1A2-02

部分構造情報を用いたグラフクラスタリング手法の検討

A Step towards New Graph Clusterring Algorithm based on Substructure Distribution

和田　貴久[*1]　大野　博之[*2]　稲積　宏誠[*2]
Takahisa Wada　Hiroyuki Oono　Hiroshige Inazumi

[*1]青山学院大学大学院　理工学研究科　理工学専攻　知能情報コース
Graduate school of Science and Engineering, Aoyama Gakuin University

[*2]青山学院大学　理工学部　情報テクノロジー学科
College of Science and Engineering, Aoyama Gakuin University

This paper presents new graph clusterring algorithm based on substructure distribution. The graph structural data expressed by the node and the link is targeted. A partial structure is extracted by applying the Chunkingless Graph-Based Induction (Cl-GBI) method. The relation between the partial structure and each graph is expressed by the matrix. The similarity between arbitrary graphs is calculated by using the matrix. We applide hierarchical clustering method to divide instances into groups based on similarity.

1. はじめに

近年，Web 上にはテキストデータや時系列データ，グラフデータなど，さまざまな形式のデータが蓄積され，それらを活用しようと，多くのマイニング手法が研究されている．しかし，従来の解析対象の多くはテキストデータや時系列データなどのデータそのものであり，構造情報を含むグラフデータに対するマイニング手法の研究は比較的新しい分野といえる．本来，グラフ構造は汎用的なデータ構造である．本稿においても一般的なグラフ構造データに適用できる手法の開発を目的とするが，最も典型的なグラフ構造データとして化学物質を取り上げる．化学の分野では，化学物質の特性を分析する際に電荷情報などの物理化学的な情報を利用することが多く，構造情報の有効活用は必ずしも実現されていない．

本稿では，構造情報の有効活用を実現するための取り組みとして，グラフ構造データから Chunkingless Graph-Based Induction（Cl-GBI）法 [高林 05] を用いて部分構造の抽出を行い，それを用いた類似度の計算方法とクラスタリング手法を提案する．さらに，他のグラフクラスタリング手法と特性の比較検討を行い，その応用について展望する．

2. 部分構造情報を用いたクラスタリング

2.1 Cl-GBI 法を用いた部分構造抽出

ノードとリンクで表現されるグラフ構造データは，Cl-GBI 法を適用することによって，高い頻度で出現する特徴的な部分構造を抽出することができる．Cl-GBI 法は，ノードペアを逐次抽出・逐次チャンクすることで部分構造を抽出するが，もとのノード情報を保持するので，部分的に重なる部分構造などのすべての部分構造を抽出することができる．また，ビーム幅やチャンク条件を工夫することによって，非常に多くの部分構造が比較的低コストで抽出可能である．ただし，これらの部分構造の中には，その部分構造を含むすべてのグラフに同時に含まれていて，かつ包含する部分構造が存在する場合がある．これを冗長な構造とし，本提案手法ではこのような冗長な部分構造は抽出された部分構造から除去することとする．

連絡先: 和田 貴久，青山学院大学大学院　理工学研究科
〒 229-8558 神奈川県相模原市淵野辺 5-10-1
t-wada@ina-lab.it.aoyama.ac.jp

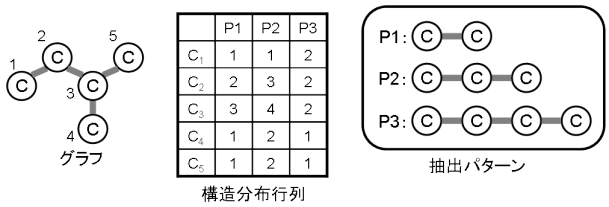

図 1: グラフと部分構造の関係を表す行列

2.2 部分構造を用いたグラフ間の類似度

各グラフを Cl-GBI 法により抽出された部分構造をもとに部分構造とノードとの関係を表し，計算可能な行列を生成する．グラフ内に存在するノード N はノードラベルとノード ID で表現され $N = \{x_i | x \in \{\mathrm{C,O,N,}\cdots\}, i = 1,2,\cdots\}$ とする．また，リンクラベルも考慮して，抽出された部分構造を $P = \{p_1, p_2, \cdots, p_j\}$ とする．部分構造 p_j に含まれるノード x_i の個数を y_{ij} とし，y_{ij} を要素として持つ行列を定義し，これを各グラフの構造分布行列と呼ぶ．3 つの部分構造 p_1, p_2, p_3 を持つグラフの行列表現例を図 1 に示す．

次に各グラフの構造分布行列を用いてグラフ間の類似度を定義する．まず，各部分構造ごとに，行列の各要素の一致度，及び不一致数に注目する．しかし，比較するグラフ内のノードラベルの種類数や各ラベルを持つノードの個数が必ずしも同じとは限らないので，その差異を考慮した計算方法が必要となる．また，部分構造によって一致していたときに与える重みを設定する．例えば，重みを部分構造のサイズにすれば，大きなサイズの部分構造を共有するほど類似度に大きく貢献することになる．また，特別な性質を有する部分構造の重みを高くすれば，その性質を反映させることができる．以上を考慮した類似度算出の手順を以下に示し，グラフ間の類似度の算出例を図 2 に示す．

1) 比較する 2 つのグラフから同じラベルを持つノードをそれぞれ一つずつ用意する．

2) 構造分布行列を利用し，2 つのノードの一致度 (C)，不一致数 (E) を計算する．ただし，

- **a)** 各部分構造毎に二つのノードに含まれている個数の最小値にその部分構造に含まれるノード数を重みとしてかけ，その総和を一致度とする．

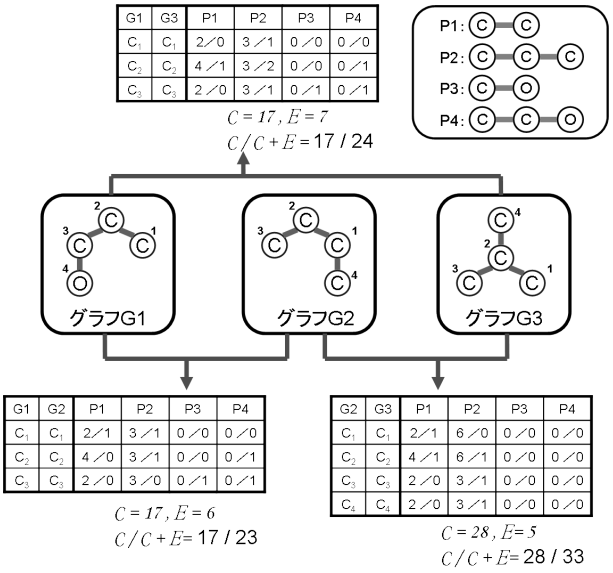

図 2: グラフ間の類似度

b) 2 つのノードにおいて，各部分構造の数の差を計算し，その総和を不一致数とする．

3) 同じノードラベルを持つ他のすべてのノードのペアについて 2) の処理を行う．

4) 計算された各ノードペアの中で，一致度の割合が最も大きなペアのノードを各のグラフより取り除き各数値を保存する．

5) 同じラベルを持つノードのペアが存在しなくなるまで 1)〜4) の処理を繰り返す．

6) 保存されている一致度と不一致数のそれぞれの総数から一致度の割合を求め，これを類似度とする．

このようにして，求められた全てのグラフ間の類似度を用いて，最短距離法による階層的クラスタリングを行う．最も類似度の高いものから順に逐次的にクラスタを形成していき，最終的に 1 つのクラスタになるまで処理を行い，デンドログラムを作成する．作成されたデンドログラムに対して閾値を設定し，複数のクラスタに分割する．

3. 実験

本稿で提案したクラスタリング手法の特性を検討するために，フラボノイド類 64 種類と、その他の 41 種類のあわせて 105 種類の物質を用いた．まず部分構造の抽出には，繰り返し数を 20，ビーム幅を 10，閾値とする含有率を 3 %とし，Cl-GBI 法を適用し，冗長な部分構造を除去し，1372 種類の部分構造を抽出した．次に抽出した部分構造から構造分布行列を生成し，類似度を求め，類似度の閾値を 0.9 としてクラスタリングを行った．その結果，20 個のクラスタが生成された．

図 3 に cluster1,2,7 内に含まれる物質の最大共通構造を挙げる．これらの最大共通構造の比較を行う．たとえば，cluster1 と cluster2 では，ベンゼン環の結合位置の違いがクラスタを特徴付けていると推測できる．また，cluster 1 の共通構造はフラボノイドの基本骨格であるフラボン，cluster 7 はイソフラボンを含んだ構造となっている．ただし，cluster 2 の共通構造はフラボノイドの基本骨格にはない構造であった．

提案手法では，例えば，フラボノイド類とそうでない物質と

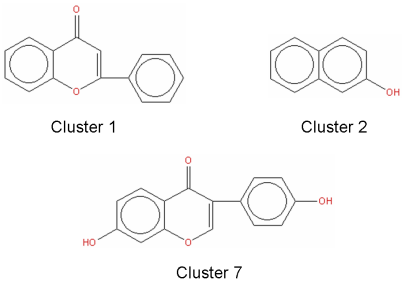

図 3: cluster1,2,7 の共通構造

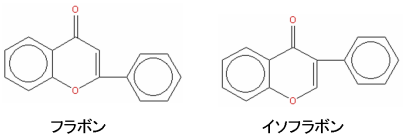

図 4: フラボノイドの基本骨格の一例

の区別や，フラボノイド類の基本骨格による分類がクラスタの違いに大きく反映していることが確認できた．

4. まとめと今後の課題

本稿で提案したクラスタリングは，グラフを構成する各ノードの特徴を，対象とする集合に存在する部分構造とどのような関わりを持つかにより定義し，各グラフはそのノードの特徴の集合体とみなすことにより実現された．類似度も同様の考え方に基づく．その結果，一定レベルの妥当な結果が得られたが，検討すべき課題が多く残っている．まず，比較するグラフのサイズの違いによる類似度の評価の曖昧さである．たとえば，ノードと関わりがある部分構造のサイズによって，ある程度グラフのサイズを反映することができるが，抽出される部分構造のサイズがもとのグラフのサイズに比べて非常に小さいものばかりの場合は類似度に反映されないことがある．次に，異なるノードラベル間の違いが類似度に反映されないことである．すなわち，本提案アルゴリズムは，異なるノードラベルは類似度計算においては無視されたるために不一致数には含まれず，積極的に類似性を高く評価するものといえる．

今後は，以上の問題を加味した類似性評価も導入することにより，より厳密なクラスタリングアルゴリズムとしての確立に向けて検討していきたい．また，他手法 [高橋 03][速水 05] との特性の違いや，より汎用的な手法としてさまざまな分野への適用も検討していきたい．

参考文献

[高林 05] 高林 健登, 他:グラフ構造データからの特徴的なパターン抽出における探索の効率化, 第 19 回人工知能学会全国大会，2F3-01 (2005).

[高橋 03] 高橋 由雅，他:化学物質の構造類似性にもとづくデータマイニング, *J.Comput.Chem.Jpn.*, Vol.2, No.4, pp.119-126(2003).

[速水 05] 速水 亜希子，他:部分構造の包含関係を指標とするグラフクラスタリングの提案－化学物質を対象として－, 人工知能学会 知識ベースシステム研究会，SIG-KBS-A405, pp.1-6 (2005)