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

白川　稜，　横山　侑政，　岡崎　文哉，　瀧川　一学　　北海道大学　email:sira@ist.hokudai.ac.jp

概要

- グラフに対する教師付き学習(分類・回帰)
- 部分グラフの総数は膨大、全列挙困難
- 適応的な部分グラフ指示子の探索・選択に基づく効率的な非線形モデル構築法の提案
- 実データ及び人工データに対する実験並びに精度、スケーラビリティの評価

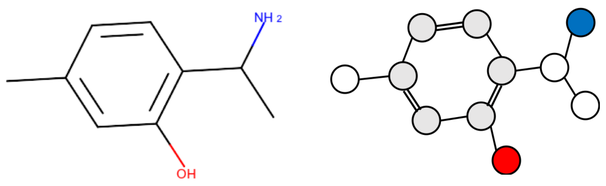
背景

グラフ は広く用いられる重要なデータ構造

- 化学構造式
- RNA二次構造
- 構文木

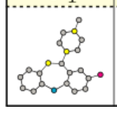
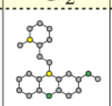
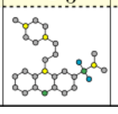
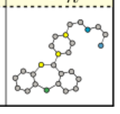
グラフに対する教師付き学習

- 様々な分野での応用
  - 創薬
  - 材料科学



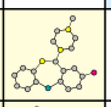
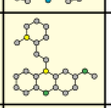
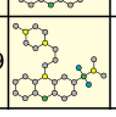
グラフに対する教師付き学習

入力 ラベル (y：離散, 実数値) 付きグラフ集合

y_1	y_2	y_3		y_n
0.1	0.7	1.2	⋯	0.9
G_1	G_2	G_3		G_n
			⋯	

出力 未知のグラフに対するラベルを予測する予測モデル

特徴量 部分グラフ指示子

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	...

既存研究

- 2-step 手法(Wale+ 2007)
  事前選択された特徴の列挙 + 任意モデルでの学習
  → 事前に選択される特徴に大きく影響
- gBoost(Saigo+ 2009)
  適応的部分グラフ指示子の探索・選択に基づく線形モデル
  → 全部分グラフ指示子の考慮が可能

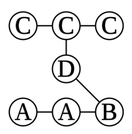
アプローチ

- branch & bound手法を用いた特徴探索により全部分グラフ指示子を考慮した回帰木の学習
- 精度及び不安定性向上のためアンサンブル学習(勾配ブースティング)を基にした非線形モデルの構築
- 実験による線形、非線形モデル間の比較

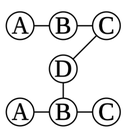
実験＆結果

精度予測(%) Graph-XOR

"Graph-XOR"：非線形な学習を要するグラフ版XOR



$y = +1$



$y = -1$

Group 1	Group 2
A–A–A	B–B–B
C–C–C	A–A–B
B–C–C	A–B–C

非線形モデルが必要

非線形モデル	線形モデル	
提案手法	提案手法 ($d1$)	gBoost
100.0	64.3	70.0

d: 木の深さ

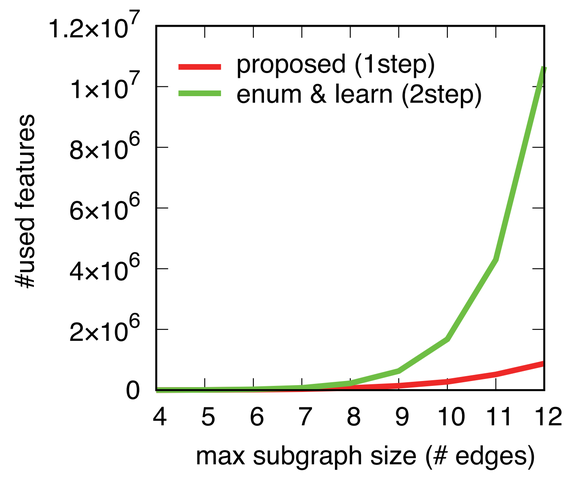
精度予測 (%) QSAR

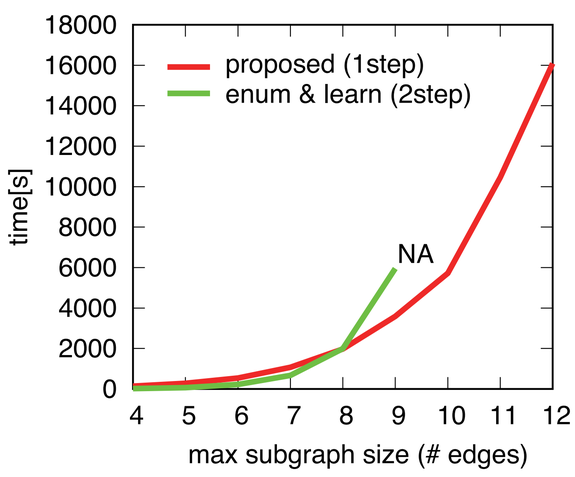
実データに対しても高精度

	CPDB	Mutag	NCI1	NCI47
非線形モデル 提案手法	**79.3**	87.8	**84.7**	**84.5**
線形モデル 提案手法 (d1)	**79.3**	87.8	83.1	82.8
gBoost	77.1	**91.4**	82.7	81.3

2-step 手法とのスケーラビリティの比較

2-step手法よりもスケールする





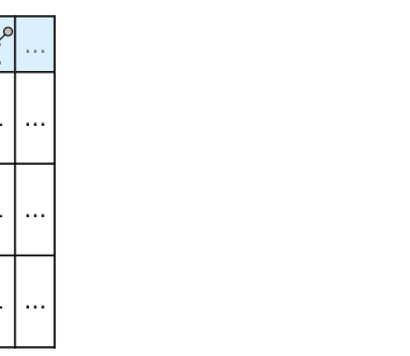
提案手法

非線形グラフ分類回帰モデル

回帰木
入力データに対して
内部ノードで質問し最適な分割を行う
葉ノードで定数値を返す

質問：
ある部分グラフを含むor含まない

勾配ブースティングは分類木ではなく回帰木が必要



勾配ブースティング
加法的アンサンブルモデル

$$F(G) = T_0(G) + sT_1(G) + sT_2(G) + sT_3(G) + \cdots \quad (1)$$

T_k：各反復における残差r_iに対する回帰木.

$$r_i = \frac{\partial \mathrm{L}(y_i, F_{k-1}(G_i))}{\partial F} \quad (2)$$

s：学習率,　L：損失関数.

内部ノードにおける分割ルールの学習
二乗誤差和を最小化する分割ルール(部分グラフ)の学習

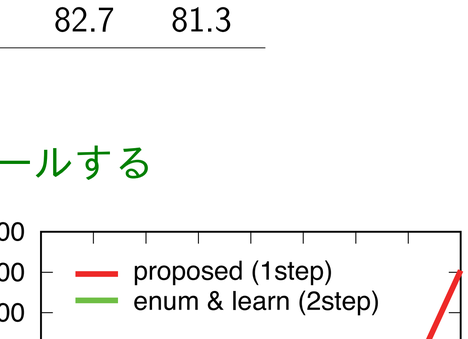
$$\arg \min_{x_j \in X} \left[\mathrm{TSS}(D_1(x_j)) + \mathrm{TSS}(D_0(x_j))\right] \quad (3)$$

X：全部分グラフ集合(全列挙は困難)
$D_1(x_j)$：{x_jを含むグラフ集合},　$D_0(x_j)$：{x_jを含まないグラフ集合}
$\mathrm{TSS}(D)$：残差r_iに対する二乗誤差和

内部ノードにおける分割ルールの学習

木探索(列挙木)

全部分グラフ集合は部分グラフ同型関係により
木状探索空間に表現が可能

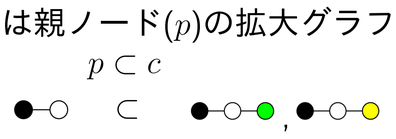


列挙木

木探索の特性

子ノード(c)は親ノード(p)の拡大グラフとなる

$p \subset c$



→ 子孫ノードcを含むグラフは常に親ノードpを含む
→ 子孫ノードcを含むグラフ集合は親ノードpを含むグラフ集合の部分集合となる

$$D_1(c) \subseteq D_1(p) \quad (4)$$

最適部分グラフの探索

- 列挙木の性質に基づいたBranch & bound探索

定理.
$D_1(g)$と$D_0(g)$が与えられる時, $g' \supset g$を満たす全ての部分グラフに対して以下が成立

$$\mathrm{TSS}(D_1(g')) + \mathrm{TSS}(D_0(g')) \geq \min_{(\diamond, k)} \left[\mathrm{TSS}(D_1(g) \setminus S_{\diamond,k}) + \mathrm{TSS}(D_0(g) \cup S_{\diamond,k})\right]$$

$(\diamond, k) \in \{\leq, >\} \times \{2, \ldots, |D_1(g) - 1|\}$, $S_{\diamond,k} \subset D_1(g)$,
$S_{\leq,k}$は$D_1(g)$を残差に関して降順にした際の上からk番目までの集合. $S_{>,k}$は昇順.

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, U.K., August 20, 2018