3C1-03

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可視化 —化学物質を対象として—

A Step Towards Visual Graph Clustering based on Inclusion measure among a set of Subgraphs

石井　雄一郎[*1]　田中　栄太朗[*1*3]　速水　亜希子[*1*3]　大野　博之[*2]　稲積　宏誠[*2]
Yuichiro Ishii　Eitaro Tanaka　Akiko Hayami　Hiroyuki Oono　Hiroshige Inazumi

[*1]青山学院大学大学院　理工学研究科
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
Department of Science and Engineering, Aoyama Gakuin University

Graph-Based Induction (GBI) was proposed as an effective approach which enables us to extract typical patterns from graph data by stepwise pair expansion. A new graph clustering method based on inclusion measure among a set of subgraphs, extracted by applying GBI, have proposed. In this paper, considering chemical databases, we discuss some visual graph clustering method representing a typical structure of each cluster using multiple GBI.

1. はじめに

事例からの知識発見において, グラフ構造情報の効率的な取り扱いが要求される例として化学物質が挙げられる. 物質の構造上の特徴から，その性質などの予測を試みることができれば, 物質を合成する前にその評価を行なうことができ, 新規化学物質の開発にとっての意義は極めて高い. このため, 複雑な構造データからのパターン抽出に対する研究が近年盛んに行われており, グラフ構造データからのマイニングに対してもさまざまな研究がなされている.

我々は, 化学合成や創薬における支援システムとして, GBI 法 [松田 01] に注目し，GUI 環境による可視化を含む柔軟な GBI 処理を目的として多段 GBI を提案した [田中 05]. また, GBI 法により得られた部分構造を用いて, グラフ構造表現された事例のクラスタリング方法を提案した [速水 05]. しかしながら，クラスタリングにより得られた各クラスタがどのような構造上の特徴を持っているかについてを理解するためには, その可視化が望まれる.

本稿では，部分構造の包含関係に基づくクラスタリング結果から，そのクラスタ中の代表的な構造を可視化する手法について検討する. 特に，多段 GBI を用いた実現方法について，化学物質であるフラボノイドデータを用いて検討する.

2. 部分構造関係に基づくグラフクラスタリング [速水 05]

2.1 準備

対象とするすべての化学物質をグラフ表現し，そのグラフ構造データから，一定頻度で存在する部分構造を抽出する. その抽出された部分構造に対して，以下の処理を行なう.

1. 部分構造間に存在する包含関係 (順序関係) に注目し，抽出された部分構造をノード, 順序関係をリンクとする有向グラフ（部分構造関係グラフ）で表現する. このとき, 種々の冗長構造は除去する.
2. 部分構造関係グラフにおける末端ノード, すなわち, その部分構造を包含する他の部分構造がないノードを特定する.
3. それぞれの末端ノードからたどれるリンクをすべてたどり, 部分グラフを抽出する. この部分グラフから閉路を取り除くことによって，末端ノードを根とする木構造（部分構造関係群）に変換する.
4. 任意の化学物質を，このようにした表現された部分構造関係群の集合として表現する.

任意のグラフ構造データは部分構造関係群の集合として表現することができることから, これを用いることによって, 化学物質の構造上の類似性を, そこに含まれる部分構造の関係を用いて定義し, その特徴に基づくクラスタリングを考える.

2.2 クラスタリング方法

対象とする全物質からは，すでに部分構造関係グラフが求められているものとする. 各化学物質は, そこから得られる部分構造関係群の, どのノードに対応する部分構造までもっているかによって表現することができる. これにより，以下の処理を行なう.

1. 各群においてその事例における最大の部分構造を表すノードを求め，代表ノードとする. ただし, その半順序性から代表ノードが群に対し複数存在することもある.
2. 各物質は，どのような代表ノードを持つかにより特徴づけられることとなり，各代表ノードには，群を特定する ID，部分構造を特定する ID，部分構造の大きさ，根ノードからの分岐レベルなどの情報を保持させる.
3. 群内の同一系列中に存在する代表ノード間の距離を定義する. それを用いて，群中の部分構造の大きさの差を相違度として求める. 全ての群についてその相違度を計算し, その値を評価尺度としてクラスタリングを実現する.

たとえば, 今回実験で用いた Dihydroxyflavone と Alpinetin を含むフラボノイド事例集合から図 1 に示すような部分構造が抽出され, それらの包含関係から, 群 ♯1, ♯2 で表現される部

連絡先: 石井雄一郎，青山学院大学大学院
〒 229-8558 神奈川県相模原市淵野辺 5-10-1
E-mail:c5604015@cc.aoyama.ac.jp
[*3]現在（株）NTT データ

分構造関係群が導かれたとする．この 2 つの物質に対しては，図 2 のように代表ノードが与えられる．ボトムアップクラスタリングにおいてクラスタ同士を統合する際は，計算された相違度を用いて次のようにクラスタの代表ノードを算出する．統合すべき 2 つの事例の代表ノードの部分構造の大きさの平均値を統合後の代表ノードの示す部分構造の大きさとする．その値が， 部分構造関係群中の順序関係を満足する位置を求め，仮想ノードとして挿入する．この仮想ノードには相違度計算可能な情報を付与する．先に示した事例が統合されてクラスタが生成された場合の仮想ノードを図 3 に示す．この仮想ノードを統合後のクラスタの代表ノードと定義し，通常の代表ノードと同等の情報を保持させた上で，クラスタリング操作を継続する．

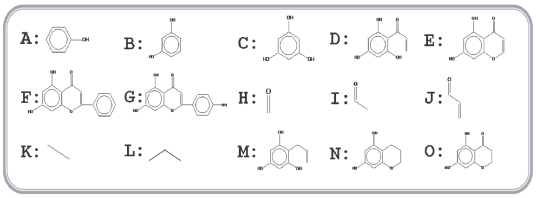

図 1: 部分構造の例

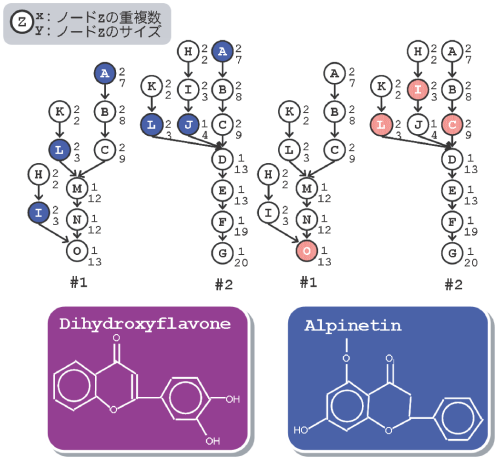

図 2: 事例に対する代表ノードの例

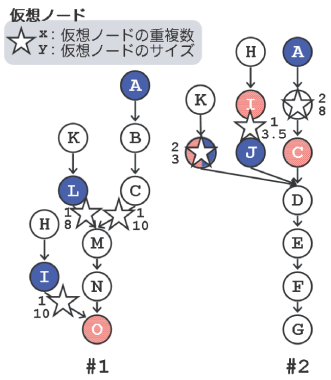

図 3: クラスタの仮想ノードの算出

3. 多段 GBI によるクラスタ特徴構造の抽出

GBI 法は，グラフ構造中の類型パターンを，ノードペアを逐次拡張することにより獲得していくものである [松田 01]．この GBI 法を用いて，入力物質群からある共通の基盤構造を抽出し，その基盤構造を含む物質群のみに対して，基盤構造とそれに付加される部分構造の組み合わせを抽出する方法を多段 GBI 処理と定義する [田中 05]．多段 GBI では，ユーザが特定した部分構造以外のすでに抽出済の類型パターンをリセットする機能 (還元処理) や，特定した部分構造に付加する類型パターンのみを逐次拡張する機能を組み合わせることができる．また，いずれの処理過程も GUI 環境が整備されているため，定量的な評価指標とユーザの選択を組み合わせることにより，処理を実現することができる．

このような多段 GBI 環境に基づき，同一クラスタ内の物質群に対して以下の処理を行なう．

1. クラスタ内全物質に対して基本 GBI 処理を実行し，その共通構造のみを抽出し，それ以外の部分構造に対して還元処理を行なう．
2. クラスタ内全物質をサブクラスタに分割し，それぞれに対して手順 1 で求めた共通構造を包含する基盤構造を多段 GBI により抽出する．
3. 手順 2 で求められた基盤構造に対して多段 GBI を実行し，その最大構造を求め，それぞれを候補構造として保持する．さらに候補構造間のの合成構造を求めて候補構造に加える．
4. 各候補構造に対して，部分構造関係群中の代表ノードを求め，クラスタ代表ノードとの相違度に基づいて候補構造からクラスタ特徴構造を選定する．

今回用いたクラスタリングでは，各クラスタの代表ノードが，クラスタのセントロイドとして計算されている．しかしながら，その代表ノードそのものが，その物質群を特徴付ける特徴構造であることが保障されているわけではない．したがって，代表ノードになるべく近いノードを保持する仮想物質であることを一定条件とはするが，なるべく多角的に仮想構造を合成していくというのが，本稿での考え方である．したがって，セントロイドから近い物質群を特定し，そこからクラスタ特徴構造を求めるのではなく，サブクラスタから基盤構造を求めていくという考え方をとった．

現段階では，経験的基準から実験を行い，化学分野の専門家からの意見を聴取している．これらに基づいて，サブクラスタの粒度，多段 GBI における基盤構造と付加構造の決定，さらに候補構造群から合成構造を求める方法を確立する予定ある．

参考文献

[松田 01] 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 A, pp.363-374 (2001)

[田中 05] 田中栄太朗，稲積宏誠：GBI 法の拡張と GUI によるグラフマイニング支援環境の構築, 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).

[速水 05] 速水亜希子, 稲積宏誠：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, 人工知能学会知識ベースシステム研究会 SIG-KBS-A405, pp.81-86(2005).