化合物データからの特徴的な外平面的グラフパターン発見アルゴリズムの実装とGUI開発

延 拓哉[1]　鈴木 祐介[1,a)]　内田 智之[1,b)]　宮原 哲浩[1,c)]

概要：化合物データの多くは外平面的グラフの構造を持つデータとみることができる．知識発見の分野では，グラフ構造を持つデータに共通する特徴的なパターンを発見する手法の開発が行われている．ブロック保存型外平面的グラフパターン (BPO グラフパターン) とは，外平面的グラフに構造的変数を導入したグラフパターンであり，外平面的グラフに共通するグラフ構造を表現することができる．外平面的グラフ構造を持つ化合物データに対する，特徴的な BPO グラフパターン発見アルゴリズムの実装と GUI 開発を行った．

キーワード：グラフアルゴリズム，機械学習，データマイニング

Implementation of Mining Algorithms for Extracting Characteristic Outerplanar Graph Patterns for Chemical Dataset

Takuya Nobu[1]　Yusuke Suzuki[1,a)]　Tomoyuki Uchida[1,b)]　Tetsuhiro Miyahara[1,c)]

Abstract: Many chemical compounds can be expressed by outerplanar graphs. Block preserving outerplanar graph patterns are graph structured patterns having structured variables and are suited to represent characteristic graph structures common to chemical dataset. We implemented the mining algorithms for extracting characteristic block preserving outerplanar graph patterns and developed a GUI of their graph mining systems.

Keywords: graph algorithm, machine learning, data mining

1. はじめに

近年グラフ構造を持つデータの量が増大しており，知識発見の分野では，グラフ構造を持つデータに共通する特徴的なパターンを発見する手法の開発が行われている．化学化合物の多くは外平面的グラフの構造を持つデータとみることができる．外平面的グラフとは，平面的グラフのうちすべての頂点が外平面に接するような平面埋め込みを持つもののことを指す．図 1 に示したグラフ G_1, G_2, G_3 はいずれも外平面的グラフである，本論文では，外平面的グラフの構造を持つ化合物データから構造的特徴を発見するパターン発見アルゴリズムの実装とその GUI の開発を行う．

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)[7] とは，外平面的グラフに構造的変数の概念を導入したグラフパターンであり，外平面的グラフに共通するグラフ構造を表現することができる．BPO グラフパターンはブリッジ変数および末端変数とよばれる 2 種類の構造変数を持っており，変数は任意の連結な外平面的グラフで置き換えることが可能である．BPO グラフパターン g の全ての変数を適切な連結外平面的グラフで置き換えることによって，外平面的グラフ G が得られるとき，g は G にマッチするという．図 1 に BPO グラフパターン，および BPO グラ

[1] 広島市立大学情報科学部
Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University, Japan
[a)] y-suzuki@hiroshima-cu.ac.jp
[b)] uchida@hiroshima-cu.ac.jp
[c)] miyares16@info.hiroshima-cu.ac.jp

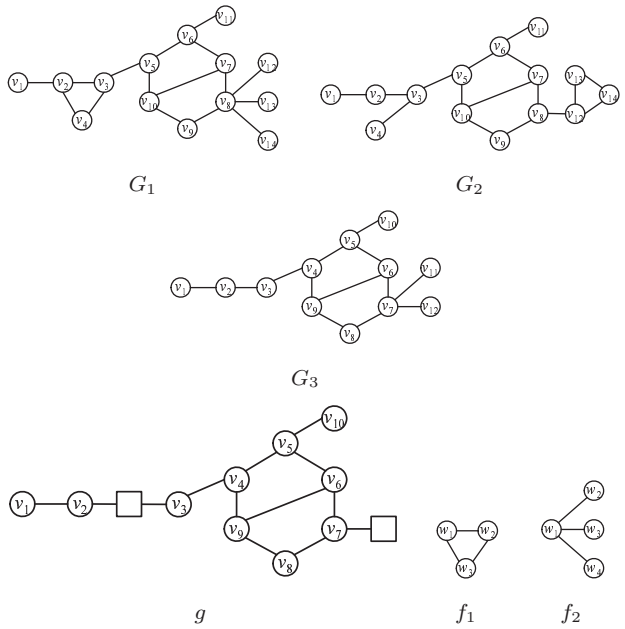

図 1 BPO グラフパターン g と外平面的グラフ G_1, G_2, G_3, f_1, f_2. BPO グラフパターンの変数は，直線と四角によって表される．BPO グラフパターン g は変数 (v_2, v_3) と変数 (v_7) を持つ．

フパターンにマッチする外平面的グラフの例を示す．BPO グラフパターン g はブリッジ変数 (v_2, v_3) と末端変数 (v_7) を持ち，それぞれを外平面的グラフ f_1, f_2 で置き換えることにより外平面的グラフ G_1 を得ることができる．ブリッジ変数および末端変数に任意の連結外平面的グラフを代入しても，代入後のグラフは外平面的グラフのままである．外平面的グラフの集合 S に対し，S を説明する極小一般化 BPO グラフパターンとは，与えられた外平面的グラフの集合 S 中の全ての要素にマッチする BPO グラフパターンのうちで，BPO グラフパターンのマッチするグラフ集合が極小であるものを言う．

山崎ら [7] は，与えられた BPO グラフパターンと外平面的グラフがマッチする否か判定するマッチングアルゴリズムを提案した．また，与えられた外平面的グラフ集合に対し，それを説明する極小一般化 BPO グラフパターンを発見するパターン発見アルゴリズムの提案を行った．本研究では，これら二つのアルゴリズムを計算機上に実装する．さらに，これらの二つのアルゴリズムを用いて，外平面的グラフの構造を持つ化合物データに対する，特徴的な BPO グラフパターン発見システムの GUI 開発を行う．

関連研究として，Horváth ら [2] は外平面的グラフに対する頻出部分グラフマイニングアルゴリズムを提案した．徳原ら [6] は遺伝的プログラミングを用いて，正事例と負事例の外平面的グラフから特徴的な BPO グラフパターンを発見する手法について研究を行った．山崎ら [7] の提案した，BPO グラフパターンと外平面的グラフに対するマッチングアルゴリズムは，グラフをブロック木という木構造データに変換してマッチするか否かを判定する．木構造データに共通する構造を表現する項木パターンに対しても，マッチングアルゴリズムや特徴的なパターンの発見アルゴリズムが提案されている [4], [5].

2. 準備

2.1 ブロック保存型外平面的グラフパターン

ラベル付きグラフとは，頂点集合と辺集合の各要素がラベル付けされた連結グラフをいう．連結グラフの切断点とは，その頂点を取り除くとグラフが非連結となる頂点のことである．連結グラフのブリッジとは，その辺を取り除くとグラフが非連結となる辺のことである．連結グラフのブロックとは，そのグラフの，頂点数 3 以上の切断点を持たない極大な連結部分グラフのことである．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．

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)[7] とは，ブリッジ変数および末端変数とよばれる 2 種類の構造変数を持つ外平面的グラフのことである．ブリッジ変数は，2 頂点からなる変数で，変数を辺とみなすと連結グラフのブリッジとなるような変数のことである．末端変数は，1 頂点からなる変数のことである．BPO グラフパターンの変数は，任意の外平面的グラフで置き換えることが可能である．ブリッジ変数および末端変数に，任意の外平面的グラフを代入しても，代入後のグラフは外平面的グラフのままである．

外平面的グラフ G と BPO グラフパターン g に対し，g の全ての変数を適切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．g と G はマッチするという．BPO グラフパターン g の言語を $L(g) = \{G \mid g$ と外平面的グラフ G がマッチする $\}$ と定義する．外平面的グラフの集合 S が与えられたとき，S を説明する極小一般化 BPO グラフパターンとは，与えられた外平面的グラフの集合 S の全ての要素にマッチする BPO グラフパターンのうち，その言語が極小であるものをいう．BPO グラフパターンと，そのマッチの例を図 1 に示す．BPO グラフパターン g は，外平面的グラフ G_1, G_2, G_3 にマッチする．

2.2 NCI データセット

NCI(米国国立がん研究所) データベース [3] には，公開時期や付加情報の違いなどによって数種類の分子化合物データ集合が公開されている．各分子化合物データ集合には，それぞれ数万の分子化合物の原子や結合情報が記載されている．

NCI データベースに記載されている分子化合物データ集合について説明する．分子化合物データ集合には数万個の分子化合物データが記載されており，各分子化合物データは区切り記号 “$$$$” によって区分されている．

```
graph GraphSample{
node[style="setlinewidth(3)"];
edge[style="setlinewidth(3)"];
V0 [label="v0:O"]
V1 [label="v1:C"]
V2 [label="v2:N"]
V3 [label="v3:N"]
V4 [label="v4:C"]
V5 [label="v5:C"]
V6 [label="v6:C"]
V7 [label="v7:O"]
V8 [label="v8:C"]
V9 [label="v9:C"]
V10 [label="v10:O"]
V11 [label="v11:C"]
V12 [label="v12:C"]
V13 [label="v13:C"]
V14 [label="v14:C"]
V15 [label="v15:O"]
V0 -- V1 [label="2"]
V1 -- V2 [label="1"]
V1 -- V3 [label="1"]
V2 -- V4 [label="1"]
V2 -- V5 [label="1"]
V3 -- V6 [label="1"]
V4 -- V7 [label="1"]
V4 -- V8 [label="1"]
V5 -- V9 [label="2"]
V6 -- V10 [label="2"]
V7 -- V11 [label="1"]
V8 -- V12 [label="2"]
V9 -- V13 [label="1"]
V11 -- V14 [label="1"]
V14 -- V15 [label="1"]
V6 -- V9 [label="1"]
V11 -- V12 [label="1"]
}
```

dot 言語ファイル例

graphviz による可視化例

図 2 NCI データセットの化合物データを，グラフとして表現したものが左の dot 言語ファイルである．また左の dot 言語ファイルを graphviz で可視化したものが右のグラフである．

分子化合物データは 3 行目に CAS 番号 (CAS-RN)，4 行目に分子数と結合の数が記載されている．5 行目から化合物中の原子について記載されている．それらの基本情報ののちに付加情報が記載されている．基本情報のフォーマットは NCI データベースの各分子化合物データ集合に共通である．付加情報は，AIDS 抗ウィルス活性化試験の結果やがん細胞増殖抑制試験の結果などであり，これらの付加情報の記載の有無は分子化合物データ集合により異なる．

本研究では，分子化合物データ集合から区切り記号 “$$$$” に従って，分子化合物データを切り出し，1 つの化合物を 1 つのテキストファイルに保存して使用している．

2.3 dot 言語

dot 言語は，ある文法によって規定されるグラフのデータ構造を表現するための言語である．dot 言語で記述されたファイルを dot 言語ファイルと呼び，拡張子として.dot が用いられる．dot 言語では頂点や辺の持つラベルなどを記述することができる．

dot 言語で記述されたグラフを可視化するためのソフトウェアとして Graphviz[1] がある．Graphviz は dot 言語ファイルを入力として，記述されているグラフのデータ構造を描画し，JPG や PNG などの画像ファイルとして出力することができる．図 2 に，dot 言語ファイルと Graphviz によって可視化されたグラフの例を示す．

3. 特徴的な外平面的グラフパターン発見システム

3.1 特徴的な外平面的グラフパターン発見問題

外平面的グラフに対するマッチング問題と，特徴的な外平面的グラフパターン発見問題を次のように定義する．

外平面的グラフに対するマッチング問題

入力： 外平面的グラフ G，BPO グラフパターン g.

問題： g と G がマッチするか否か判断する．

特徴的な外平面的グラフパターン発見問題

入力： 外平面的グラフの集合 S.

問題： S を説明する極小一般化 BPO グラフパターンを発見する．

外平面的グラフに対するマッチング問題は，BPO グラフパターンに対する所属性問題として，特徴的な外平面的グラフパターン発見問題は，BPO グラフパターンに対する極小言語問題として，それぞれ知られている．図 1 に特徴的な外平面的グラフパターン発見問題の例を示す．入力として外平面的グラフの集合 $S = \{G_1, G_2, G_3\}$ が与えられたとき，g は S を説明する極小一般化 BPO グラフパターンである．山崎ら [7] は，BPO グラフパターンに対する所属性問題を多項式時間で解くマッチングアルゴリズム，及び BPO グラフパターンの極小言語問題を多項式時間で解く MINL アルゴリズムを提案した．本研究では，マッチングアルゴリズムと MINL アルゴリズムを，JAVA を用いて計算機上に実装した．

3.2 化合物データに対する特徴的な BPO グラフパターン発見システムの GUI

本研究では，実装したマッチングアルゴリズムと MINL アルゴリズムを用いて，外平面的グラフの構造を持つ化合物データに対する特徴的な BPO グラフパターン発見システムの GUI 開発を行った．GUI は視覚的にわかりやすく，主にマウスで扱えるよう設計した．GUI では，グラフデータを graphviz を用いて画像ファイルに変換し表示を行う．図 3 に開発した GUI の起動時と実行時画面を示す．

このシステムは 3 つの機能を有する．1 つ目は，複数の外平面的グラフを表すファイルを入力し，それを Graphviz を用いて画像ファイルに変換し、視覚的に表示する．ここで入力ファイルは，グラフを dot 言語方式で記述した dot 言語ファイル (.dot)，または NCI データセットのファイルを区切り記号 “$$$$” に従って切り出したテキストファイル (.txt) を使用することができる．2 つ目は，選択した複数の外平面的グラフを入力として，MINL アルゴリズムを実行し，入力グラフを説明する極小一般化 BPO グラフパターンを計算する．さらに，その極小一般化 BPO グラフパターンを Graphviz を用いて画像ファイルに変換し，視

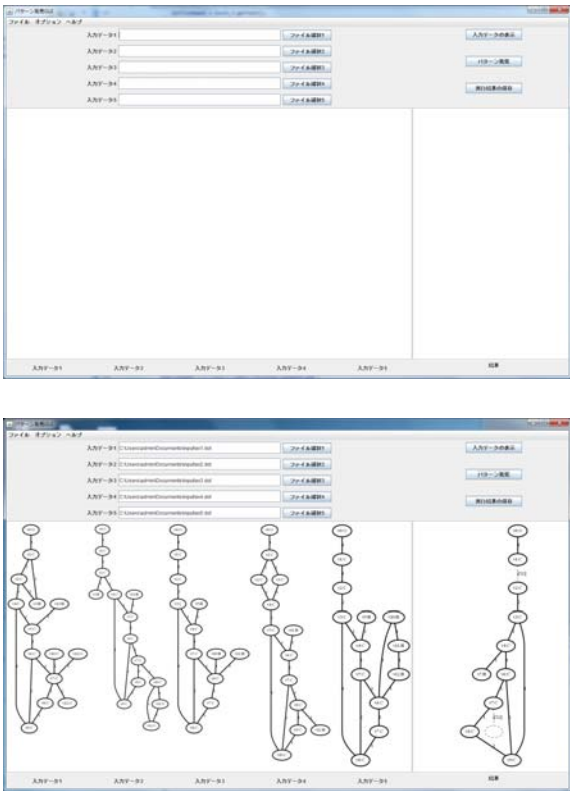

図 3 開発した GUI の起動時と実行時の画面．上部は GUI 起動時の画面である．下部は，5 つの dot 言語ファイルを入力として与え，入力グラフを説明する極小一般化 BPO グラフパターンを計算した結果を表示したものである．

覚的に表示する．3 つ目は，出力した極小一般化 BPO グラフパターンを dot 言語ファイルとして保存する．

補助的な機能として，実験を最初からやり直す開始機能，実験を終了する終了機能，入力ファイルの形式を切り替える機能，ヘルプ機能を実装している．これらの機能を備えた GUI を JAVA により実装した．

4. おわりに

本研究では，山崎ら [7] が提案した，BPO グラフパターンに対するマッチング問題を多項式時間で解くマッチングアルゴリズム，及び特徴的な外平面的グラフパターン発見問題を多項式時間で解く MINL アルゴリズムを計算機上に実装した．さらに，これらの二つのアルゴリズムを用いて，外平面的グラフの構造を持つ化合物データに対する特徴的な BPO グラフパターン発見システムの GUI 開発を行った．

今後の課題としては，入力ファイル数をユーザーが任意に変更できる機能が必要と思われる．また，より視覚的に見やすくするために，画面のサイズに合わせて，GUI 各部のサイズを変更するような機能や，実験結果の保存の際に dot 言語ファイルだけでなく，出力した画像データも保存する機能の追加が考えられる．

本研究の発展として，実装したマッチングアルゴリズムと MINL アルゴリズムの改良によるパターン発見の高速化があげられる．また開発した GUI を改良し，化合物データからのデータマイニングツールの作成などが考えられる．

参考文献

[1] Graphviz http//www.graphviz.org/
[2] T. Horváth, J. Roman, and S. Wrobel, Frequent subgraph mining in outerplanar graphs. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 197–206, 2006.
[3] National Cancer Institute Chemical Dataset, http//cactus.nci.nih.gov/
[4] Y. Suzuki, T. Shoudai, T. Uchida,T. Miyahara, Ordered Term Tree Languages Which Are Polynomial Time Inductively Inferable from Positive Data, Theoretical Computer Science, Vol. 350(1), pp. 63–90, 2006.
[5] Y. Suzuki, T. Shoudai, T. Uchida,T. Miyahara, An Efficient Pattern Matching Algorithm for Ordered Term Tree Patterns, IEICE Trans. Inf. Syst., Vol. E98-A(6), pp. 1197–1211, 2015.
[6] F. Tokuhara, T. Miyahara, Y. Suzuki, T. Uchida, T. Kuboyama, Using Canonical Representations of Block Tree Patterns in Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns, Proceedings of the 9th IEEE International Workshop on Computational Intelligence and Applications (IWCIA 2016) pp. 93–99, 2016.
[7] H. Yamasaki, Y. Sasaki, T. Shoudai, T. Uchida, Y. Suzuki, Learning Block-Preserving Graph Patterns and Its Application to Data Mining, Machine Learning, Vol.76 No.1, pp.137-173, 2009.