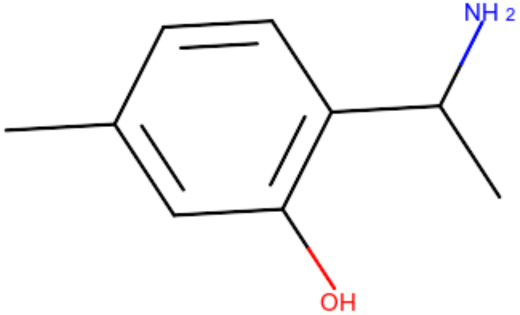
Subgraph-Feature Search for Learning Classifiers and Regressors under Fixed Budget Constraint

情報認識学研究室　白川　稜

背景

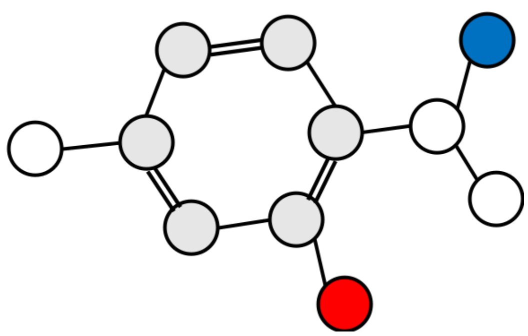
グラフは広く用いられる重要なデータ構造

- 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

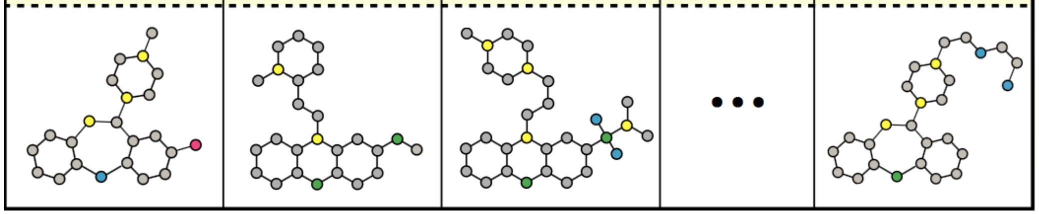
- 創薬の分野
- 生命科学や物質化学の分野



グラフ分類・回帰問題

Input: グラフデータ

G_1	G_2	G_3		G_n



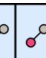
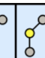
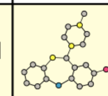
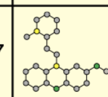
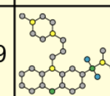
予測器
f

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

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	…
⋮	⋮									…

<u>特徴量</u>

　部分グラフの有無

<u>問題点</u>

　グラフサイズに対して部分グラフの総数は組合せ爆発

<u>既存研究</u>

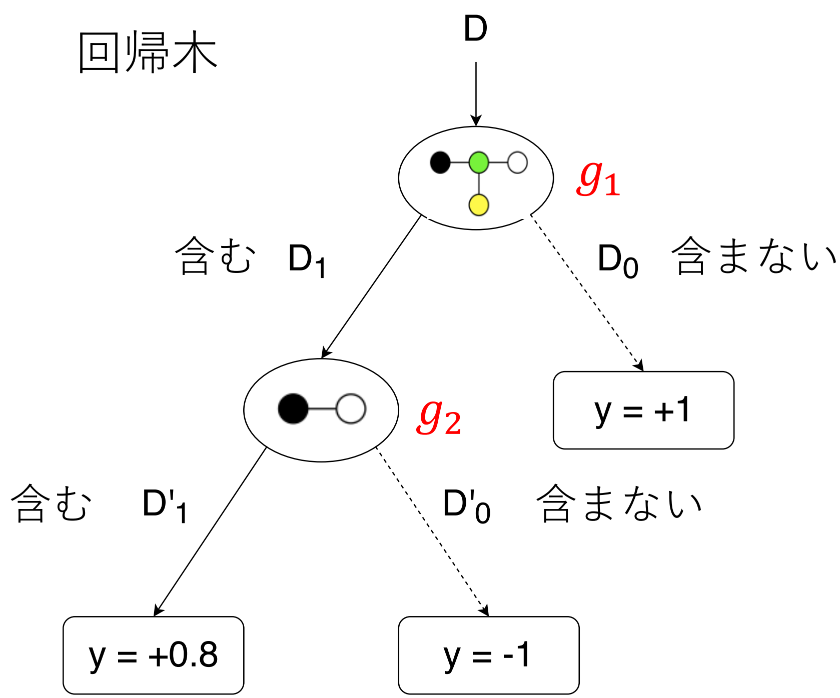
・2-step approach [Wale et al., 2007]
　ー 制約付き部分グラフ列挙 + 任意モデルの学習
・Simultaneous approach [Saigo et al., 2009][Shirakawa et al., 2018]
　ー モデルの学習と部分グラフ探索・選択の同時手法

GTB [Shirakawa et al., 2018, MLG (KDD workshop)]

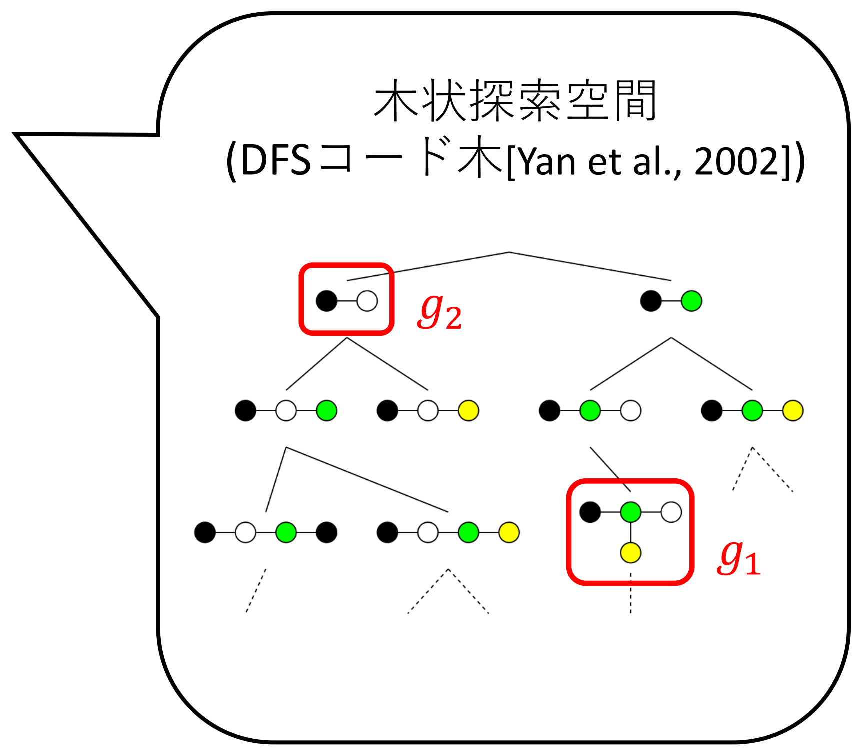
モデルの学習と部分グラフ探索・選択を同時に行う

モデル (Gradient Tree Boosting)

回帰木



特徴探索 (本研究)



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

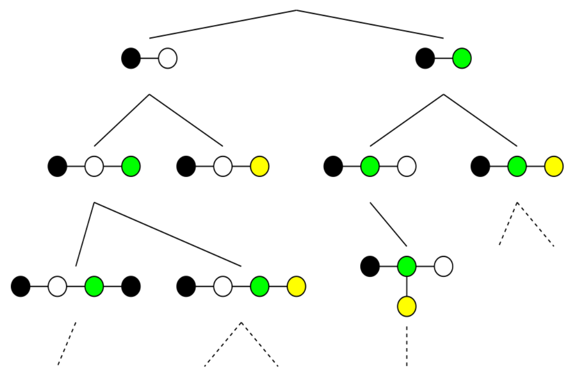
$$\min_g [\, \mathrm{TSS}(\mathrm{D}_0(g)) + \mathrm{TSS}(\mathrm{D}_1(g)) \,]$$

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

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

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

$\sqsupseteq$: 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の下限値（Bound）を計算

➡　枝刈りを利用した厳密探索

目的・提案

<u>問題点</u>

・枝刈りをしてもまだ厳密計算のコストが大きい

探索の計算コスト＝（1回の探索コスト）×(出力決定森のノード数)

<u>目的</u>

決定森の精度の劣化なしに探索コストの削減

<u>提案</u>

・厳密探索 ➡ 近似探索

・深さ優先 ➡ 有望方向優先

提案手法

近似探索

事前に１特徴探索あたりのコスト(ノード数)を設定

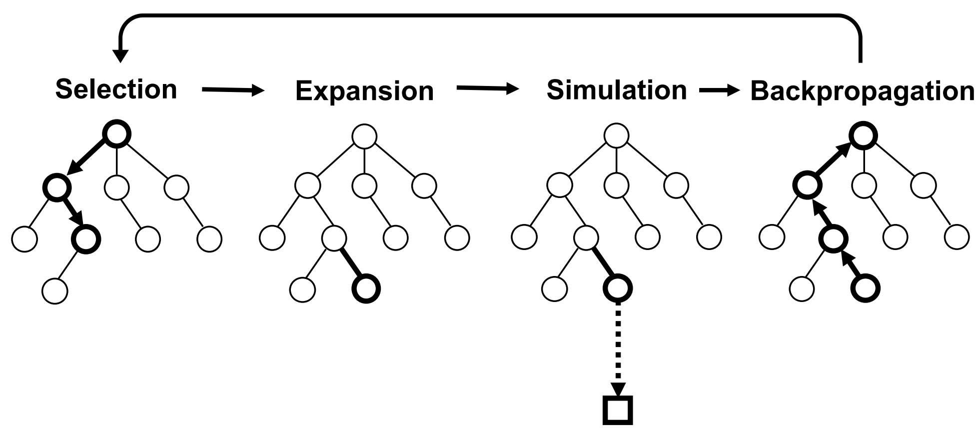
探索方針

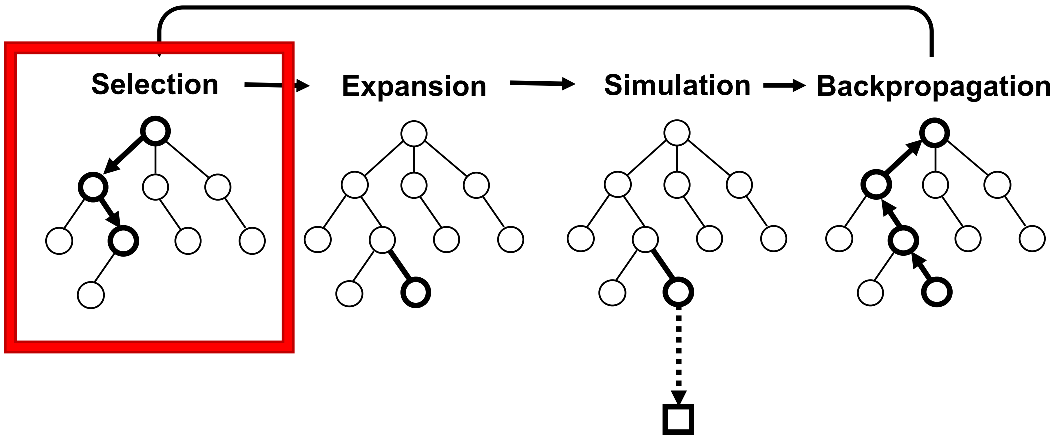
- 最良優先探索
- モンテカルロ木探索（MCTS）

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

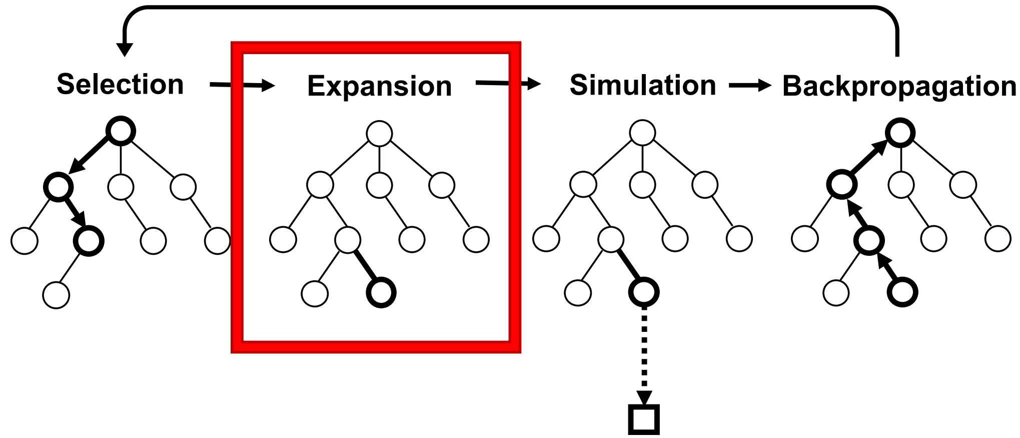
  根ノードを始点にUCBの値に基づき

  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

- 有望な方向を優先して探索



- Expansion

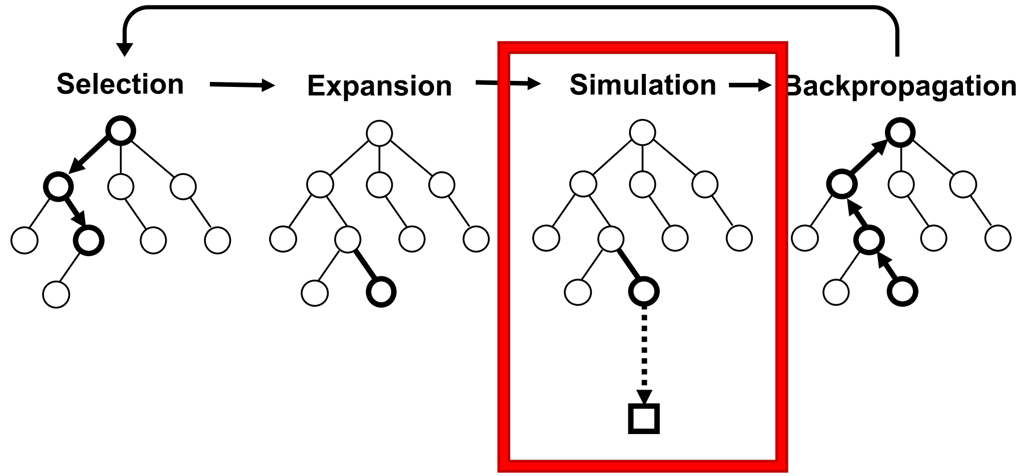
末端ノードが**初訪問**

➡　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

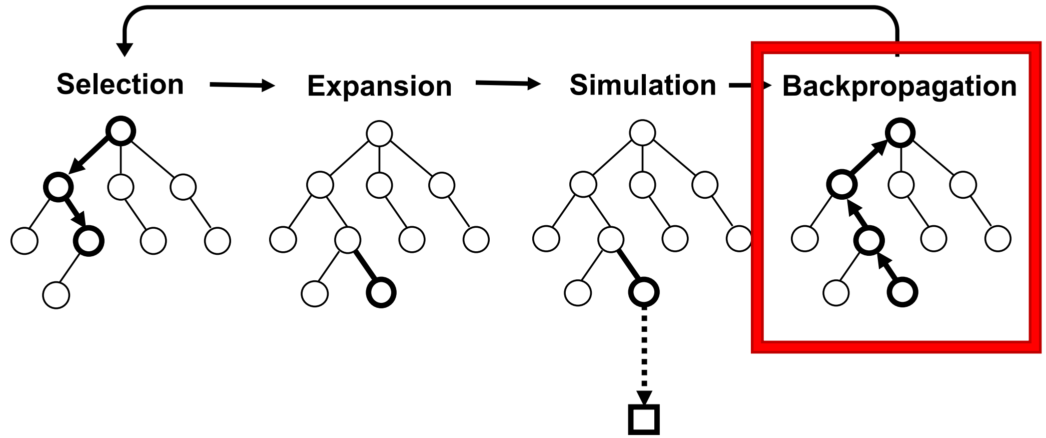
モンテカルロシミュレーションによりパスを降下

× 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

$$報酬 = -\frac{\mathrm{TSS}\big(\mathrm{D}_0(g)\big) + \mathrm{TSS}(\mathrm{D}_1(g))}{\mathrm{TSS}(\mathrm{D}_0(g) \cup \mathrm{D}_1\ (g))}$$

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

実験データセット

実データセット

Dataset	CPDB	Mutag	AIDS(CAvsCM)	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

CAS ➡ 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
$y \in [-1, +1]$

実験1

目的

提案法が少ないコストで良い特徴を見つけることの確認

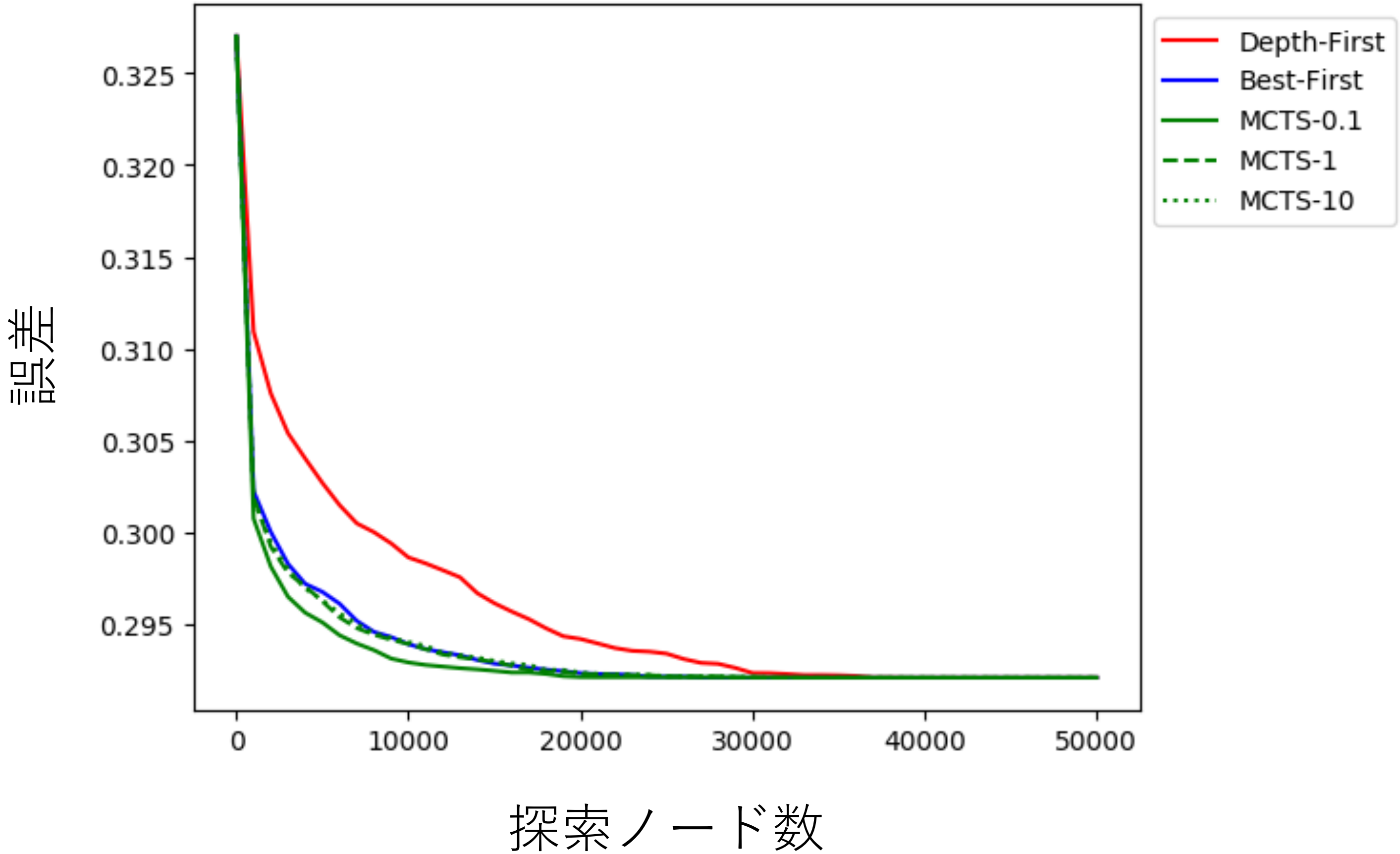
手法

人工データセットに対してコストに対するTSS改善の様子を各探索手法で比較

- 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

近似探索に基づく学習モデルの性能の比較

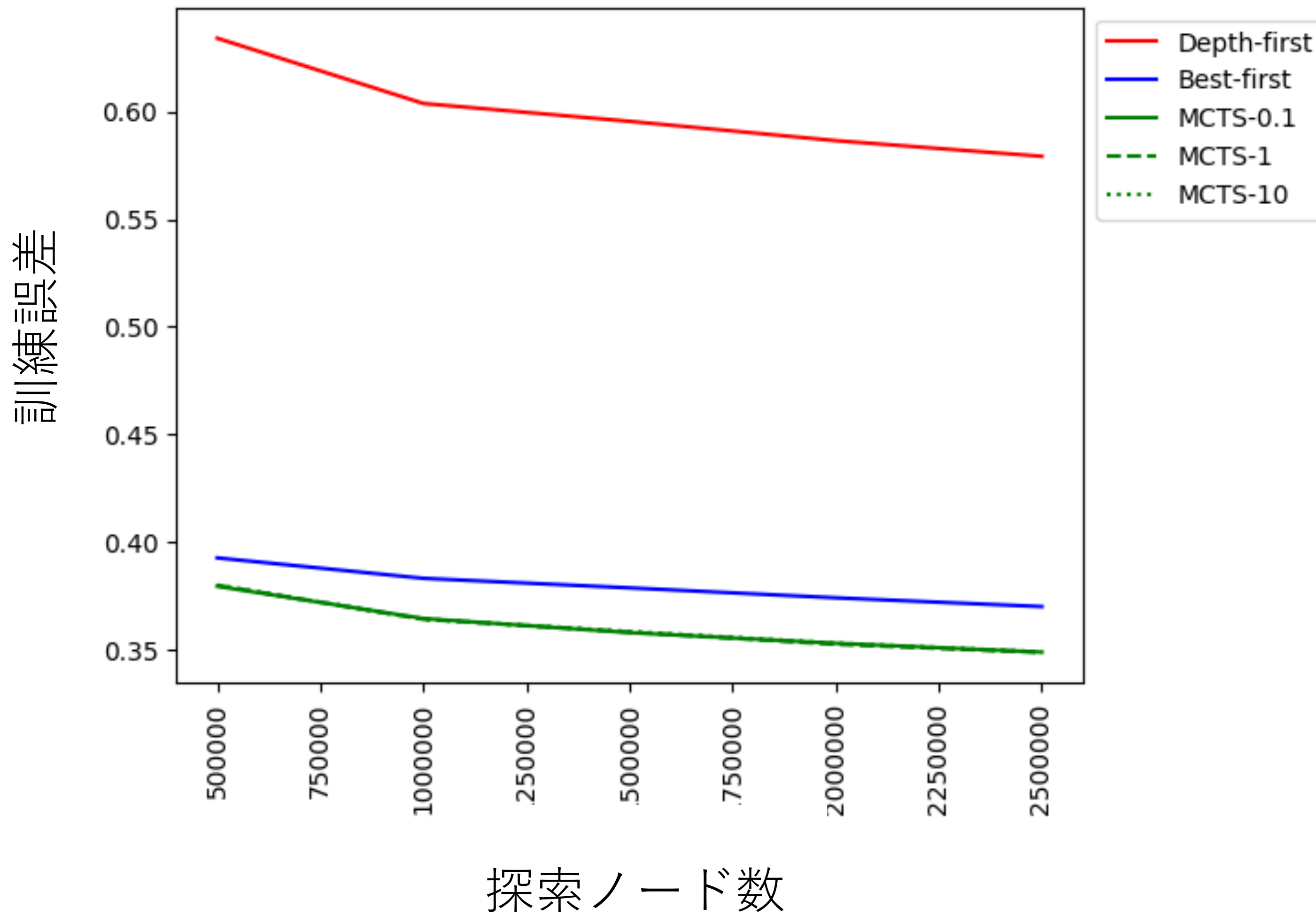
手法

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GTBによる出力決定森の精度比較

学習パラメータ
木の本数：100
木の深さ：１
ステップ幅：１

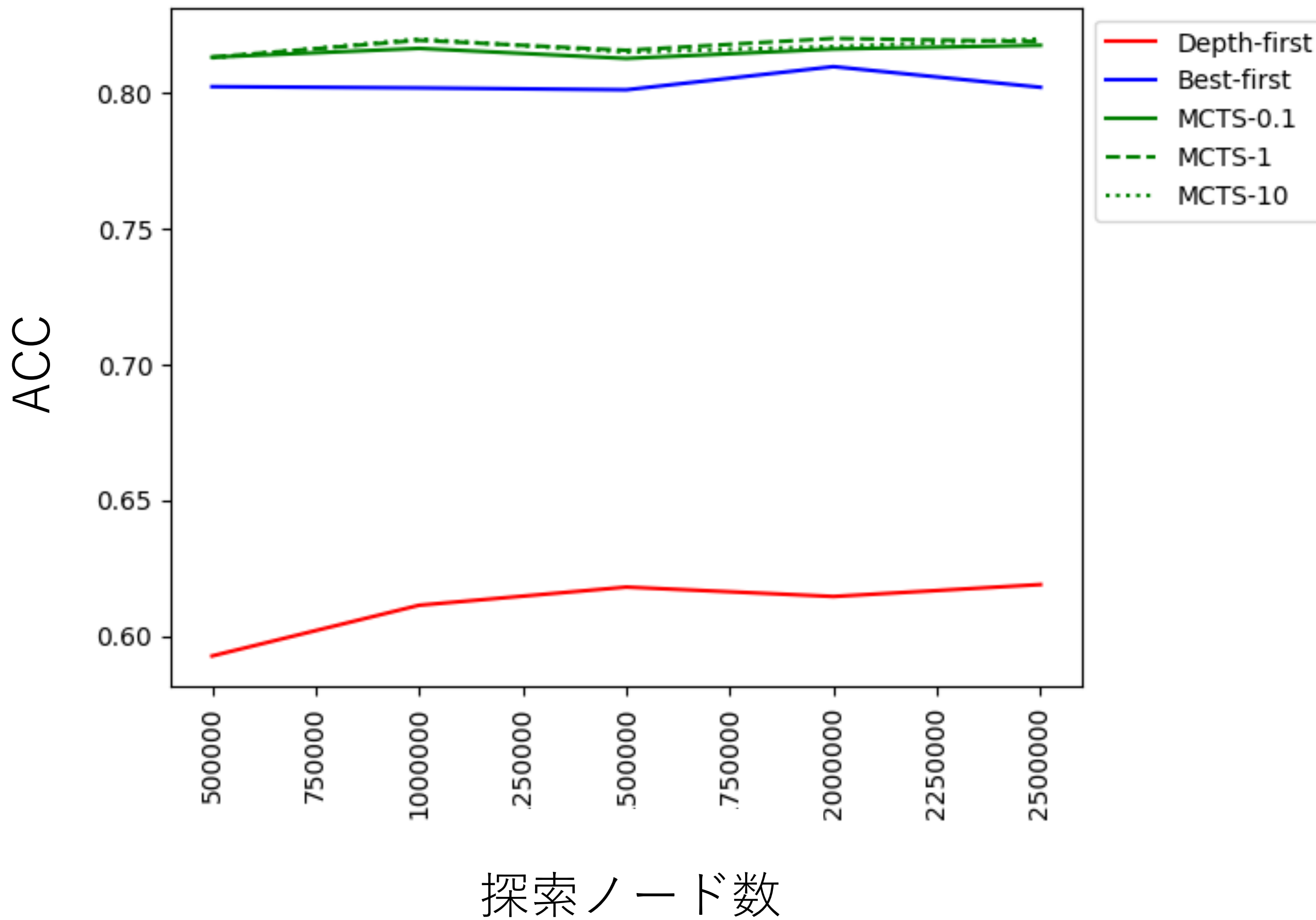
コスト制約(一回の特徴探索にかけるノード数)
CPDB：（1000, 2000, 3000, 4000, 5000）
Mutag：（200, 400, 600, 800, 1000）
AIDS：（1000, 2000, 3000, 4000, 5000）
CAS：（5000, 10000, 15000, 20000, 25000）

実験2 (CAS)



Depth-first
Best-first
MCTS-0.1
MCTS-1
MCTS-10
0.60
0.55
0.50
0.45
0.40
0.35
訓練誤差
500000
750000
1000000
250000
500000
750000
2000000
2250000
2500000
探索ノード数

実験 2 (CAS)



ACC
0.80
0.75
0.70
0.65
0.60
500000
750000
1000000
250000
500000
750000
2000000
2250000
2500000
探索ノード数
Depth-first
Best-first
MCTS-0.1
MCTS-1
MCTS-10

実験３

目的

従来厳密探索と**提案近似探索**に基づく学習モデルの性能の比較

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GTBによる出力決定森の精度・計算時間比較

・従来：GTB＋深さ優先厳密探索

・提案：GTB＋MCTS近似探索

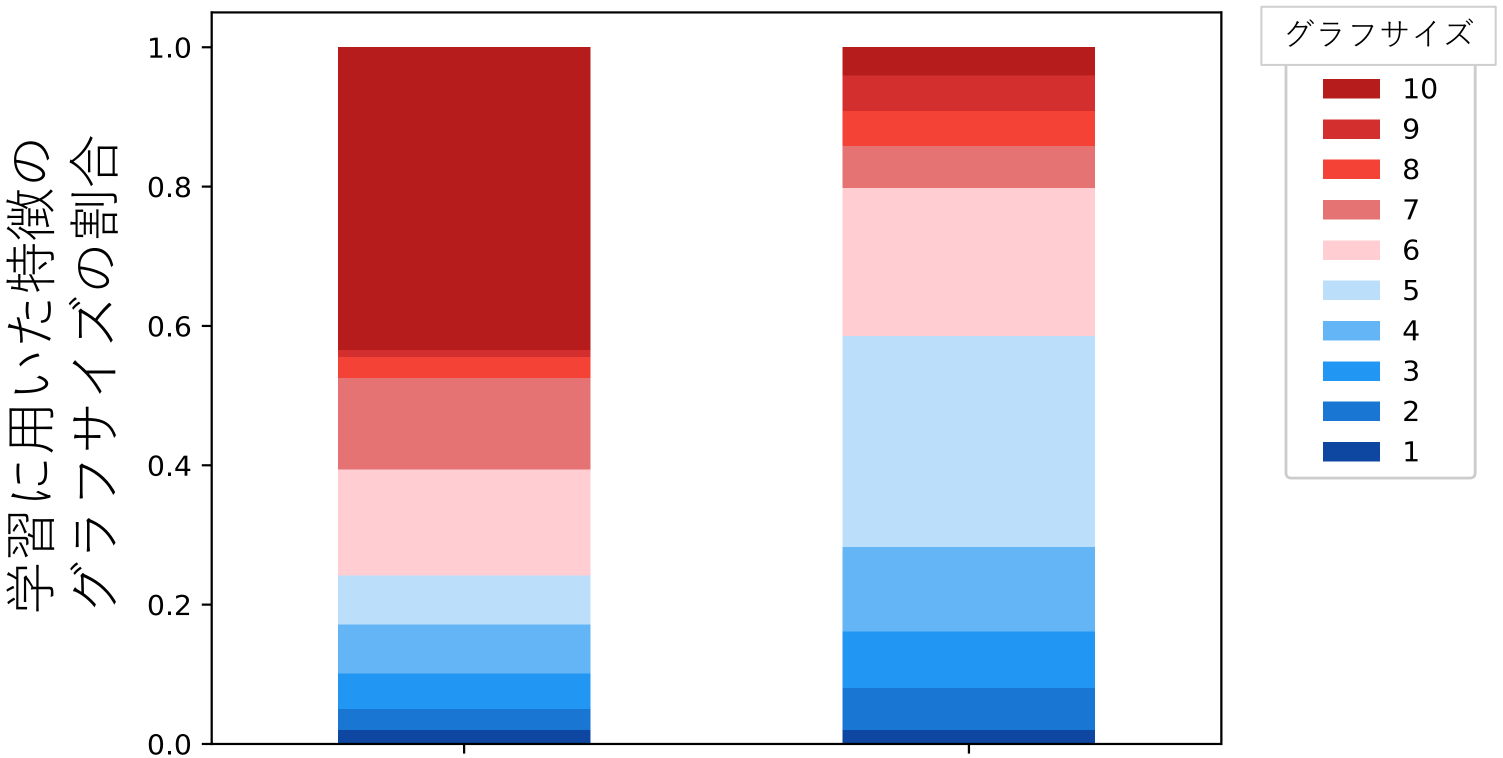
※MCTSの探索強度パラメータ、コスト制約は実験２の最良値

実験３

データ	ACC[%]		探索ノード数		実行時間[s]	
	従来	提案	従来	提案	従来	提案
CPDB	77.78	**78.35**	7.2×10^6	$\mathbf{5.0\times10^5}$	8.2×10^2	$\mathbf{6.2\times10}$
Mutag	85.03	**87.73**	3.8×10^5	$\mathbf{6.0\times10^4}$	2.3×10^2	$\mathbf{3.7}$
AIDS	81.37	**81.84**	7.9×10^7	$\mathbf{2.0\times10^5}$	2.5×10^4	$\mathbf{1.1\times10^2}$
CAS	80.82	**81.99**	6.9×10^7	$\mathbf{2.0\times10^6}$	8.0×10^4	$\mathbf{1.7\times10^3}$

- 精度の劣化なしに省コスト化を達成

提案法のACCが高い理由は？



従来　　提案

- 特徴のグラフサイズの偏りが少ない
  ➡ 過学習が避けられているのでは

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 ~1/200のコストで
同等、それ以上の精度のモデルを構築

発表実績

- 白川稜, 中村篤祥, 工藤峰一, モンテカルロ木特徴探索に基づく非線形グラフ分類回帰 第22回情報論的学習理論ワークショップ(IBIS2019), 名古屋.
- R Shirakawa, A Nakamura, M Kudo, “Learning a Nonlinear Model of Subgraph Features Using Monte Carlo Tree Search.” ACML 2019 Workshop on Statistics & Machine Learning Researchers in Japan, (2019), Nagoya.
- Shirakawa R, Yokoyama Y, Okazaki F, Takigawa I, “Jointly Learning Relevant Subgraph Patterns and Nonlinear Models of Their Indicators.” 2018 International Workshop on New Frontiers in Convergence Science and Technology,The 21th Hokkaido University-Seoul National University Joint Symposium, Sapporo.
- 白川 稜・横山 侑政・岡崎 文哉・瀧川 一学, 適応的な部分グラフ指示子の探索・選択に基づく非線形グラフ分類回帰 第21回情報論的学習理論ワークショップ(IBIS2018), 札幌.
- Shirakawa R, Yokoyama Y, Okazaki F, Takigawa I, “Jointly Learning Relevant Subgraph Patterns and Nonlinear Models of Their Indicators.” The 14th International Conference on Mining and Learning with Graphs (MLG 2018), KDD'18 Workshop, London.