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

白川 稜, 中村 篤祥, 工藤 峰一 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)
- 特徴量に部分グラフ指示子を利用
- 部分グラフの総数は膨大、全列挙困難
- 特徴探索及びモデル構築の同時学習
- モンテカルロ木探索を利用した効率的な部分グラフ指示子の探索・選択
- 回帰木勾配ブースティングによる非線形モデルの構築

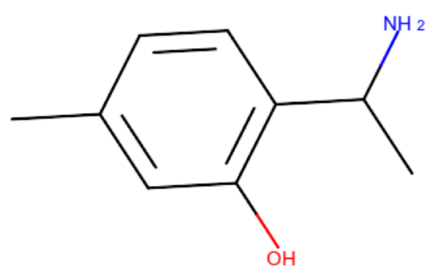
背景

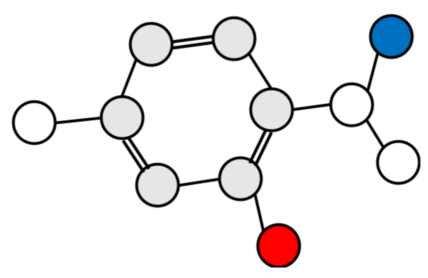
グラフ は広く用いられる重要なデータ構造

- 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
  厳密探索により、 探索コスト大

アプローチ

- 適応的な特徴探索に基づく回帰木モデルの学習（全部分グラフ指示子を考慮）
- 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
- 特徴探索において、 モンテカルロ木探索を利用することで探索コストを削減

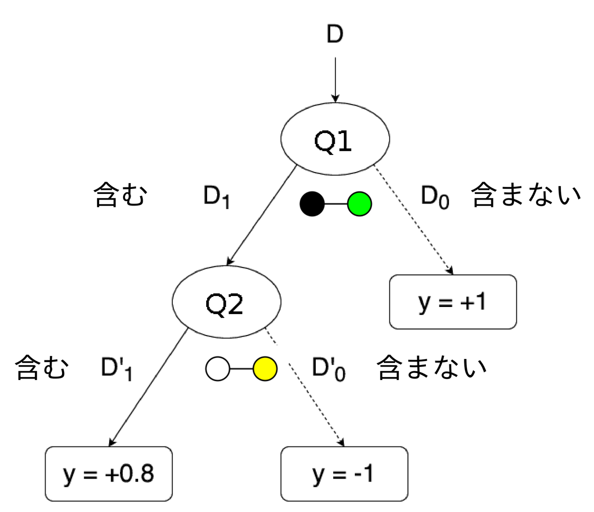
提案手法

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

加法的アンサンブルモデル

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

T_k：各反復における残差r_iに対する回帰木.

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

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

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

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

枝刈り規則

探索空間の特性

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

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

評価値の上限

$D_1(g)$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.

内部ノードにおける分割ルールの学習

探索空間