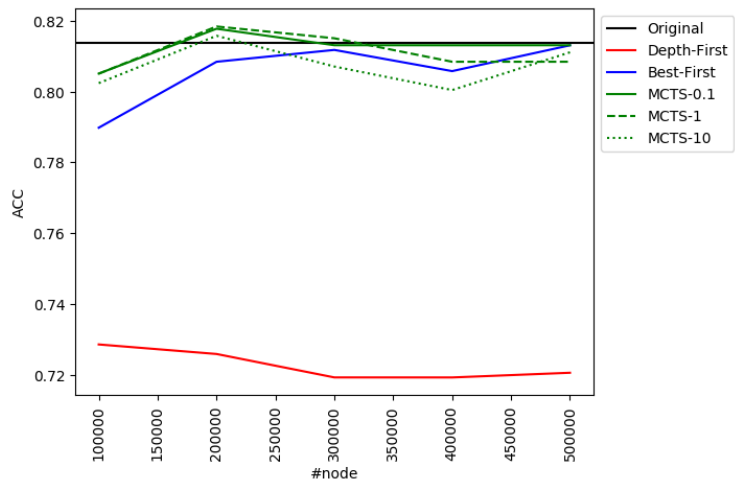
CPDB



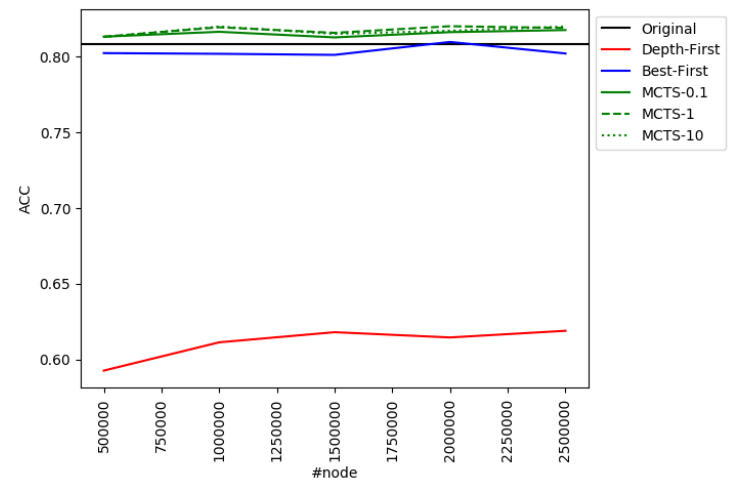
Mutag



AIDS



CAS



実験 2（厳密探索との比較）

データ	探索ノード数			実行時間[s]			精度[%]		
	従来	提案	比	従来	提案	比	従来	提案	差
CPDB	7.2×10^6	5.0×10^5	0.070	8.2×10^2	6.2×10	0.076	77.78	78.35	+0.57
Mutag	3.8×10^5	6.0×10^4	0.015	2.3×10^2	3.7	0.016	85.03	87.73	+2.70
AIDS	7.9×10^7	2.0×10^5	0.003	2.5×10^4	1.1×10^2	0.004	81.37	81.84	+0.47
CAS	6.9×10^7	2.0×10^6	0.029	8.0×10^4	1.7×10^3	0.040	80.82	81.99	+1.17

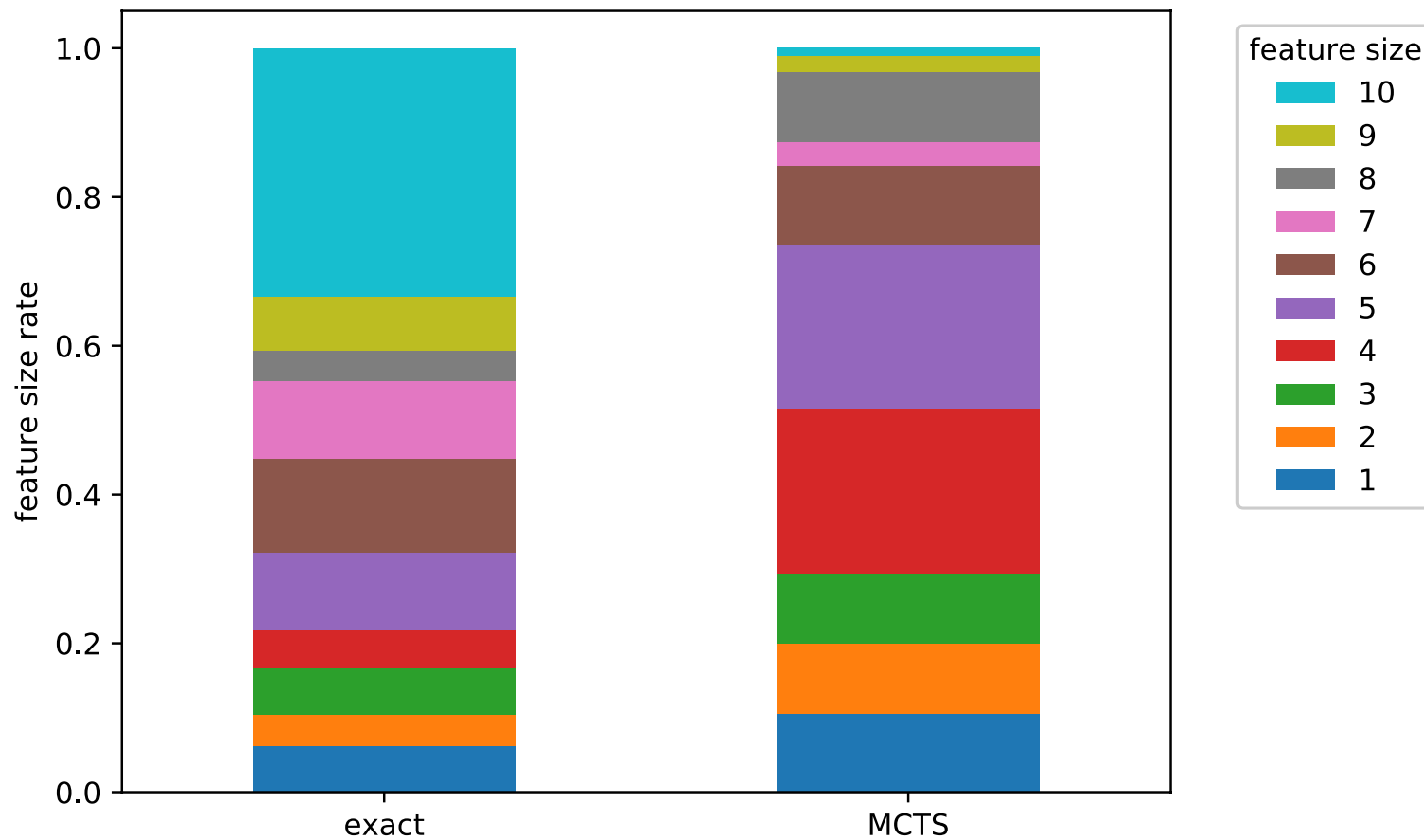
- 精度の低下なしに省コストを達成

まとめ

- ・ 既存のグラフ分類・回帰問題の学習アルゴリズムの探索にモンテカルロ木探索を利用した手法を提案
- ・ 従来手法である深さ優先探索、最良優先探索に対してより少ない探索コストでより良い解を発見
- ・ 特に従来の深さ優先厳密探索の場合と比較すると約 $1/10 \sim 1/100$ の探索コストで同等、それ以上の精度のモデルを構築

質疑

汎化性能

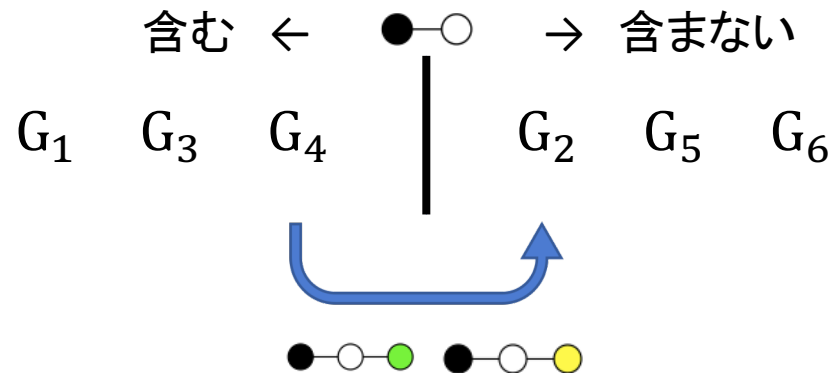


下限値の計算

探索木の特徴：子孫 (g') は親 (g) の拡大グラフ

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

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

$$\begin{aligned} TSS(D_0(g')) + TSS(D_1(g)) \\ \geq \min_{(\circ, k)} [TSS(D_0(g) \setminus S_{(\circ, k)}) + TSS(D_1(g) \cup S_{(\circ, k)})] \end{aligned}$$

$(\circ, k) \in \{\leq, >\} \times \{2, \dots, |D_1(g) - 1|\}$

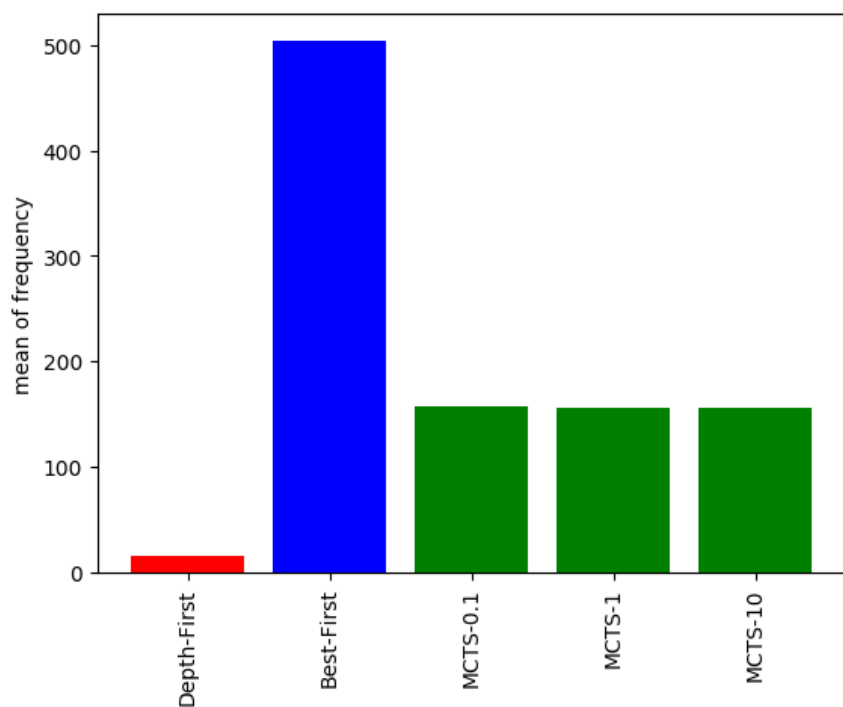
$S_{(\leq, k)}$ is a set of k pair (G_i, y_i) selected from $D_1(g)$ in descending order of y

$S_{(>, k)}$ is a set of k pair (G_i, y_i) selected from $D_1(g)$ in increasing order of y

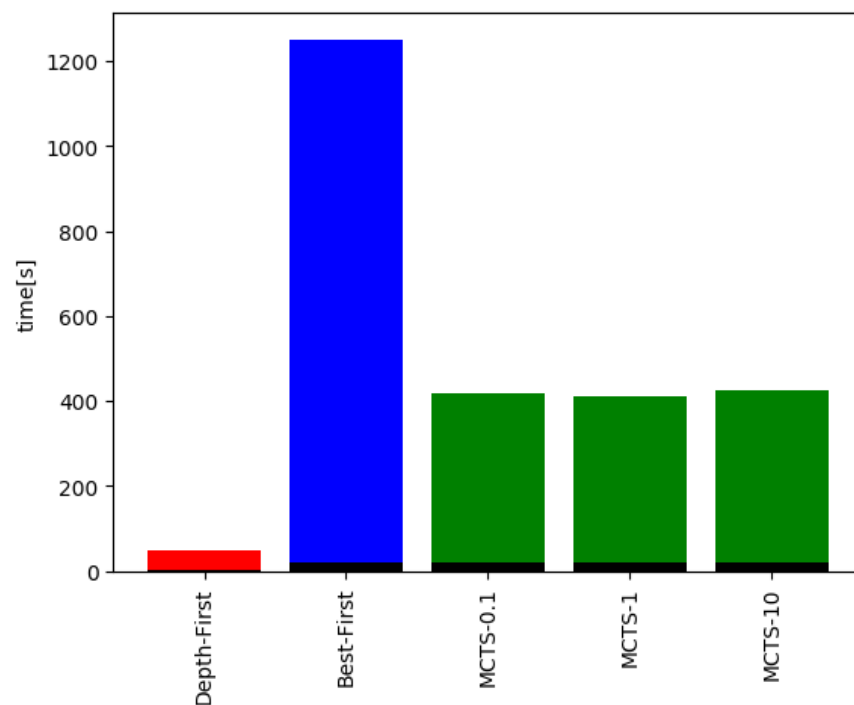
計算量：グラフ (g) の頻出度に対して線形オーダー

探索速度

頻出度累計



実行時間[s]



深さ優先は頻出度の低いノードを多く探索
最良優先は頻出度の高いノードを多く探索