化学物質の部分構造とその包含関係からの知識発見

橋本 桂　　津田 哲夫　　吉澤 有美　　稲積 宏誠
青山学院大学理工学部

１．はじめに

対象とする問題をグラフ表現し,その各要素間の関係構造の中から有効な知識を直接的に抽出しようとするアプローチとして,グラフマイニングがある.その手法の一つとして,任意の条件のもとに逐次的に部分構造を生成することによって意味のある構造を抽出しようとする手法である Graph Based Induction(GBI 法)[1] が提案されている.

本稿では化学物質データの構造情報および非構造情報を用いてその物質の性質を決定するルールを発見することを目的とし,GBI 法の考え方に基づいて,化学的性質に応じた評価基準による共通部分構造の抽出アルゴリズムを検討する.さらに抽出された部分構造間の関係を用いた知識発見を検討する.

２．GBI 法を用いた特徴的パターン抽出

GBI 法は，グラフ中に現れるペア情報を逐次拡張することによって特徴的なパターンを発見するための手法である．GBI 法では,入力グラフに対して下記 3 ステップを繰り返すことでペアを逐次拡張し，グラフ構造データに含まれる特徴を類型パターンとして抽出する.

Step 1．グラフ中に存在する二つのノードの組み合わせからなるすべてのペアを抽出する.

Step 2．抽出したペアのうち,ある評価指標によりチャンクすべきペアを一つ選び,抽出パターンとして登録する.このとき,ペアを構成するノードが Step 1．で書き換えられたノードであれば,もとのパターンに復元して登録する.

Step3.Step2 で選ばれたペアを一つのノードに書き換えることにより,グラフを書き換える.

ただし，STEP 2．でペアを抽出する際，一つのノードに書き換えられたパターンをもとのパターンに復元しての評価は行わない．すなわち,GBI 法は逐次的な探索・チャンクを行う Greedy 探索手法であり，入力グラフデータ中に存在するすべての「類型的なパターン」を抽出できるわけではない．そのため,チャンクするペアの選択基準が抽出パターンに大きく影響を及ぼす.従来の GBI 法では頻度に基づく評価基準によってペアを選択し,頻出パターンのみを抽出する.一方,本稿ではクラス分類能力を有するパターンを抽出するよう,評価指標を(1)式のように定義する.

$$\underbrace{\left(\frac{N_h}{N_H}+\frac{N_l}{N_L}\right)}_{①}+\alpha\times\underbrace{\max\left\{\frac{N_h}{N_H},\frac{N_l}{N_L}\right\}\Big/\left(\frac{N_h}{N_H}+\frac{N_l}{N_L}\right)}_{②} \quad \cdots(1)$$

Nh,Nl： パターンを含むクラスごとの物質数
NH,NL： 入力データ中のクラスごとの物質数
$0 \leqq \alpha < 1$

(1)式は次のような考え方に基づいている.対象データから 2 つのクラス(high，low)へのクラス分類能力を有するパターンを抽出するには，

1)high，low いずれかの特徴を示すパターン (higher，lower)

2) 1)をチャンクする際に基盤となるパターン (common)

をチャンクする必要がある.1)の抽出を促すためには各クラスの支持度の最大値((1)－②)を用い,2)の抽出を促すにはクラス別物質数を考慮したパターン含有率（(1)－①）に注目する.一般的に 1)は 2)が抽出された後に多く抽出されることが予想される.そのため,(1)－②はチャンクの前半では重く考慮する必要はない.逆に,2)が十分に抽出された後は,(1)－①よりも(1)－②の評価を重視したい.そこで,この(1)－①に重みαを掛けた(1)－②を足し合わせたものをチャンク時のペア選択基準として用いた.(1)－①は評価対象パターンが単純な構造であれば大きな値を示し,(1)－②の評価による影響を抑えるが,対象パターンが複雑になるにつれて値が下がり,それに伴い評価指標全体に対する影響力も低くなる.したがって(1)式は,対象パターンが単純な構造であるチャンク前半では(1)－①の評価による 1)の抽出を,対象パターンが単純な構造となるチャンク後半では(1)－②の評価による 2)の抽出を狙う

Knowledge discovery from substructures and the inclusive relations of molecules
† Katsura HASHIMOTO, Tetsuo TSUDA, Yumi YOSHIZAWA, Hiroshige INAZUMI
School of Science and Engineering, Aoyama Gakuin University

ことができると考えられる.

さらに,抽出され たパターンに対して,以下のようにそのクラス分類能力を評価し,high, low, common の３クラスに分類する.

$$common:\ \max\left\{\frac{N_h}{N_H},\frac{N_l}{N_L}\right\}\Big/\left(\frac{N_h}{N_H}+\frac{N_l}{N_L}\right)\le n$$

$$higher:\ \frac{N_h}{N_H}\Big/\left(\frac{N_h}{N_H}+\frac{N_l}{N_L}\right)> n$$

$$lower:\ \frac{N_l}{N_L}\Big/\left(\frac{N_h}{N_H}+\frac{N_l}{N_L}\right)> n$$

$$0.5< n \leqq 1$$

このようにして抽出された部分構造の包含関係にもとづく情報からクラスの変化点を抽出することができる.

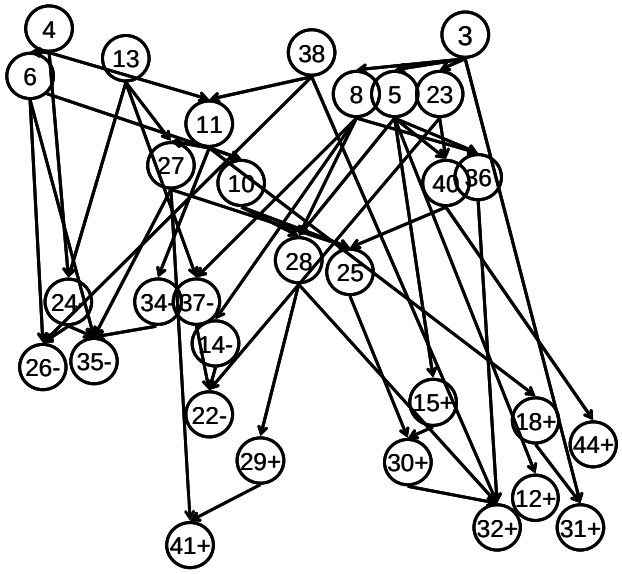

図 1. 部分構造間の含有関係

３．抗菌活性物質への適用

抗菌活性物質であるフラボノイド 325 種類を対象として前述のアルゴリズムを用いた実験を行い,44 個の部分構造を抽出した.その際,各物質の抗菌活性値を以下のように離散化し,これをクラスとして用いた.

Low ：activity ＞32
High ：activity ≦32

図１は抽出されたパターンをパターン間の包含関係を元にグラフ化し,異なるクラスへ隣接しているノード以外を除外したものである.各ノードの数字は各部分構造に割り当てられたラベルである.また,図１においてクラスの変化している隣接ノードを比較して図 2 にまとめた.これによりクラスが High となる特徴を推定することができる.例えば,34→41,28→41 から,28 や 34 で示されているような部分構造は単独では活性度に対して意味のない構造であるにもかかわらず,41 のような組み合わせになると活性度を高める結果になることがわかる.

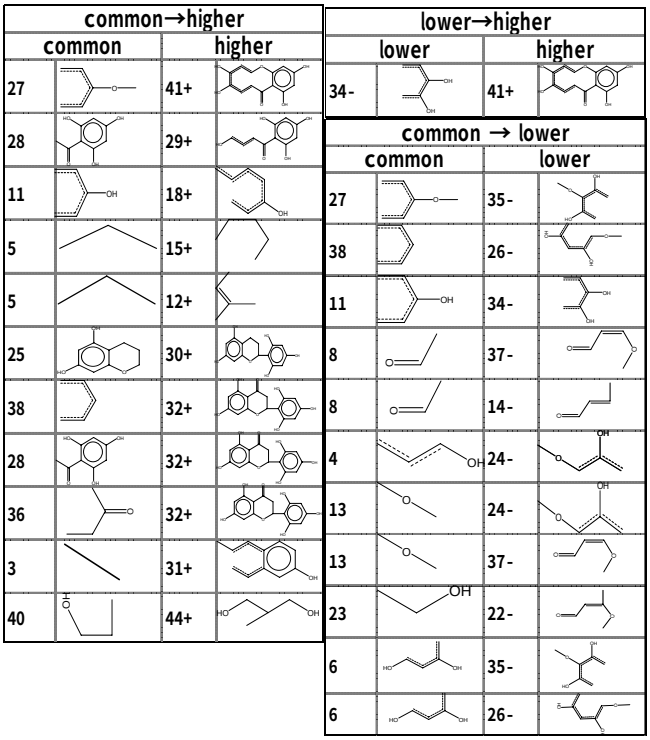

図 2　活性度の変化

４．結論

本稿では,クラス分類能力を持つ部分構造抽出を実現した.さらに,抽出された部分構造の関係情報が有効な知識発見につながる可能性を示した.今後の課題として,GBI 法そのものの改良,及びクラス分類能力判定の精度向上に加えて,関係構造の活用方法を洗練化することが挙げられる.

参考文献

[1]松田喬,元田浩,鷲尾隆：”一般グラフ構造データに対する Graph-Based Induction とその応用”,人工知能学会論文誌,Vol.16, No.4 , pp.363-374 (2001)