構造的類似性を用いる適合度設計に基づく ブロック保存型外平面的グラフ対パターンの進化的獲得

細木 翔太 (2467010)

知能工学専攻 データ科学講座
指導：宮原 哲浩 准教授

1 はじめに

多くの化合物は外平面的グラフの構造を持つデータとみなすことができるため，現在までに機能がわかっている化合物データの構造的特徴を獲得することができれば，創薬などの分野において大きな意義があるといえる．ブロック保存型外平面的グラフ対パターン(BPOグラフ対パターン)は，変化前の構造を表すBPOグラフパターンと変化後の構造を表すBPOグラフパターンを一組として扱い，変化前後の構造的特徴を同時に保持するためのパターンである．これにより，構造の変化箇所を分かりやすく表現できる．先行研究[3]では，化学変化を可視化するためにBPOグラフ対パターンを導入し，進化的学習によって特徴的なBPOグラフ対パターンを獲得する手法を提案した．

しかし，先行研究の適合度設計は分類性能（Balanced Accuracy）のみに基づいていたため，BPOグラフ対パターンを構成する2つのBPOグラフパターンのうち，一方のみで分類性能が確保できる場合には，他方のパターンの構造が単純化しても適合度が低下しにくい．その結果，進化過程において，一方が極端に単純化された個体が選択され，変化前後のパターンを見比べても構造の変化箇所を十分に把握しにくいという課題があった．そこで本研究では，BPOグラフ対パターンの変化前後における構造対応の妥当性を評価に反映させるため，木編集距離の下界に基づく構造類似性指標を用いたペナルティ付き適合度を設計する．これにより，従来手法では高評価となり得た一方が極端に単純化された個体の選択を抑制し，構造変化の可視化に適したBPOグラフ対パターンの獲得を目指す．また，進化的学習の効率化を図るため，根となる頂点の次数や頂点ラベル情報に基づき，マッチングの条件を満たさない計算を早めに打ち切ることで，計算コストを削減する．

2 BPO グラフ対パターン

ブロック保存型外平面的グラフパターン(Block Preserving Outerplanar graph pattern,BPOグラフパターン)とは，ブリッジ変数，末端変数と呼ばれる 2 種類の構造的変数を持つ外平面的グラフの構造をしたパターンである． また， ブロック保存型外平面的グラフ対パターン(BPOグラフ対パターン)とは，2つのBPOグラフパターンの順序対であり，変化前後の構造的特徴を対として表現する．外平面的グラフ対(G_1, G_2)とBPOグラフ対パターン(p_1, p_2)に対し，p_1の変数を適当な外平面的グラフで置き換えたときにG_1と同型のグラフが得られ，かつp_2の変数を適当な外平面的グラフで置き換えたときにG_2と同型のグラフが得られるとき，(p_1, p_2)と(G_1, G_2)はマッチするという．図1にBPOグラフ対パターンと外平面的グラフ対のマッチ関係の例を示す．BPOグラフ対パターン(p_1, p_2)は外平面的グラフ対(G_1, G_2)にはマッチするが，$(G_1^{'}, G_2^{'})$にはマッチしない．

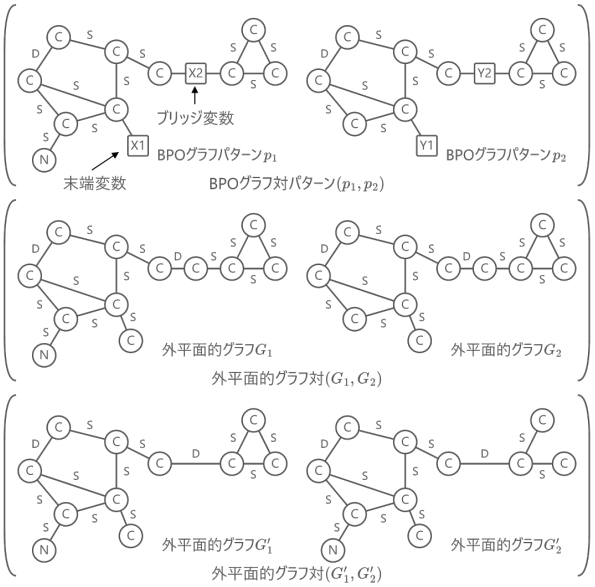


図 1: BPO グラフ対パターンと外平面的グラフ対のマッチ関係

3 特徴的な BPO グラフ対パターンの獲得問題

特徴的なBPOグラフ対パターン獲得問題を次に示す．

入力： 正事例および負事例からなる外平面的グラフ対の有限集合D.

問題： Dに関する適合度の高いBPOグラフ対パターン$p = (p_1, p_2)$を獲得する．

BPOグラフ対パターンpの外平面的グラフ対の有限集合Dに関する評価指標として，Balanced Accuracyに基づく基礎適合度$f(p)$を次のように定義する．

$$f(p) = \frac{1}{2}\left(r^{+}(p) + r^{-}(p)\right)$$

ここで，$r^{+}(p)$はpがDの正事例にマッチする割合，$r^{-}(p)$はpがDの負事例にマッチしない割合である．

4 木編集距離の下界に基づくペナルティ付き適合度

前章の基礎適合度$f(p)$は分類性能のみに基づくため，変化前後の構造対応の妥当性を十分に評価できない．本研究ではこの課題に対処するため，編集距離の下界を導入し，構造差に基づくペナルティを追加した新たな適合度を設計する．

編集距離の下界とは，任意の2つの木に対して，真の木編集距離以下であることが保証された値を返す距離関数であり，厳密な編集距離計算に比べて計算量が小さい．本研究では，BPOグラフ対パターン$p = (p_1, p_2)$を構成する2つのBPOグラフパターンp_1, p_2をその木表現であるブロック木パターンに変換し，Kailingらの下界距離関数[1]により，葉距離ヒストグラム(Leaf Distance Histogram)，次

数ヒストグラム(Degree Histogram)，ラベルヒストグラム(Label Histogram)の3種類のヒストグラムからブロック木パターン間の木編集距離の下界値を計算する．具体的には，2つのブロック木パターンから3種類のヒストグラムをそれぞれ作成し，種類ごとにL_1距離(マンハッタン距離)を算出する．得られた3つの距離に下界性を保つための係数(葉距離は1，次数は1/3，ラベルは1/2)を乗じ，乗じた結果の最大値を下界値として用いる．ここで葉距離ヒストグラムは，各ノードを根とする部分木の高さ(葉までの最長距離)の度数分布，次数ヒストグラムは各ノードの子ノード数の度数分布，ラベルヒストグラムはラベルの度数分布を表す．Kailing らの手法は根付き無順序木を対象とするのに対し，本研究で扱うブロック木パターンは根無し無順序木であるため，根の取り方によって下界値が大きく変動し得る．そこで本研究では，木の中心を根とした根付き無順序木に変換することでこの変動を抑える．また構造的特徴をより正確に捉えるため，ラベルヒストグラムには頂点ラベルに加えて辺ラベルと変数の情報も反映させる．この下界値を構造ヒストグラム距離(Structural Histogram Distance,SHD)と呼び，構造差の指標$L(p)$として用いる．

提案するペナルティ付き適合度$f'(p)$を次式で定義する．

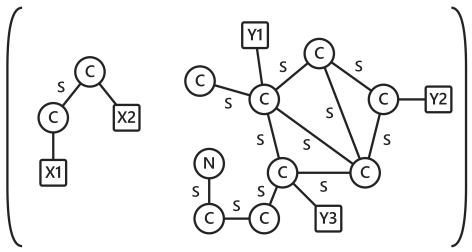
$$f'(p) = f(p) - w \max(0, L(p) - \tau)$$

ここで，wはペナルティの重み，τは許容閾値である．この定義により，構造差が過度に大きい個体に対してペナルティが課される．なお，ペナルティ付き適合度は，多目的最適化の考え方およびbloat抑制で用いられる罰則法の考え方を参考に，進化過程の選択に用いる指標として設計した[2]．最終的な分類性能の評価は基礎適合度$f(p)$により行う．また探索初期の多様性を損なわないため，本研究では基礎適合度$f(p)$が一定以上の個体に対してのみペナルティを適用する．

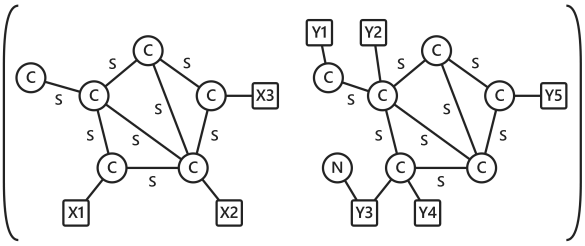
5 実験と結果

入力データとして人工データを用いた． 生成パターンと呼ぶBPOグラフ対パターンを一つ用意し，これにマッチする外平面的グラフ対500個を正事例，マッチしない外平面的グラフ対500個を負事例とする．提案するペナルティ付き適合度の効果を確認するため，作成した人工データを用いて遺伝的プログラミング(Genetic Programming,GP)を10試行行った．GPは，世代数1000，個体数80，エリート保存数3，遺伝操作確率は複製0.05，交叉0.50，突然変異0.45とした．ペナルティは$w = 0.03, \tau = 1.5$とし，基礎適合度$f(p)$≥0.85の個体にのみ適用した．比較は適合度にペナルティなし(従来手法)と適合度にペナルティあり(提案手法)の2条件とし，それ以外の設定は同一とした．

図2に，最終世代の最良個体の例を示す．図に示す個体の基礎適合度はいずれも$f(p) = 1.0$であり，分類性能を満たした上で得られた個体である．ペナルティなしでは，変化前のBPOグラフパターンが単純化した個体が得られたのに対し，ペナルティありではそのような単純化が抑制され，変化前後の構造対応をより把握しやすいBPOグラフ対パターンが得られた． また，10試行の最終世代の最良個体の平均基礎適合度は，ペナルティの有無に関わらず0.9163であり，差は見られなかった．図3に従来手法，提案手法の各世代における最良個体の10試行の平均基礎適合度の推移を示す．さらに，GPの世代数200，個体数80の設定(適合度にペナルティあり)でマッチング判定の高速化処理の有無を比較したところ，10試行の平均実行時間は高速化処理なしの設定の1745.847秒から高速化処理ありの設定の614.035秒へ短縮された(約65%減)．



(a) 従来手法 (ペナルティなし)



(b) 提案手法 (ペナルティあり)

図 2: 最終世代の最良個体の例

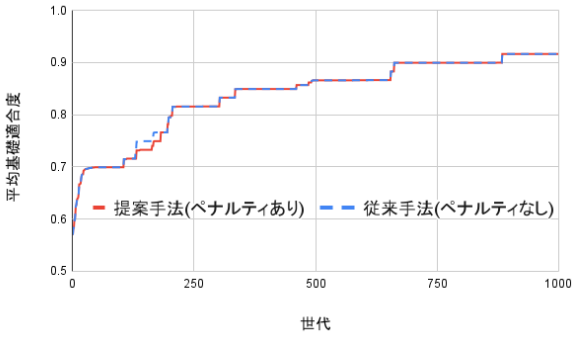


図 3: 従来手法，提案手法の各世代における最良個体の平均基礎適合度の推移

6 おわりに

本研究では，特徴的なBPOグラフ対パターンを獲得するために，分類性能のみに基づく適合度において一方のパターンが単純化しやすい課題に対して，木編集距離の下界に基づく構造差ペナルティを導入した適合度を設計した．また，実験により，分類性能を維持したまま，変化前後の構造対応を把握しやすいBPOグラフ対パターンを獲得できることを確認した．

参考文献

[1] Karin Kailing, et al. “Efficient similarity search for hierarchical data in large databases.” EDBT 2004, LNCS 2992, pp.676-693, 2004

[2] Riccardo Poli, et al. A Field Guide to Genetic Programming. Lulu Enterprises UK Ltd, 2008.

外部発表情報

[3] 細木 翔太 他 “進化的学習による特徴的なブロック保存型外平面的グラフ対パターンの獲得”2024 IEEE SMC Hiroshima Chapter 若手研究会講演論文集，pp.27-31，2024.