

外平面的グラフのクラスタリングによる 複合的なブロック保存型外平面的グラフパターンの進化的獲得

Evolutionary Acquisition of Multiple Block Preserving Outerplanar Graph Patterns by Clustering Outerplanar Graphs

徳原 史也 *¹
Fumiya Tokuhara

宮原 哲浩 *¹
Tetsuhiro Miyahara

久保山 哲二 *²
Tetsuji Kuboyama

鈴木 祐介 *¹
Yusuke Suzuki

内田 智之 *¹
Tomoyuki Uchida

*¹広島市立大学情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

*²学習院大学計算機センター
Computer Centre, Gakushuin University

Machine learning and data mining from graph structured data are studied intensively. Many chemical compounds can be expressed by outerplanar graphs. We report a two-stage evolutionary learning method for acquiring characteristic multiple block preserving outerplanar graph patterns from positive and negative outerplanar graph data, by clustering outerplanar graphs.

1. はじめに

グラフ構造を持つデータからの機械学習やデータマイニングが注目されており、多くの化合物は外平面的グラフの構造を持つことが知られている [Horvath 10]. 外平面的グラフの構造的特徴を表現するために、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) が提案され、正事例の外平面的グラフの構造的特徴を表現する BPO グラフパターンを学習する手法が報告されている [Sasaki 08, Yamasaki 09].

我々は、正負の分類ラベルを持つグラフデータを分類する、グラフ分類問題 [Rehman 12] に取り組んでおり、正事例と負事例の外平面的グラフから、遺伝的プログラミングを用いて特徴的な BPO グラフパターンを獲得する進化的手法を実現している [Ouchiyama 15, Tokuhara 16a, Tokuhara 16b]. 遺伝的プログラミング (Genetic Programming, GP) [Koza 92, Banzhaf 98] とは、遺伝的アルゴリズムの遺伝子型を木構造のような構造的表現を扱えるように拡張した進化的手法である。さらに、一つの BPO グラフパターンで特徴をとらえることが難しいような外平面的グラフデータ集合の構造的特徴を表現するため、正事例と負事例の外平面的グラフから、遺伝的プログラミングを拡張した二段階構造の進化的獲得 [Nakai 14, Yamagata 17] による特徴的な複合的 BPO グラフパターンを獲得する手法を提案し、外平面的グラフの正事例集合のクラスタリング情報が与えられた入力データを対象にした実験結果を報告している [徳原 17a, Tokuhara 17]. ここで、複合的 BPO グラフパターンとは、BPO グラフパターン集合のことをいう。

本稿では、二段階構造の進化的獲得手法のために外平面的グラフの正事例集合をクラスタリングする手法を提案する。さらに、提案クラスタリング手法を用いて、正事例と負事例の外平面的グラフから特徴的な複合的 BPO グラフパターンを獲得する進化的手法の実験結果を報告する。

関連研究として、正事例と負事例からなる木構造データまたはグラフ構造データから特徴的な木パターンまたはグラフパターンを獲得する遺伝的プログラミングによる手法 [Nagamine 07, Nagai 12, Nakai 13, Miyahara 14], 及びグラフ構造データに対する進化的手法 [Katagiri 00, Shirakawa 07] がある。

2. 準備

2.1 BPO グラフパターン

BPO グラフパターンとブロック木パターン [Sasaki 08, Yamasaki 09] の説明をする [Tokuhara 16a, Tokuhara 16b, 徳原 17b, 徳原 17a, Tokuhara 17].

G をグラフ、 Λ と Δ をアルファベットとする。ラベル付きグラフとは、頂点集合 $V(G)$ と辺集合 $E(G)$ の各要素が、それぞれ Λ と Δ の要素によってラベル付けされたグラフをいう。ラベル付きグラフのすべての頂点が外平面に接するように平面埋め込みが可能であるとき、そのグラフを外平面的グラフとよぶ。連結グラフにおいて、削除することでグラフを非連結にすることができる頂点をカット点という。外平面的グラフのブロックとは、頂点数 3 以上のカット点をもたない極大な二重連結成分をいう。ブロックに属さない辺をブリッジとよぶ。BPO グラフパターンとは、ブリッジ変数および末端変数とよばれる 2 種類の構造的変数を持つ連結な外平面的グラフである。本稿では、連結な外平面的グラフのみを扱う。以後、連結な外平面的グラフを単に外平面的グラフとよぶ。外平面的グラフ G と BPO グラフパターン p に対し、 p の全ての変数を適当な外平面的グラフで置き換えることによって G が得られるとき p と G はマッチするという。外平面的グラフと BPO グラフパターンの例を図 1 に示す。図 1 において、BPO グラフパターン p のブリッジ変数 X , Y と末端変数 Z をそれぞれ外平面的グラフ g_1 , g_2 , g_3 に置き換えることで外平面的グラフ G を得ることができるので、 p と G はマッチする。

BPO グラフパターン p のブロック部分を、ブロックの辺ラベルの情報を保持したブロック頂点へと置き換えることで得られる、根なし無順序木の構造を持つグラフパターンを p のブロック木パターンとよび、 $t(p)$ で表す。ブロック木パターンの例を図 1 に示す。

2.2 遺伝的プログラミングによる複合的な BPO グラフパターンの獲得

本研究では、遺伝的プログラミングを拡張した二段階構造の進化的獲得手法により、特徴的な BPO グラフパターン集合 (複合的 BPO グラフパターン) を個体として獲得する。特徴的な BPO グラフパターン集合獲得手法では、以下の特徴的な BPO グラフパターン獲得問題を解く、BPO グラフパターンを個体とする遺伝的プログラミングによる手法 [Tokuhara 16a, Tokuhara 16b] を手続きとして使う。

連絡先: 宮原哲浩, 広島市立大学情報科学研究科, 〒731-3194, 広島市安佐南区大塚東 3-4-1, miyares18@hiroshima-cu.ac.jp

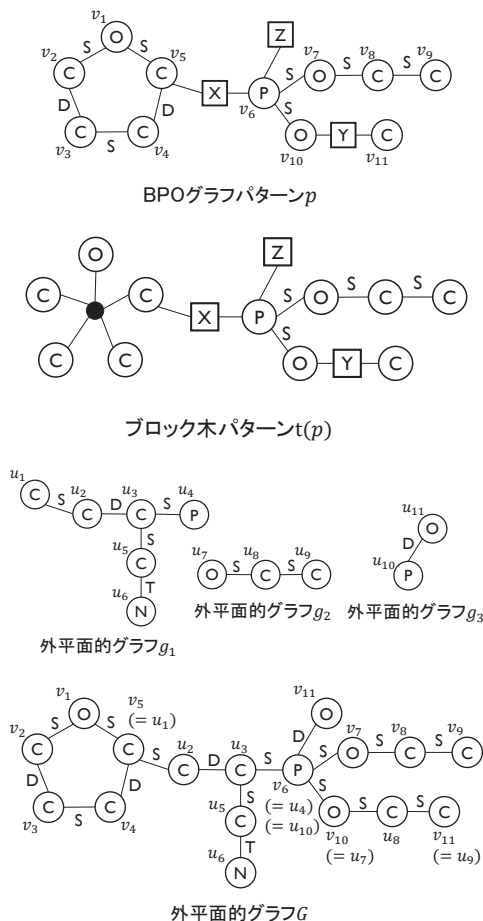


図 1: BPO グラフパターン p . p のブロック部分を変換して得られた p のブロック木パターン $t(p)$. 外平面的グラフ g_1, g_2, g_3, G . 頂点内の文字列は頂点ラベル, 辺のそばの文字列は辺ラベル, 正方形は変数を表すものとする. p と G はマッチする.

特徴的な BPO グラフパターン獲得問題

入力: 正事例及び負事例からなる外平面的グラフの有限集合 D .

問題: D に関する適合度の高い BPO グラフパターンを獲得する.

ここで, BPO グラフパターン p の D に関する適合度 $fitness_D(p)$ は, $fitness_D(p) = (p \text{ が } D \text{ の正事例にマッチする割合} + p \text{ が } D \text{ の負事例にマッチしない割合})/2$ で定義する. 遺伝的プログラミングの遺伝操作は BPO グラフパターン p のブロック木パターン $t(p)$ に適用する.

以下の特徴的な BPO グラフパターン集合獲得問題を解く, 特徴的な BPO グラフパターン集合を獲得するための進化的手法を提案している [Tokuhara 17, 徳原 17a].

特徴的な BPO グラフパターン集合獲得問題

入力: 正事例集合 P 及び負事例集合 N からなる外平面的グラフの有限集合 D , 正整数 $c (1 \leq c < |P|)$.

問題: D に関する適合度の高い BPO グラフパターン集合 $\Pi (1 \leq |\Pi| \leq c)$ を獲得する.

ここで, 集合 A の要素数を $|A|$ と表す. BPO グラフパターン集合 Π に含まれる BPO グラフパターンの少なくとも 1 つと外平面的グラフ G がマッチするとき, Π と G はマッチす

るという. BPO グラフパターン集合 Π の D に関する適合度 $fitness_D(\Pi)$ は, $fitness_D(\Pi) = (\Pi \text{ が } D \text{ の正事例にマッチする割合} + \Pi \text{ が } D \text{ の負事例にマッチしない割合})/2$ で定義する. よって, D に関する適合度の高い BPO グラフパターン集合とは, D の多くの正事例にマッチし, D の負事例にあまりマッチしないような特徴的な BPO グラフパターン集合であるといえる. BPO グラフパターン集合を獲得する進化的手法 (メインルーチン) は, 特徴的な BPO グラフパターンを獲得する遺伝的プログラミングをサブルーチンとして使う二段階の構造をしている. メインルーチンでは, 個体である BPO グラフパターン集合 $\{\pi_1, \pi_2, \dots, \pi_c\}$ を, BPO グラフパターン列 $[\pi_1, \pi_2, \dots, \pi_c]$ として扱う.

3. 外平面的グラフのクラスタリングによる複合的なブロック保存型外平面的グラフパターンの進化的獲得

特徴的な BPO グラフパターン集合獲得手法のための外平面的グラフの正事例データのクラスタリング手順の概要を以下に示す.

外平面的グラフの正事例データのクラスタリングの手順

1. ブロック木パターンの正規形表現を用いて, 外平面的グラフを根付き順序木へと変換する.
2. 根付き順序木へと変換した外平面的グラフからグラム分布で特徴分布の表を作成する.
3. 特徴選択を用いて, 正事例と負事例を分類するのに選択された特徴以外を特徴分布の表から削除する.
4. 正事例データの特徴ベクトルをカーネル PCA により, 指定した次元の空間の点に変換し, ユークリッド距離を用いて距離行列を計算する.
5. 距離行列に対して, ソフトクラスタリングを行う.

次にこの手順の詳細を説明する. まず, 外平面的グラフ間の距離を計算するために, グラム分布 (gram distribution) [Kuboyama 06] を用いる. グラム分布は木の指定した長さのパスと同型な部分木の出現頻度をベクトルにしたものである. グラム分布は頂点にのみラベルが付いた木を対象としているので, 出現部分木の辺ラベルは無視して同型判定を行う. またグラム分布は根付き順序木に対して計算を行うので, 外平面的グラフを根付き順序木へと変換する必要がある. 外平面的グラフをブロック木へと変換し, ブロック木パターンの正規形表現 [Tokuhara 16b, 徳原 17a] を用いて, ブロック木の正規形表現を求めることで根付き順序木へと変換することができる. 求めたブロック木の正規形表現に対して, パスの長さが 1 から 4 までのグラム分布を計算し特徴分布の表を作成する. 特徴的な BPO グラフパターン集合獲得手法では, 正事例と負事例を分類する特徴によって正事例が分類されている方がより高い適合度の個体を獲得できると考えられる. よって, CWC 特徴選択法 [Shin 15] を用いて, 正事例と負事例を分類するのに必要な特徴だけ選択し, 分類するのに必要のない特徴は特徴分布の表から削除する. 選択された特徴のみを残した特徴分布の表から求めた正事例データの特徴ベクトルを, カーネル PCA を用いて, 指定した次元 (d) の空間の点へと変換する. ユークリッド距離を用いて距離

行列を計算し、求めた距離行列に対してソフトクラスタリングを行う。特徴的な BPO グラフパターン集合獲得手法では、正事例が複数のクラスに属することを許すのでソフトクラスタリングを行い、正事例が閾値 (m) 以上の帰属度を持つクラスに属するものとする。

4. 実験

外平面的グラフのクラスタリングによる特徴的な複合的ブロック保存型外平面的グラフパターンの進化的獲得手法を、Intel Xeon CPU E5-2630 v2 2.60GHz プロセッサ、実装メモリ 32.0GB の Windows10 Pro 64bit OS 上に Java 言語で実装した。比較のため、特徴的な BPO グラフパターン獲得手法も同じ環境で実装した。

人工データ D を次のように生成して、実験データとした。設定した BPO グラフパターン集合 $\Pi = \{\pi_1, \pi_2, \pi_3\}$ について、人工的に生成した外平面的グラフのうち Π にマッチするものを正事例、マッチしないものを負事例として、500 個の正事例と 500 個の負事例からなる外平面的グラフの集合 D を作成した。

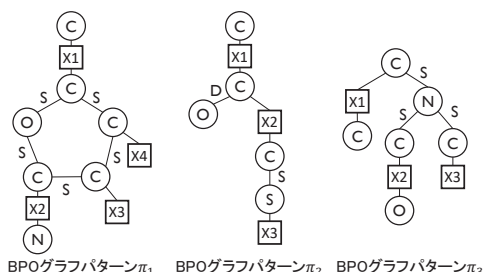


図 2: 人工データ作成のために設定した BPO グラフパターン集合 $\Pi = \{\pi_1, \pi_2, \pi_3\}$

外平面的グラフの正事例データのクラスタリング手順 4 において、変換後の正事例データ空間の次元を $d = 10$ とした。手順 5 のソフトクラスタリングを行った後に、正事例のクラスに属すると指定する帰属度の閾値を $m = 0.2$ とした。ソフトクラスタリングは fuzzy k-means を行う R 言語のパッケージである fclust[CRAN] を用いて行った。

fclust にはクラスタ数の妥当性を評価する 6 つの指標が用意されている。その指標で評価を行ったところ、クラスタ数 $c = 2, 3, 4, \dots$ と数を大きくするにつれ、全ての指標で妥当性の評価が低下した。よってクラスタ数 $c = 2, 3$ の 2 つの設定で実験を行った。

特徴的な BPO グラフパターン集合獲得手法 (メインルーチンの進化的手法) のパラメータを次のように設定した。BPO グラフパターンの上位数 (k):10($c = 2$ のとき),5($c \neq 2$ のとき), 集団サイズ (b):50, エリートサイズ (e):5, 最大世代数 (n):200, 加算適合度の最大値 (C_{add}):0.1。

特徴的な BPO グラフパターン獲得手法 (サブルーチン及び比較実験の遺伝的プログラミング) のパラメータを次のように設定した。集団サイズ (b'):50, エリートサイズ (e'):3, 最大世代数:200, トーナメントサイズ:2, 複製確率:0.05, 交叉確率:0.50, 突然変異確率:0.45。

特徴的な BPO グラフパターン獲得手法を 10 試行、特徴的な BPO グラフパターン集合獲得手法を正事例のクラスタリングのクラスタ数 c を $c = 2, 3$ として、それぞれ 10 試行の実験を行った。特徴的な BPO グラフパターン獲得手法、特徴的な

BPO グラフパターン集合獲得手法 (クラスタ数 $c = 2, 3$) の全 10 試行の実行時間の平均はそれぞれ、6843 秒、29341 秒、44980 秒であった。また特徴的な BPO グラフパターン獲得手法、特徴的な BPO グラフパターン集合獲得手法の各試行における最終世代の最良個体の適合度を表 1 に、各世代における最良個体の適合度の 10 試行の平均を図 3 に示す。最終世代の最良個体である BPO グラフパターン集合と BPO グラフパターンを図 4 に示す。

表 1: 各試行における最終世代の最良個体の適合度

試行	BPO グラフパターン獲得	BPO グラフパターン集合獲得	
		$c = 2$	$c = 3$
1	0.719	0.753	0.881
2	0.719	0.880	0.852
3	0.715	0.733	0.719
4	0.715	0.880	0.881
5	0.715	0.822	0.881
6	0.715	0.797	0.847
7	0.715	0.872	0.817
8	0.719	0.822	0.717
9	0.719	0.745	0.717
10	0.715	0.793	0.720
平均	0.717	0.810	0.803

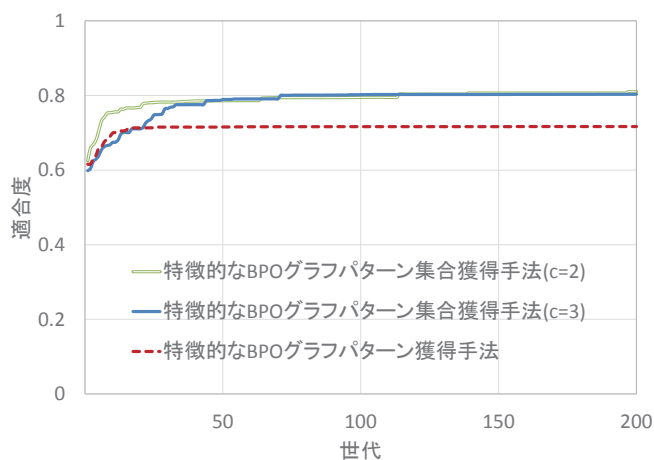


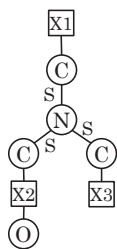
図 3: 各世代における最良個体の適合度の平均

5. おわりに

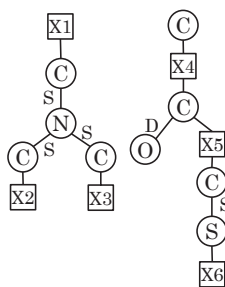
本稿では、二段階構造の進化的獲得手法のために外平面的グラフの正事例集合をクラスタリングする手法を提案し、正事例と負事例の外平面的グラフから特徴的な複合的 BPO グラフパターンを獲得する進化的手法を人工データに適用した実験結果を報告した。今後の課題として、計算時間の短縮などが挙げられる。

謝辞

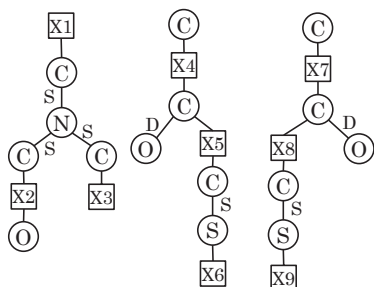
本研究は JSPS 科研費 JP15K00312, JP15K00313, JP17K00321 の助成を受けたものです。



特徴的なBPOグラフパターン
獲得手法の第1,8試行目の最終世
代最良個体
適合度0.719



特徴的なBPOグラフパターン集合
獲得手法(クラス数 $c=2$)の
第4試行目の最終世代最良個体
適合度0.880



特徴的なBPOグラフパターン集合
獲得手法(クラス数 $c=3$)の
第5試行目の最終世代最良個体
適合度0.881

図 4: 獲得した特徴的な BPO グラフパターン集合と特徴的な
BPO グラフパターン

参考文献

- [Banzhaf 98] W.Banzhaf et al.: *Genetic Programming: An Introduction : On the Automatic Evolution of Computer Programs and Its Applications*, Morgan Kaufmann (1998)
- [CRAN] CRAN: Package fclust,
<https://cran.r-project.org/web/packages/fclust/index.html>
- [Horvath 10] T.Horvath et al.: Frequent Subgraph Mining in Outerplanar Graphs, *Data Mining and Knowledge Discovery*, Vol. 21, pp. 472–508 (2010)
- [Katagiri 00] H.Katagiri et al.: Genetic Network Programming - Application to Intelligent Agents, *Proc. IEEE SMC 2000*, pp. 3829–3834 (2000)
- [Koza 92] J.R.Koza: *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press (1992)
- [Kuboyama 06] T.Kuboyama et al.: A Gram Distribution Kernel Applied to Glycan Classification and Motif Extraction, *Genome Informatics*, Vol. 17(2), pp. 25–34 (2006)
- [Miyahara 14] T.Miyahara and T.Kuboyama: Learning of Glycan Motifs Using Genetic Programming and Various Fitness Functions, *JACIII*, Vol. 18(3), pp. 401–408 (2014)
- [Nagai 12] S.Nagai et al.: Acquisition of Characteristic TTSP Graph Patterns by Genetic Programming, *Proc. IIAI AAI 2012*, pp. 340–344 (2012)
- [Nagamine 07] M.Nagamine et al.: A Genetic Programming Approach to Extraction of Glycan Motifs using Tree Structured Patterns, *Proc. AI-2007, Springer-Verlag, Vol. 4830*, pp. 150–159 (2007)
- [Nakai 13] S.Nakai et al.: Acquisition of Characteristic Tree Patterns with VLDC's by Genetic Programming and Edit Distance, *Proc. IIAI AAI 2013*, pp. 147–151 (2013)
- [Nakai 14] S.Nakai et al.: Acquisition of Characteristic Sets of Tree Patterns with VLDC's using Genetic Programming and Edit Distance, *Proc. IWCIA 2014*, pp. 113–118 (2014)
- [Ouchiyama 15] Y.Ouchiyama et al.: Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns from Positive and Negative Data using Genetic Programming and Tree Representation of Graph Patterns, *Proc. IWCIA 2015*, pp. 95–101 (2015)
- [Rehman 12] S.Rehman et al.: Graph Mining: A Survey of Graph Mining Techniques, *Proc. ICDIM 2012*, pp. 88–92 (2012)
- [Sasaki 08] Y.Sasaki et al.: Mining of Frequent Block Preserving Outerplanar Graph Structured Patterns, *Proc. ILP-2007, Springer-Verlag LNAI 4894*, pp. 239–253 (2008)
- [Shin 15] K.Shin et al.: Super-CWC and Super-LCC: Super Fast Feature Selection Algorithms, *Proc. IEEE Big Data 2015*, pp. 61–67 (2015)
- [Shirakawa 07] S.Shirakawa et al.: Graph Structured Program Evolution, *Proc. GECCO 2007*, pp. 1686–1693 (2007)
- [Tokuhara 16a] F.Tokuhara et al.: Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns by Genetic Programming using Label Information, *Proc. IIAI AAI 2016*, pp. 203–210 (2016)
- [Tokuhara 16b] F.Tokuhara et al.: Using Canonical Representations of Block Tree Patterns in Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns, *Proc. IWCIA 2016*, pp. 93–99 (2016)
- [Tokuhara 17] F.Tokuhara et al.: Acquisition of Multiple Block Preserving Outerplanar Graph Patterns by an Evolutionary Method for Graph Pattern Sets, *Proc. IWCIA 2017*, pp. 191–197 (2017)
- [Yamagata 17] Y.Yamagata et al.: Acquisition of Multiple Graph Structured Patterns by an Evolutionary Method using Sets of TTSP Graph Patterns as Individuals, *Proc. IIAI AAI 2017* (2017)
- [Yamasaki 09] H.Yamasaki et al.: Learning Block-Preserving Graph Patterns and Its Application to Data Mining, *Machine Learning*, Vol. 76, pp. 137–173 (2009)
- [徳原 17a] 徳原史也ほか: 遺伝的プログラミングによる複合的なブロック保存型外平面的グラフパターンの獲得, 2017 年度人工知能学会全国大会論文集, 4H2-2 (2017)
- [徳原 17b] 徳原史也ほか: 特徴的なブロック保存型外平面的グラフパターンの獲得におけるブロック木パターンの深さラベル列の利用, 2017 年度人工知能学会全国大会論文集, 4E1-2 (2017)