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 大学院 情報科学研究科)†

1　はじめに

グラフ構造は、事物間の関係を表現することに適しているため、様々な分野のデータ表現として用いられている。特にケモインフォマティクスの分野では、化学構造式やRNA 二次構造などのデータの表現としてグラフを用いた研究が盛んに行われている。

本研究では、そのうちの 1 つとしてグラフ分類回帰問題を扱う。グラフ分類回帰問題とは、ラベル付きグラフ集合をトレーニングデータとして与えられた際に未知のグラフのラベルを予測する予測器を学習する問題であり、主に創薬や材料科学分野におけるスクリーニングとしての応用がある。

本問題に利用する特徴量には一般的に部分グラフの有無（部分グラフ指示子）が挙げられるが、部分グラフの総数はグラフサイズに対して指数関数的に増加するため特徴構築が困難である。本研究では、回帰木を利用したアンサンブル学習モデルとモンテカルロ木特徴探索を組み合わせることで、表現力の高い学習モデルを低コストで構築する手法を提案する。最後に、グラフ分類回帰問題におけるベンチマークのデータを利用し、既存手法との精度及びコスト比較の実験を考える。

2　問題設定

まずはじめにグラフ分類回帰問題の問題設定から記述する。グラフ分類回帰問題は、あるグラフクラス $\mathcal{G}$ 上の関数 $f : \mathcal{G} \to \mathbb{N}\ or\ \mathbb{R}$ をトレーニングデータ $\mathcal{D}$ から学習する問題である。出力が離散値の場合は分類問題、実数値の場合は回帰問題となる。

3　既存研究

特徴量に部分グラフ指示子を利用した既存手法として gBoost[1] という手法を説明する。

3.1　モデル

gBoost は LPBoost と呼ばれる線形計画問題によって定式化されたアンサンブル学習モデルと木状の部分グラフ空間における探索アルゴリズムを組み合わせた手法である。gBoost では特徴探索とモデル構築を独立に行わずにモデル構築に必要な特徴のみを逐次的に探索する。

本手法では弱仮説に決定株（Fig.1）を用い、弱仮説の重み付き線形和を最終的な予測器とする。決定株の分割ルールに用いる特徴を gSpan 木（Fig.2）と呼ばれる部分グラフ空間を探索することで決定する。本手法では、特徴探索において gSpan 木の親子ノードに関する包含関係を利用して有効な枝狩りを行い探索コストの削減を行う。

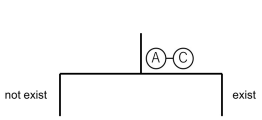

Fig. 1　決定株

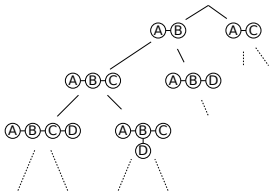

Fig. 2　gSpan 木

3.2　問題点

既存手法の問題点としては以下の二点が挙げられる。

- モデルの表現力

既存手法の弱仮説は決定株であり全体のモデルもその線形和で表現されるため、既存手法は線形モデルとなる。線形モデルは非線形モデルと比較して表現力で劣るため、学習精度の低下につながりうる。

- 探索コスト

探索効率のため枝刈りを利用するが、既存手法では膨大な部分グラフ空間を厳密に探索する。スケールの大きな問題では依然探索のコストが膨大である。

4　提案手法

本研究では回帰木アンサンブルモデルとモンテカルロ木探索アルゴリズムを組み合わせることで非線形モデルを効率的に構築する手法を提案する。

4.1　学習モデル

提案手法では弱仮説として回帰木を用いる。回帰木とは再帰的に集合分割を行い、集合のラベルを実数値（平均値）で予測するモデルである。分割ルールは以下の式に従い決定される。

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

X : 全部分グラフ集合, TSS : 二乗誤差和,

$D_1(x_j) : \{x_j$を含むグラフ集合 $\}$, $D_0(x_j) : \{x_j$を含まないグラフ集合 $\}$

回帰木は非線形モデルであり決定株よりも表現力の高いモデルである。また勾配ブースティングによりアンサンブルを取ることで精度及び安定性の向上を図る。

*sira@ist.hokudai.ac.jp
†札幌市北区北 14 条西 9 丁目北海道大学大学院情報科学研究科

4.2 モンテカルロ木探索

本手法では弱学習器を回帰木に拡張したことにより集合分割の回数が既存手法よりも増加する。

そこで本手法ではモンテカルロ木探索により特徴探索を行うことで一度の探索におけるコストの削減を図る。モンテカルロ木探索 [2][3] とはモンテカルロシミュレーションと解の評価値のの見積もりをうまく組み合わせた強化学習の一つであり、膨大な探索空間から限られたコスト内でより良い解が得られる手段として注目されている。

本探索ではモンテカルロ木探索の一種である UCT アルゴリズムを利用し分割特徴を決定する。UCT アルゴリズムは大きく分けて 4 つの操作を反復することにより探索を行う。最終的な分割ルールはモンテカルロ木探索によって拡大される探索空間のうち、最も回帰誤差の小さい特徴を用いる。

- Selection

Selection では UCB(Upper Confidence Bound) の値を利用して、根ノードから各反復時点での探索空間の末端まで子ノードの選択を繰り返す。

$$\underset{c_j \in \{\text{ 子ノード集合 }\}}{\arg\max} \quad UCB$$

$$UCB = \bar{V} + C\sqrt{2\frac{logN}{n_j}}$$

$\bar{V}$: c_jの選択による報酬平均, C : 探索強度パラメータ,

N : 親ノードの選択回数, n_j : c_jの選択回数,

また本探索の報酬に関しては、TSS の値による回帰誤差の符号を負にしたものを利用し、誤差が大きいほど報酬が低く誤差が小さいほど大きな報酬を得られるよう設計する。

- Expansion

Expansion では Selection によって選択された末端ノードについて考える。末端ノードの選択回数がある閾値を上回る際には末端ノードから子ノードを展開し各時点での探索空間に追加する。加えて子ノードの中から 1 つを選択する。

- Simulation

Simulation では Selection 及び Expansion によって選択された末端ノードからモンテカルロシミュレーションによりノードを展開する。ここで愚直に gSpan 木を展開するのではなく、トレーニングデータ中の 1 つのグラフを無作為に選択しそのグラフ上で拡大しうるシミュレーションのみを扱うことで 1 回のシミュレーションにかかるコストの削減を図る。またシミュレーションによる拡大を制限するため、一定確率で拡大を停止する停止条件を設ける。

- Backpropagation

Backpropagation では Simulation により決定されたノードに関して報酬値を計算し、Selection により選ばれたノードの報酬平均値 $\bar{V}$ を更新する。

4.3 探索空間

ここでは探索空間の形状を考える。既存手法では gSpan 木を探索空間として利用した。しかし、gSpan 木では固有の辞書順を定め重複無しに部分グラフパターンを列挙していることから、木全体の形が大きく偏るという特徴を持つ。モンテカルロ木探索においてこの木の偏りは探索の偏りにつながるため、本提案手法では gSpan 木を DAG 状の空間に拡張することで探索空間の偏りを減らし探索が有効に働くよう設計する。

5 実験

本実験ではグラフ分類回帰問題におけるベンチマークである QSAR データセットを用いて、既存手法との学習精度及び学習コストの比較を行うことで、提案手法の有用性を検証する。また、モンテカルロ木探索におけるハイパーパラメータをそれぞれ変動させ、各ハイパーパラメータの学習結果に与える影響を調べる実験を行いたいと考える。

6 まとめ

回帰木アンサンブルモデルとモンテカルロ木特徴探索を組み合わせることでより軽量な非線形学習モデルを構築した。本探索は UCT アルゴリズムによるナイーブな手法であるため、本問題における適切なヒューリスティクスや RAVE などの発展的な手法を取り入れることで一層のパフォーマンスの向上が期待される。

参考文献

[1] Hiroto Saigo, Sebastian Nowozin, Tadashi Kadowaki, Taku Kudo, and Koji Tsuda. gboost: a mathematical programming approach to graph classification and regression. *Machine Learning*, Vol. 75, No. 1, pp. 69–89, 2009.

[2] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 4, No. 1, pp. 1–43, March 2012.

[3] Romaric Gaudel and Michèle Sebag. Feature Selection as a One-Player Game. In *International Conference on Machine Learning*, ICML 2010 Conference Proceedings Book, pp. 359–366, Haifa, Israel, June 2010.