

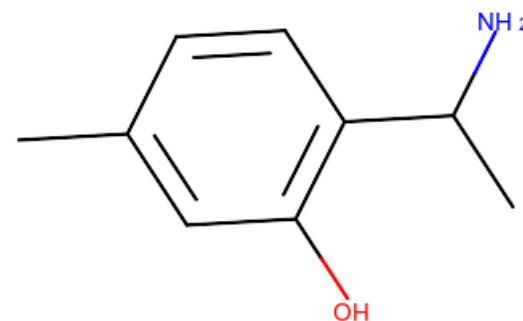
Subgraph-Feature Search for Learning Classifiers and Regressors under Fixed Budget Constraint

情報認識学研究室 白川 稜

背景

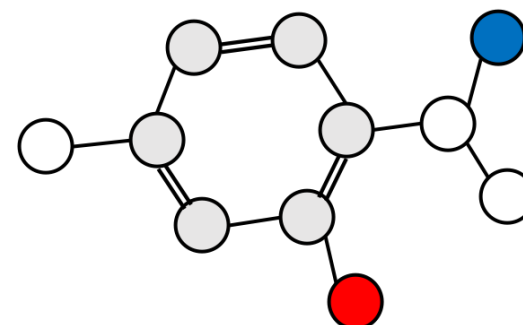
グラフは広く用いられる重要なデータ構造

- 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




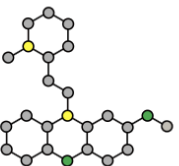
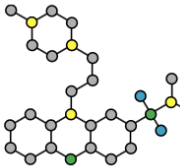
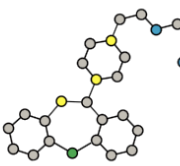
グラフデータからの教師付き学習

- 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

活性の有無、物性値 etc...

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]
 - ー 制約付き部分グラフ列挙 + 任意モデルの学習
- Simultaneous approach [Saigo et al., 2009][Shirakawa et al., 2018]
 - ー モデルの学習と部分グラフ探索・選択の同時手法

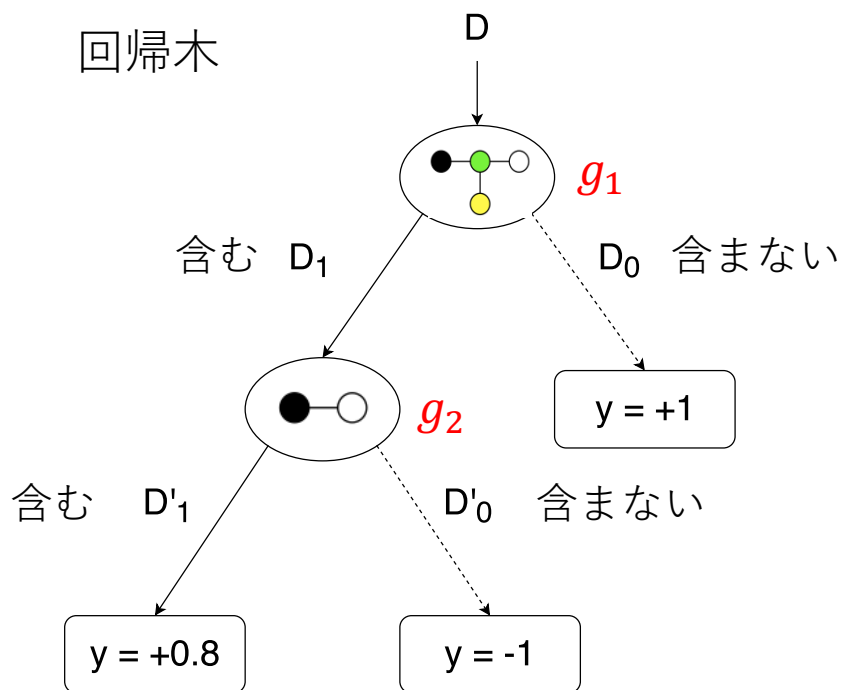
GTB [Shirakawa et al., 2018]

モデルの学習と部分グラフ探索・選択を同時に行う

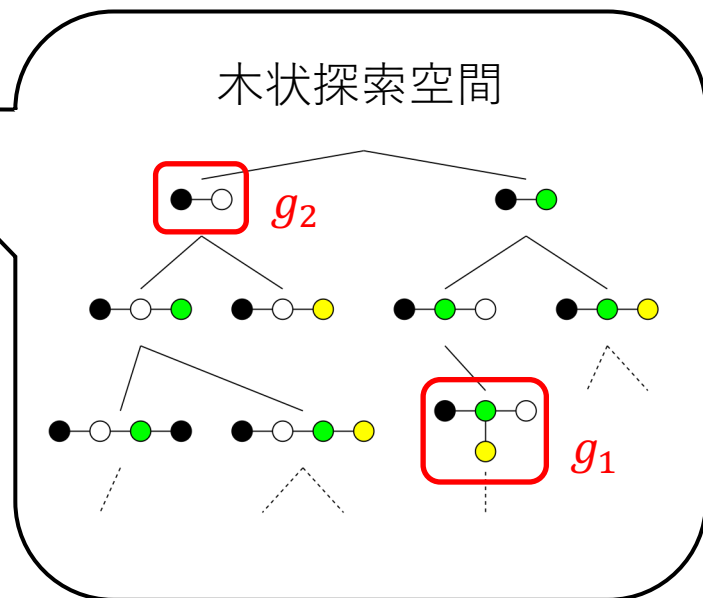
モデル (Gradient Tree Boosting)

特徴探索 (本研究)

回帰木



木状探索空間



従来手法 (exact Depth-first)

分割後の二乗誤差和 (TSS) が最小になる
部分グラフ特徴 (g) を深さ優先的に探索

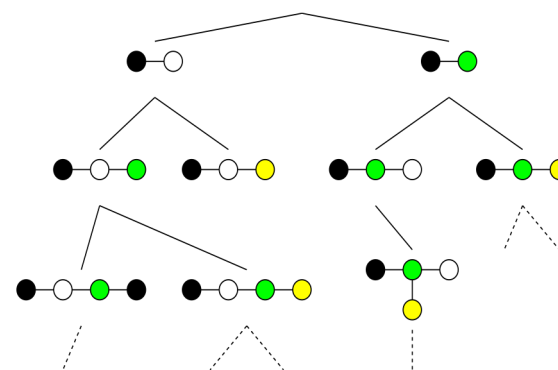
$$\min_g [\text{TSS}(D_0(g)) + \text{TSS}(D_1(g))]$$

$$\text{TSS}(D) = \sum_{i \in [|D|]} (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{|D|} \sum_{i \in [|D|]} y_i$$

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

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

\supseteq : 部分グラフ同型



探索木の包含関係を利用し

子孫ノードでのTSSの下限值 (Bound) を計算

➡ 枝刈りを利用した厳密探索

目的・提案

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

目的

学習モデルの精度の劣化なしに探索コストの削減

提案

- ・ 厳密探索 ➡ 近似探索
- ・ 深さ優先 ➡ TSS, Boundを利用した探索方針

提案手法

近似探索

事前に探索コスト（ノード数）を設定し
コストを使い果たしたら終了

探索方針

- 最良優先探索
- モンテカルロ木探索（MCTS）

提案手法

モンテカルロ木探索（MCTS）の一つである

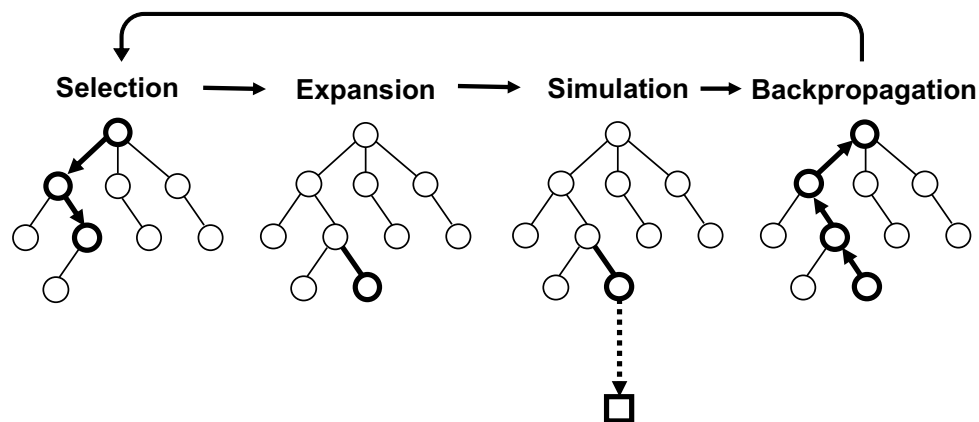
UCTアルゴリズム[Levente et al., 2006] をグラフ探索に適用

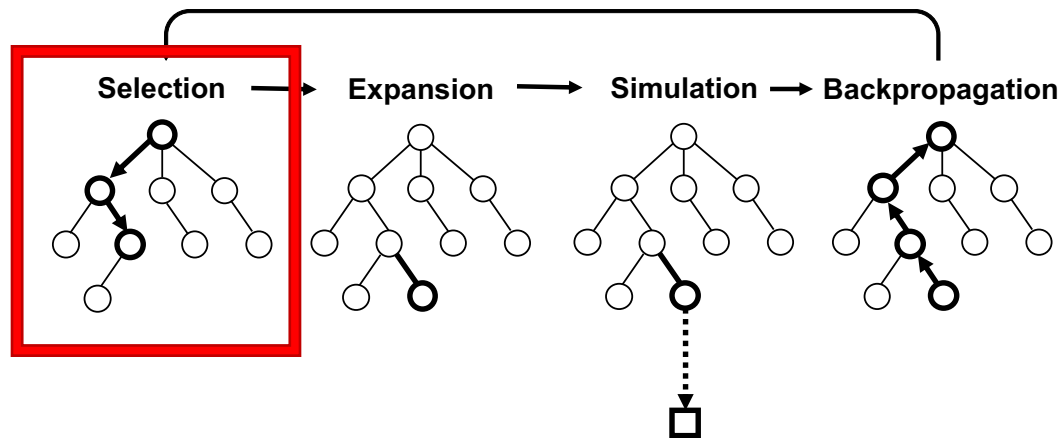
UCB（Upper Confidence Bound）の値をもとに探索

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

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

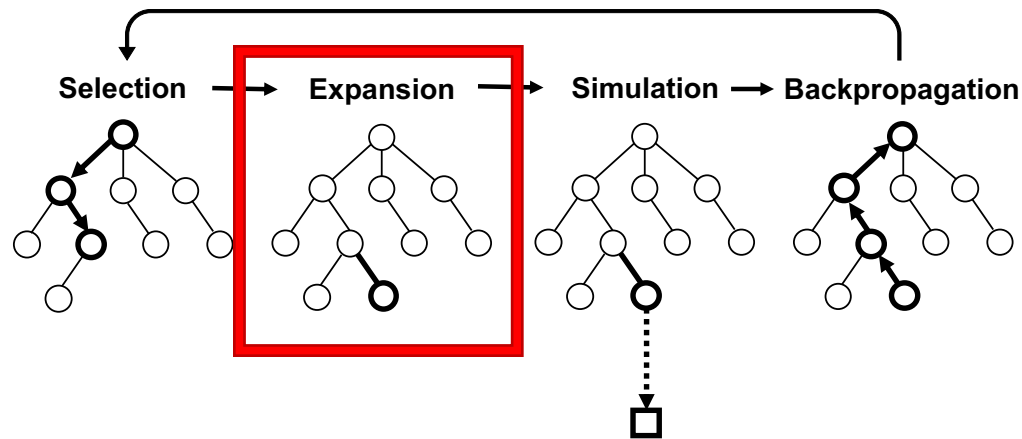
i : 子ノード番号

\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

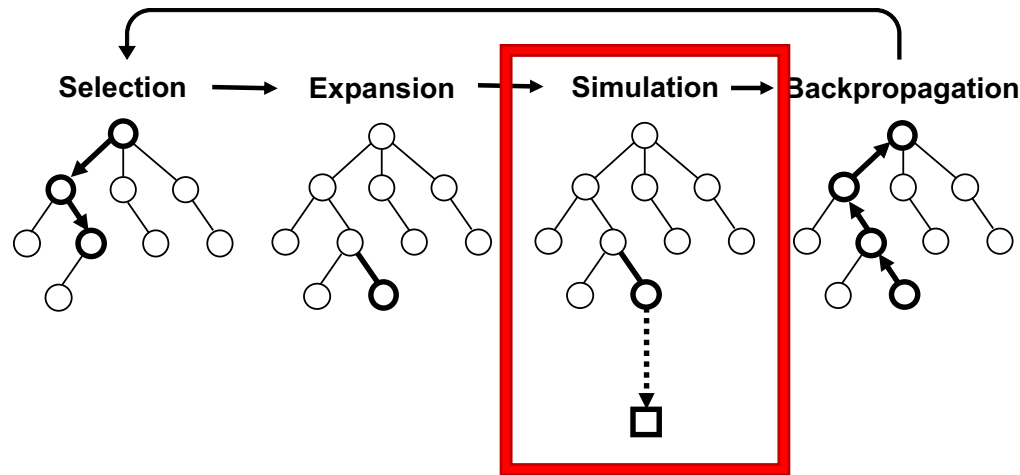
- 末端ノードが**初訪問**

- ➡ 子ノードを列挙

- ➡ ランダムに子ノードを 1 つ拡大

- 末端ノードが**既訪問**

- ➡ ランダムに未拡大の子ノードを 1 つ拡大



- Simulation

モンテカルロシミュレーションによりパスを降下

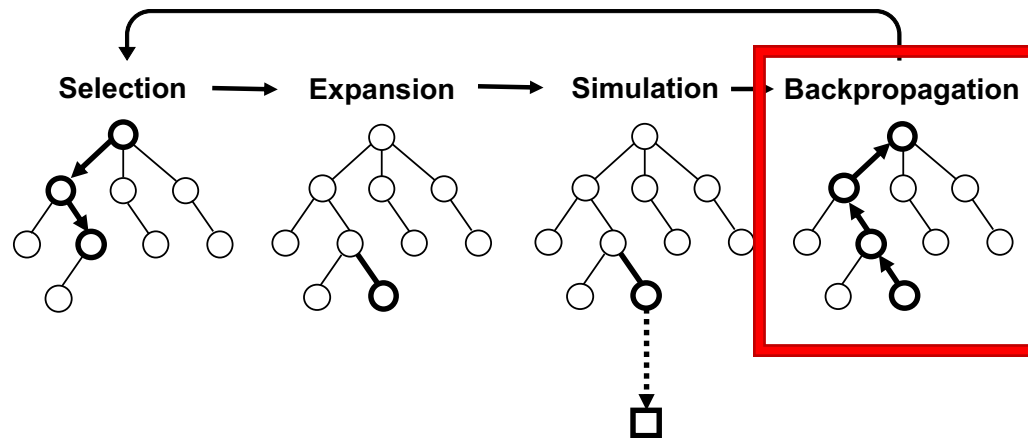
✗ 列挙 + ランダム選択

○ ランダムにグラフを選択

➡ グラフ上でランダムに1エッジ拡大

➡ 拡大グラフが子ノードになる：OK

└ 拡大グラフが子ノードにならない：戻る



- Backpropagation

Simulationによって選択されたノードの報酬を計算

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

TSS : 二乗誤差和

$D_0(g)$: g を含むグラフ集合

$D_1(g)$: g を含まないグラフ集合

報酬をSelectionで選択したパスに逆伝搬

実験準備

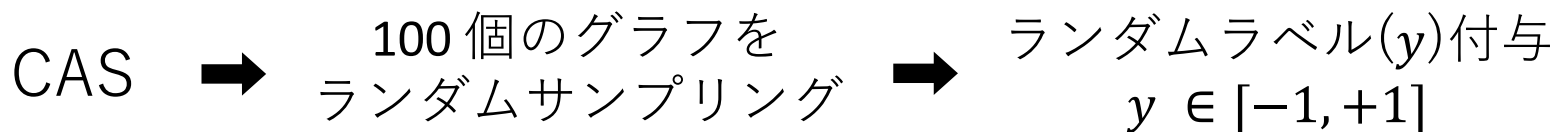
実データセット

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

人工データセット

様々な問題設定を考慮するため

以下の操作で**100個**のデータセットを準備



実験 1

目的

厳密探索での各探索方針の比較

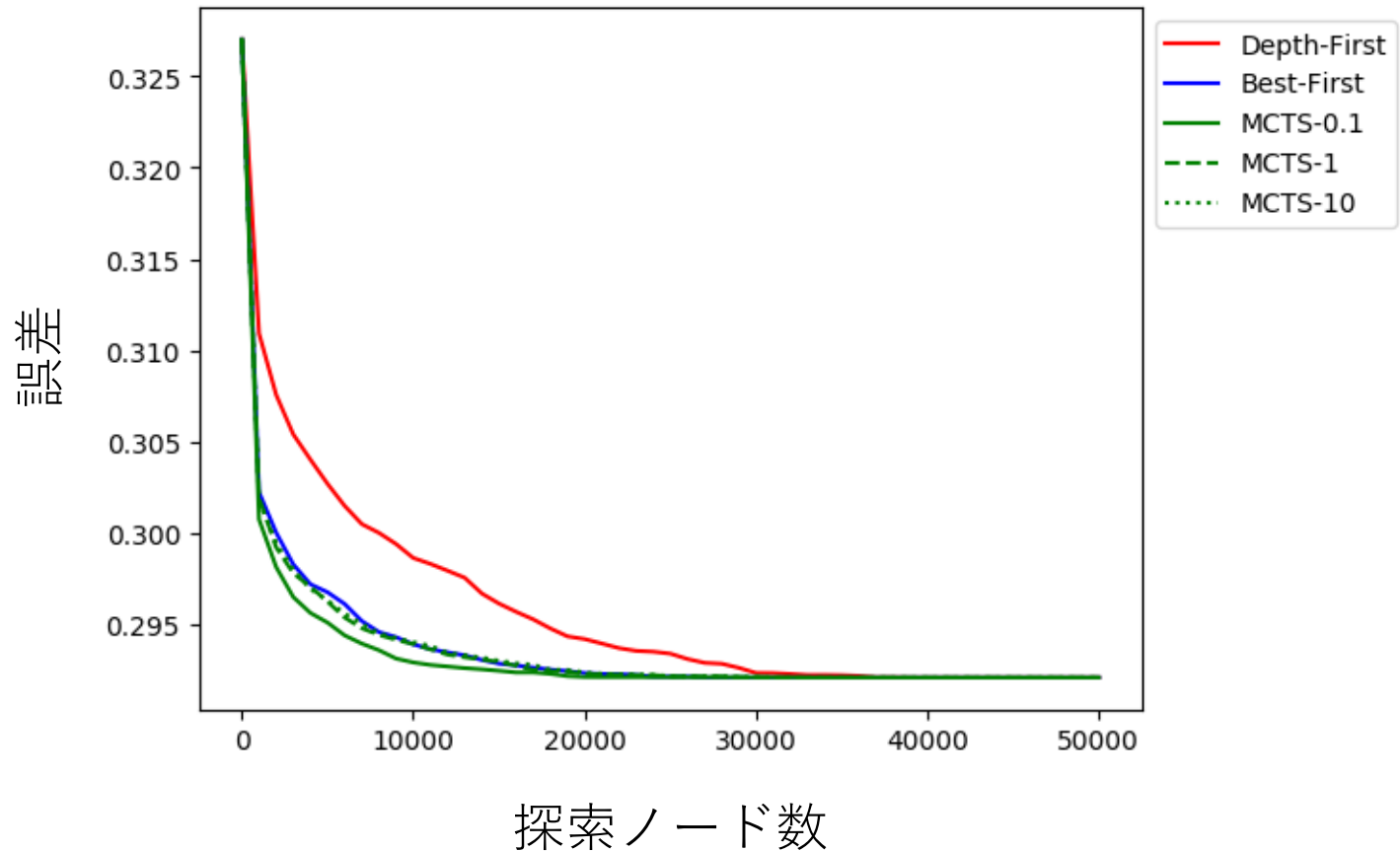
手法

人工データセットに対して厳密特徴探索を 1 回行い
各探索手法での解の更新の様子を比較

- 深さ優先探索
- 最良優先探索
- モンテカルロ木探索

※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



- 提案手法がより早くに良い特徴を発見
- 後半の探索の改善度は低い

実験 2

目的

近似探索での各探索方針の比較

手法

実データセットに対して各探索方針に基づく
近似探索を用いたアンサンブルモデルの学習・比較

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)

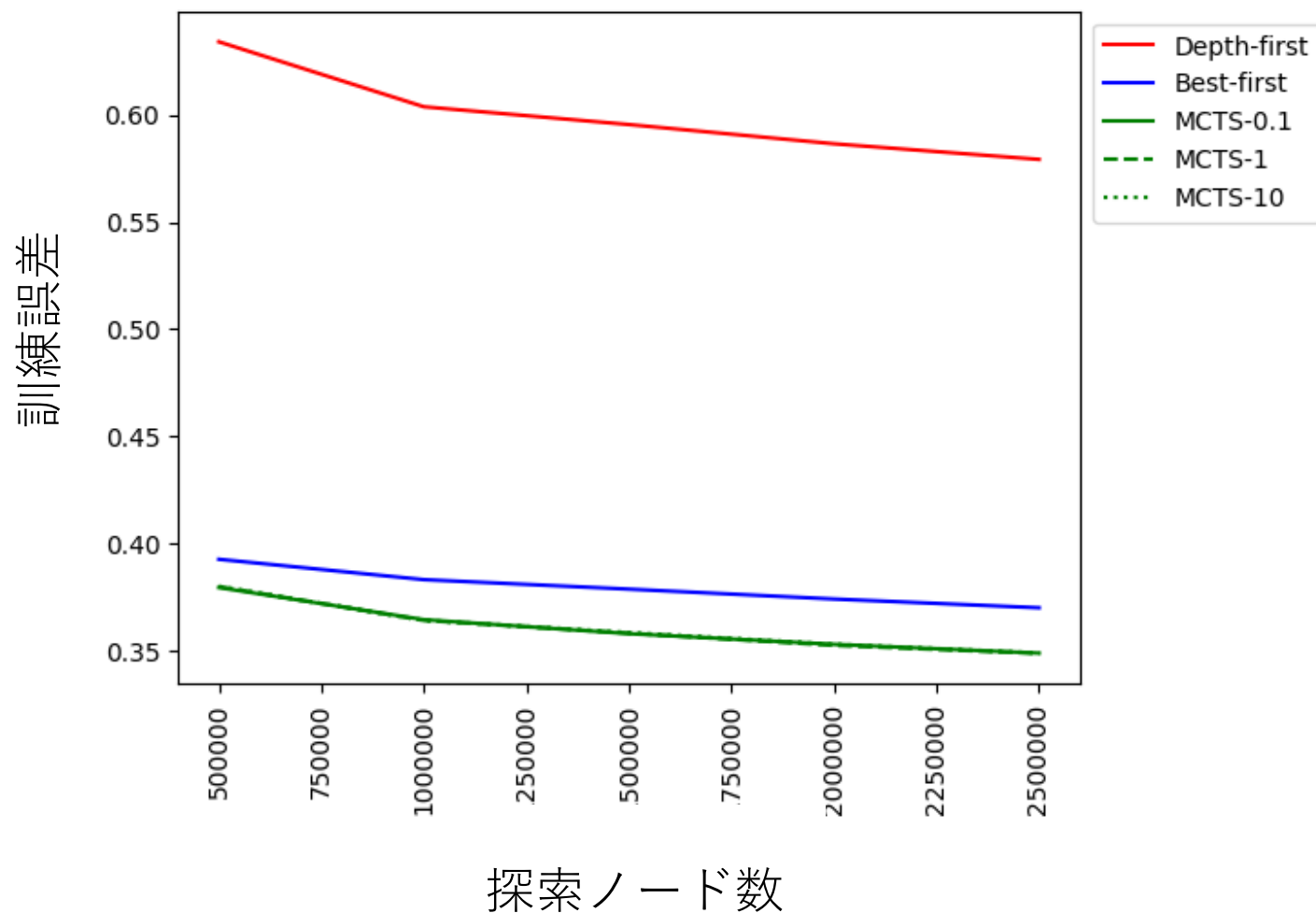
CPDB： (1000, 2000, 3000, 4000, 5000)

Mutag： (200, 400, 600, 800, 1000)

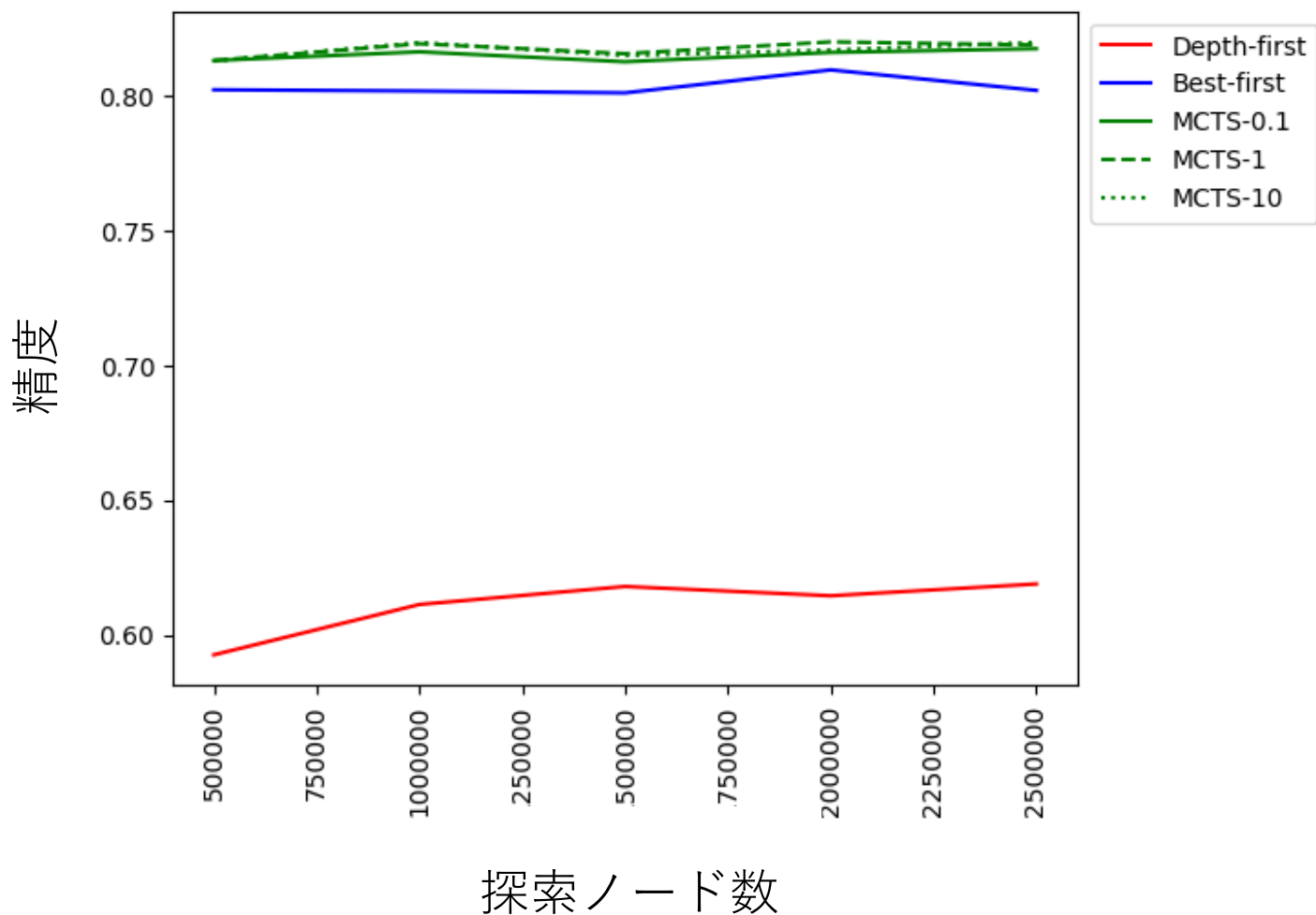
AIDS： (1000, 2000, 3000, 4000, 5000)

CAS： (5000, 10000, 15000, 20000, 25000)

実験 2 (CAS)



実験 2 (CAS)



実験 3

目的

従来厳密探索と提案近似探索の比較

手法

実データセットに対して従来厳密探索と提案近似探索に基づくアンサンブルモデルの学習・比較

- ・ 従来：GTB + 深さ優先厳密探索
- ・ 提案：GTB + MCTS近似探索

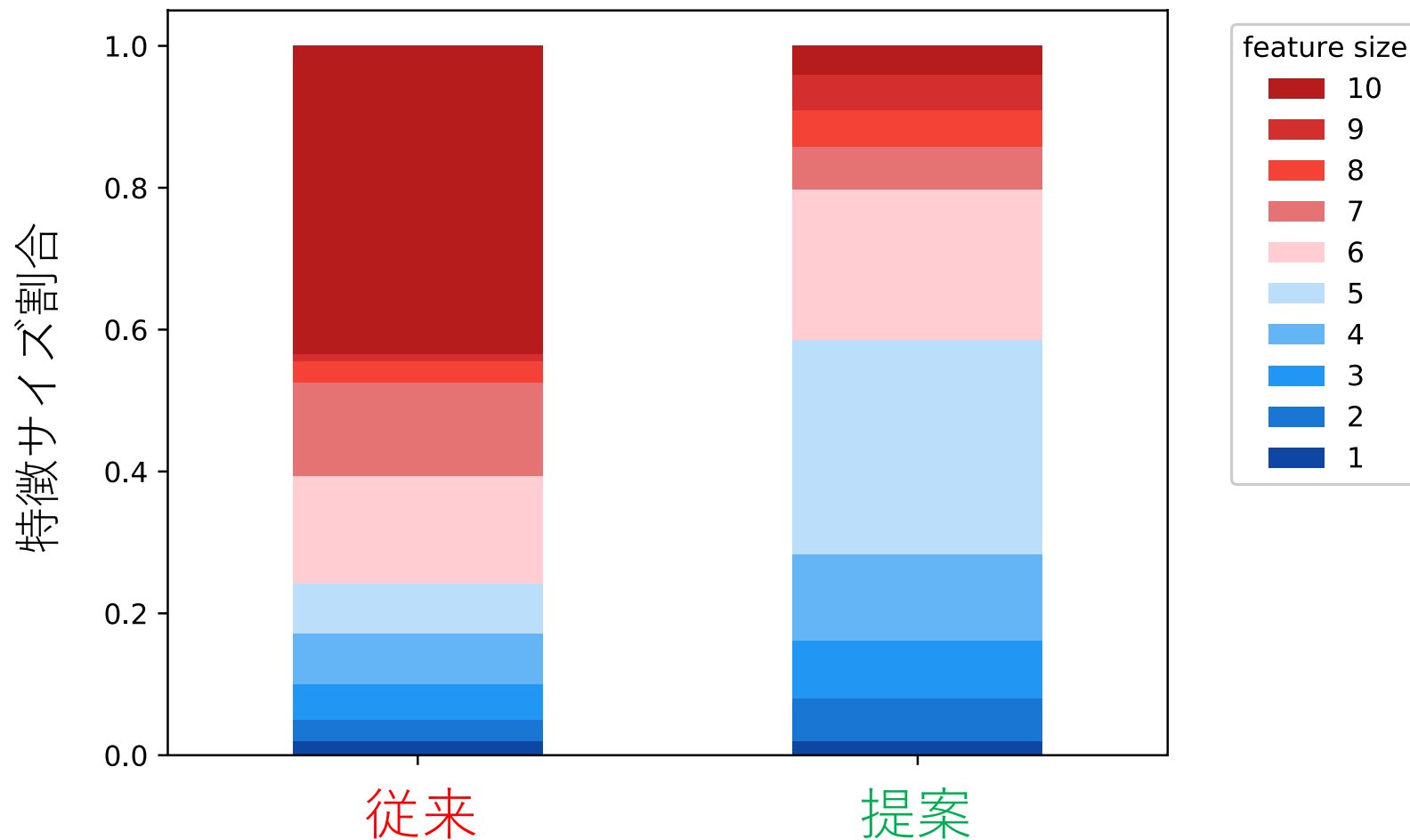
※MCTSの探索強度パラメータ、コスト制約は実験2の最良値

実験 3

データ	探索ノード数		実行時間[s]		精度[%]	
	従来	提案	従来	提案	従来	提案
CPDB	7.2×10^6	5.0×10^5	8.2×10^2	6.2×10	77.78	78.35
Mutag	3.8×10^5	6.0×10^4	2.3×10^2	3.7	85.03	87.73
AIDS	7.9×10^7	2.0×10^5	2.5×10^4	1.1×10^2	81.37	81.84
CAS	6.9×10^7	2.0×10^6	8.0×10^4	1.7×10^3	80.82	81.99

- 精度の劣化なしに省コスト化を達成

汎化性能



- 提案手法の方が汎化性能が高い

まとめ

既存のグラフ分類・回帰アルゴリズムの特徴探索に
最良優先, MCTSを利用した近似探索手法を提案

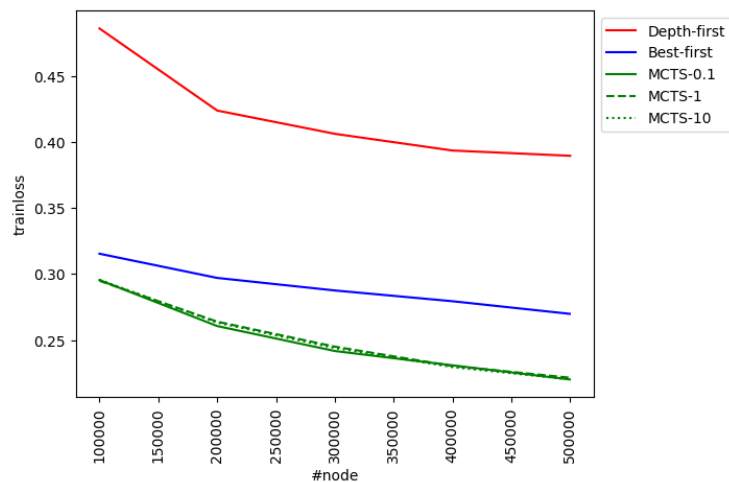
探索方針に関して従来手法である深さ優先方針に比べ
より少ない探索数でより良い解を発見

従来の深さ優先厳密探索モデルと
提案のMCTSを利用した近似探索モデル比較すると
約 $1/10 \sim 1/200$ のコストで
同等、それ以上の精度のモデルを構築

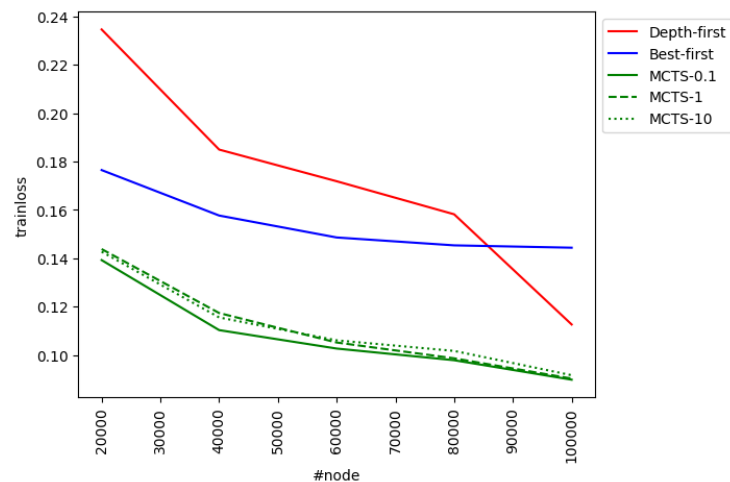
質疑

実験 2 (Training Loss : #node)

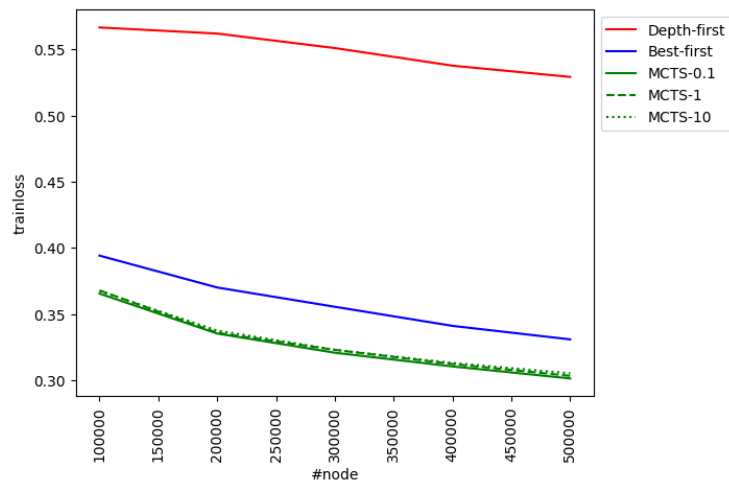
CPDB



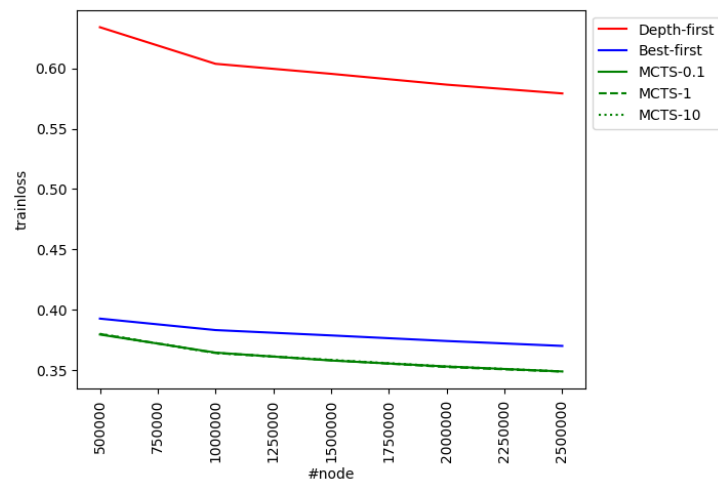
Mutag



AIDS

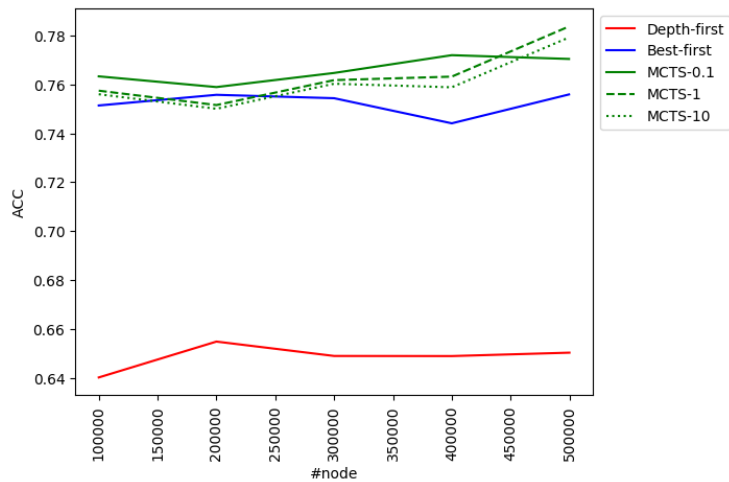


CAS

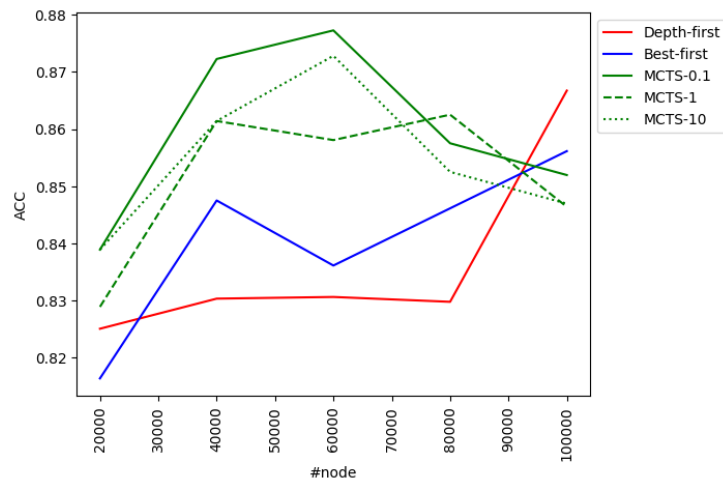


実験 2 (ACC : #node)

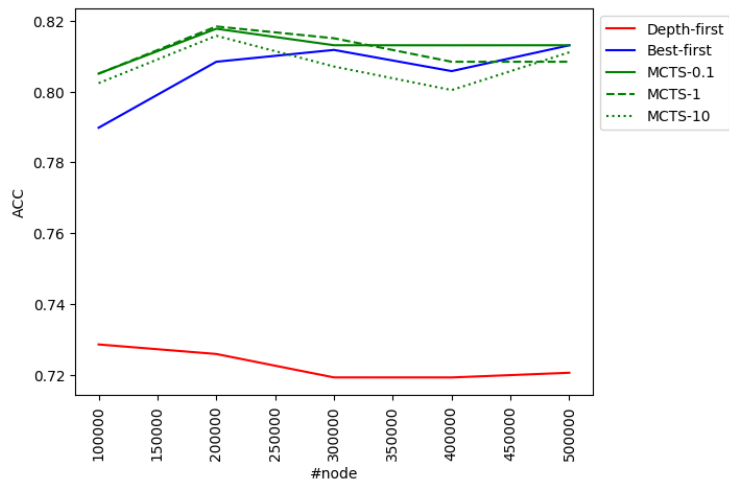
CPDB



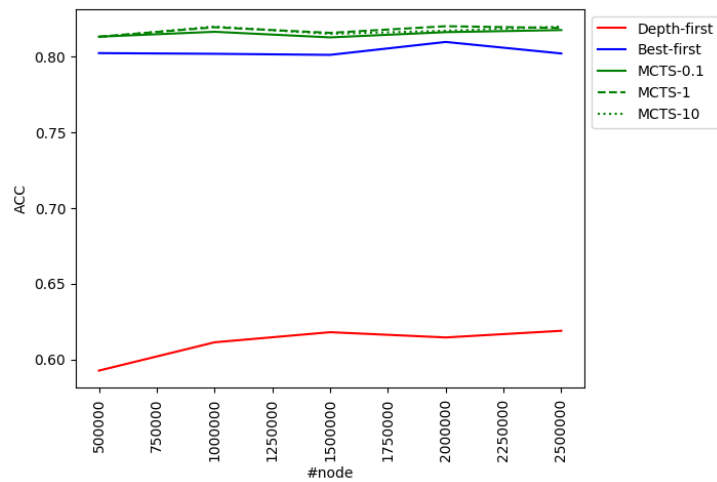
Mutag



AIDS

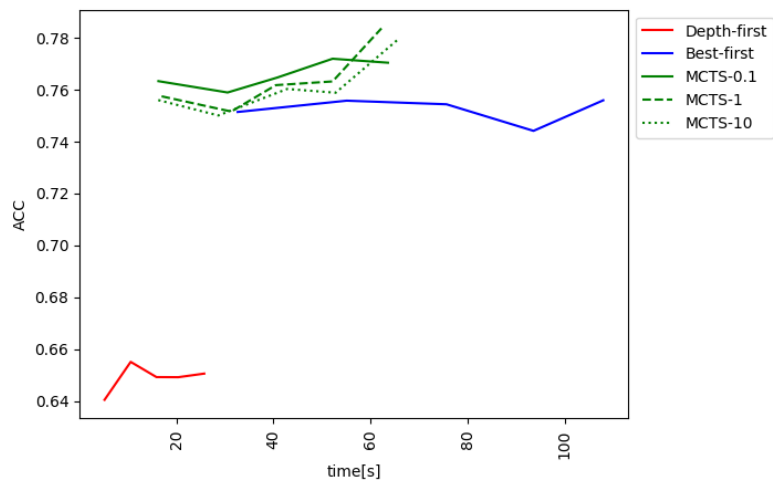


CAS

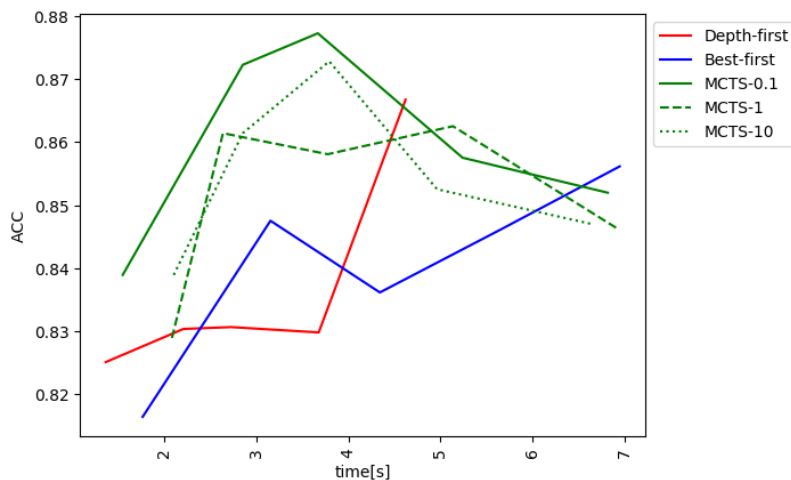


実験 2 (ACC : time[s])

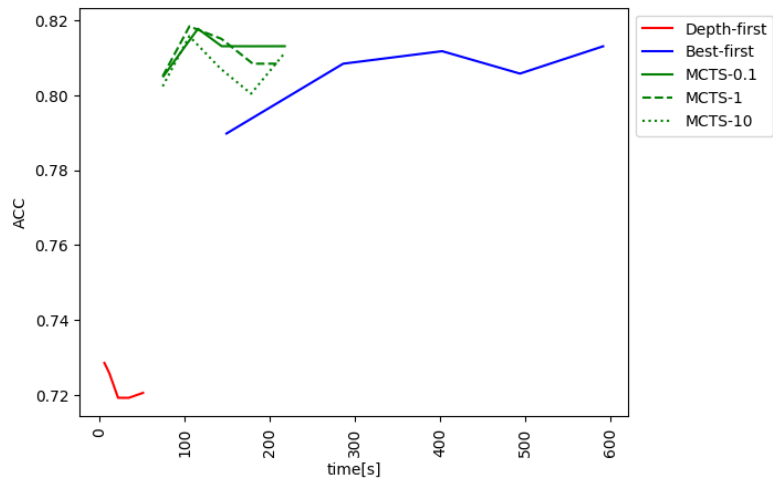
CPDB



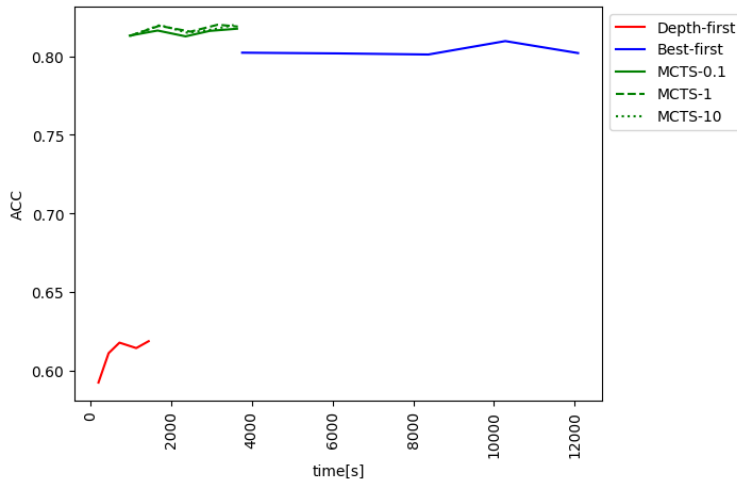
Mutag



AIDS



CAS

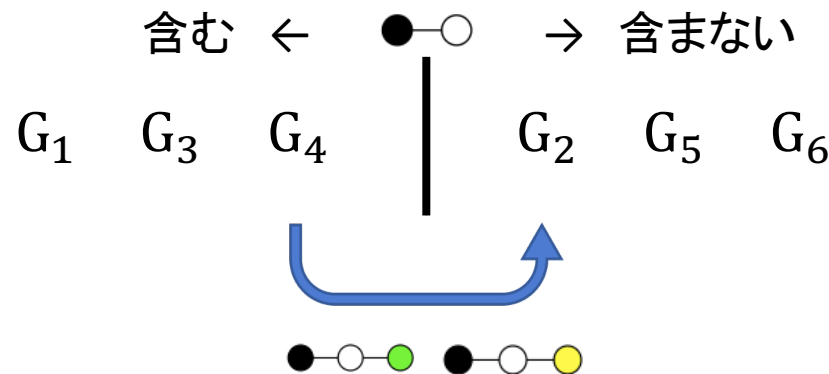


下限値の計算

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

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

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

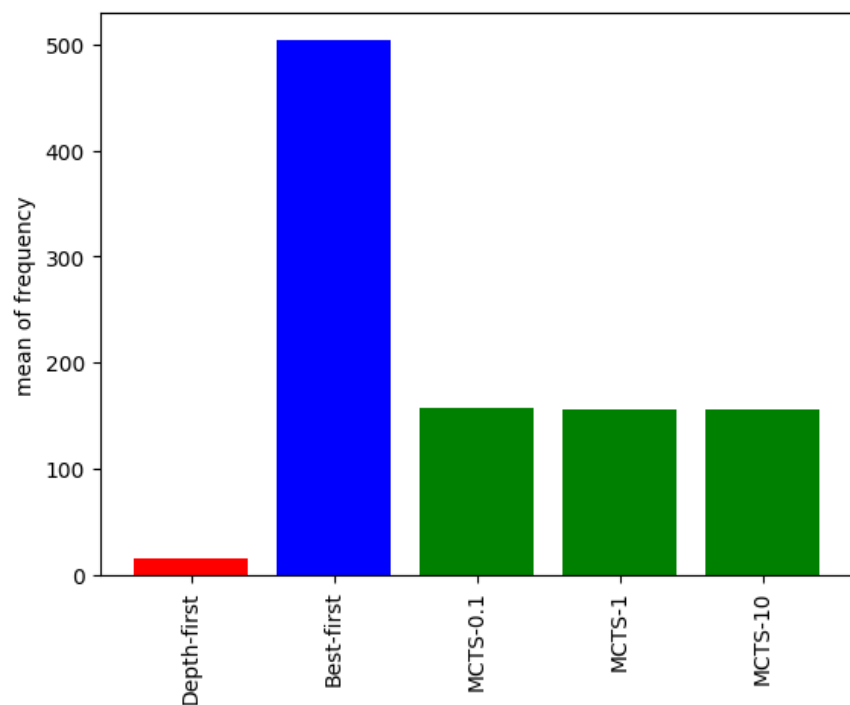
$S_{(\leq, k)}$ is a set of k pair (G_i, y_i) selected from $D_1(g)$ in descending order of y

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

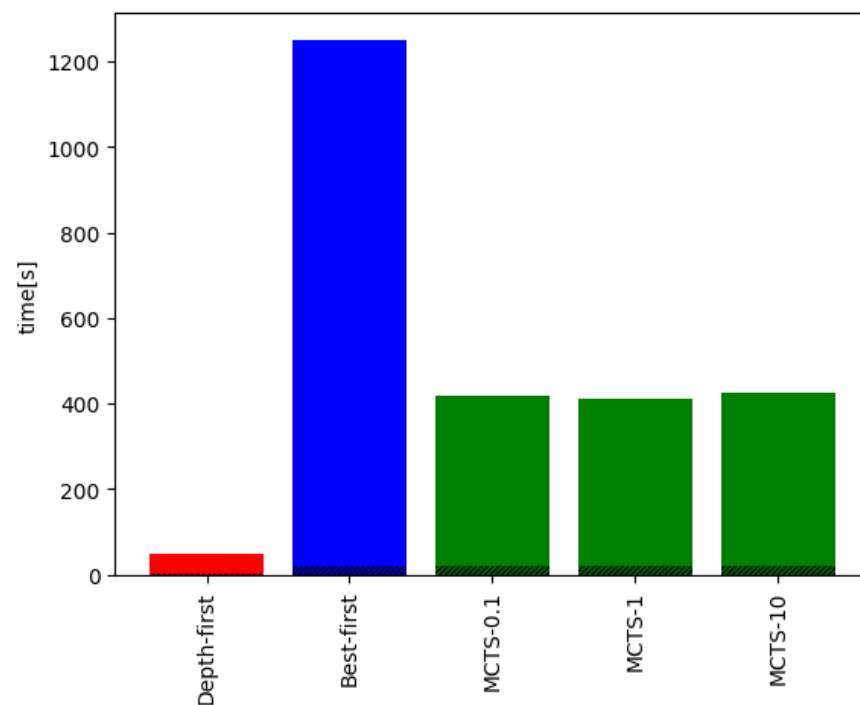
計算量：グラフ (g) の頻出度に対して線形オーダー

探索速度

探索グラフ頻出度平均

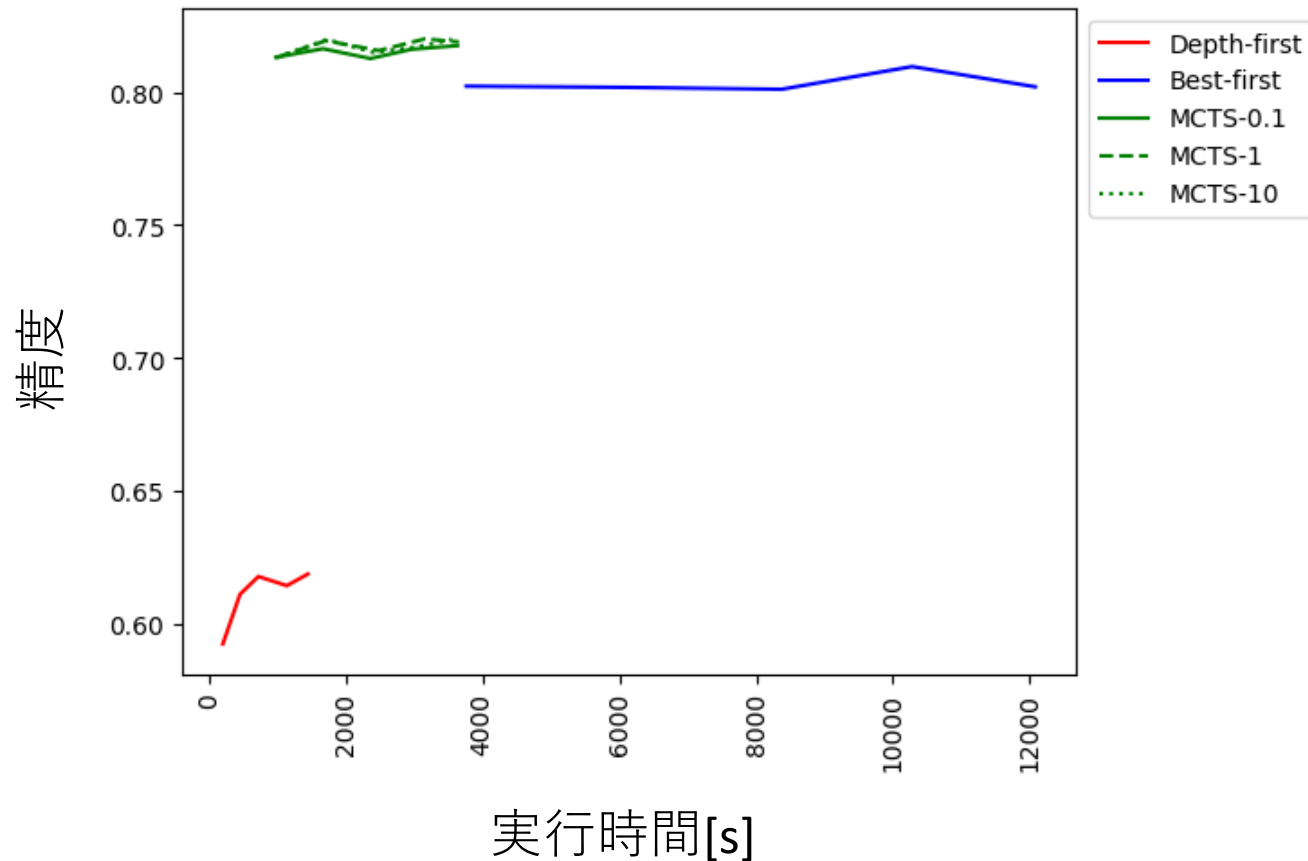


実行時間[s]



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

実験 2 (CAS)



- 同探索数での実行時間：深さ優先 < MCTS < 最良優先
- 同実行時間での精度はMCTSが最良