D_032

構造類似性を基にしたグラフクラスタリング手法の検討

A Step towards New Graph Clustering Algorithm Based on Structural Similarity

和田　貴久†　大野　博之‡　稲積　宏誠‡
Takahisa Wada　Hiroyuki Oono　Hiroshige Inazumi

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 [1] 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．抽出された部分構造は，Cl-GBI 法の特性により以下のような情報を持つ．

a) 抽出部分構造の中には構造間に包含関係の情報．

b) グラフ構造データ内の各ノードは、抽出された部分構造の中のどの構造と関係を持っているかという情報．

本提案手法では、特に b) のノードと部分構造の関係情報を利用する．

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 $N = \{x_i | x_i \in \{\mathrm{C,O,N,\cdots}\}, i = 1,2,\cdots\}$ 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 $P = \{\mathrm{P_1,P_2,\cdots,P_j}\}$ とする．部分構造 $\mathrm{P_j}$ 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．4 つの部分構造 $\mathrm{P_1,P_2,P_3,P_4}$ を持つグラフの行列表現例を図 1 に示す．

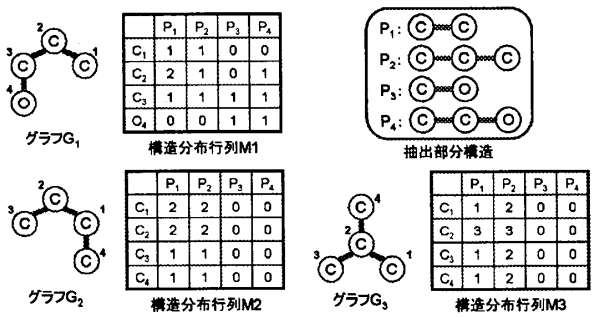

図 1: 構造分布行列の例

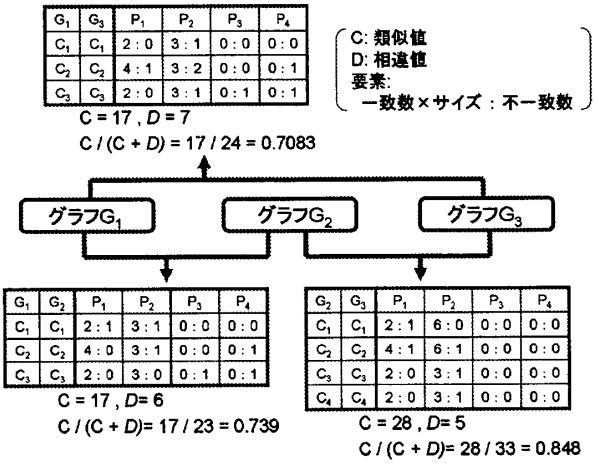

図 2: グラフ間類似度の算出例

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．ここで，グラフ間の類似度とはグラフに含まれている各ノード毎の類似度によって求められる．まず，ノード間の類似度を求めるためにラベルの同じノードを 2 つ取り上げ，各部分構造毎に関連している数を比較する．各関係数の一致している数に重みとして注目している部分構造のサイズの積を計算しその総和を類似値，不一致数の総和を相違値とする．これらを用いて，ノード間の類似度は，類似値/(類似値 + 相違値) と定義する．同じラベルのノードペアすべてに対して類似度を求め，各々で最も高い類似度を取るような組み合わせを探す．グラフ間類似度は，各ノード間類似度を求める際に利用した類似値と相違値のそれぞれの総和を用いて，$\sum$類似値/($\sum$類似値 + $\sum$相違値) を計算することで求める．このとき，互いにラベルが異なるノード同士は比較しない．しかし，図 1 のグラフ $\mathrm{G_1}$ とグラフ $\mathrm{G_2}$ のように，$\mathrm{O_4}$ と $\mathrm{C_4}$ のノードは，$\mathrm{C_2}$ や

†青山学院大学大学院　理工学研究科　理工学専攻
‡青山学院大学　理工学部　情報テクノロジー学科

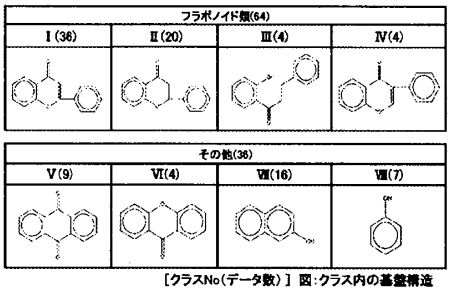

図 3: 実験データの既知情報による分類

C_3 のノードの比較の際に使用される O を含む部分構造 P_3 や P_4 の情報によってサポートされているといえる．

以上のような一連の類似度算出の手順を 2 つのグラフ G_A , G_B を用いて以下に示し，グラフ間の類似度の算出例を図 2 に示す．

1) グラフ G_A からノード N_A を選び，グラフ G_B から N_A と同じラベルを持つノード N_B を用意する．
2) 構造分布行列を利用し，2 つのノード（N_A, N_B）の類似値，相違値を計算する．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，その総和を類似値とする．
   b) 2 つのノードにおいて，各部分構造の数の差を計算し，その総和を相違値とする．
3) 求めた類似値・相違値を用いて，類似値/(類似値＋相違値) を計算し，それをノード間類似度とする．
4) G_B 中に含まれる N_A と同じラベルを持ち，かつまだ類似度を計算していないすべてのノードについて 2) 3) の処理を行う．
5) 各ノードペアの中で，類似度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．
6) 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)〜5) の処理を繰り返す．
7) 保存されている類似値と相違値のそれぞれの総数から $\sum$類似値/($\sum$類似値＋$\sum$相違値) を計算し，これをグラフ間類似度とする．

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

本稿で提案した手法の特性を検討するため，Cl-GBI 法によって抽出される部分構造の数を変えてクラスタリングを行う．実験データとして，図 3 にある I〜VIII のように基盤構造によって分類される計 100 種類の化学物質を用いる．部分構造の数は，Cl-GBI 法のパラメータ（繰り返し回数，ビーム幅）を変化させることで調節が可能である．本実験では，A:40 個，B:70 個，C:195 個，D:544 個の 4 つの部分構造セットを用意した．各セット間では，部分構造の数だけではなく部分構造のサイズにも違いが出てきているため，セット D の中の部分構造はセット A の中の部分構造のサイズより大きなものが多い．これは，Cl-GBI 法が逐次的にチャンクし部分構造の拡張による抽出を行っているためである．

表 1: 部分構造数別のクラスタリングの結果

	A(40)	B(70)	C(195)	D(544)
I	A1(21/22) A2(14),A3(1)	B1(36)	C1(36)	D1(34),D2(2)
II	A4(19/24) A5(1)	B2(19),B3(1)	C2(20)	D3(20/24)
III	A6(4)	B4(4)	C3(4)	D3(4/24)
IV	A4(4/24)	B5(3),B6(1)	C4(4)	D4(4)
V	A7(6), A8(3)	B7(9)	C5(9)	D5(9)
VI	A9(4)	B8(4)	C6(4)	D6(4)
VII	A10(8), A11(1) A12(1) ,A13(6)	B9(9),B10(5) B11(1)	C7(10),C8(4) C9(2)	D7(15),D8(1)
VIII	A4(1/24) ,A14(1) A15(4) ,A1(1/22)	B12(4),B13(2) B14(1)	C10(3),C11(2) C12(1),C13(1)	D9(5),D10(1) D11(1)

クラスタ名(数)or クラスタ名(該当数 / 総数)

各部分構造セットを利用してクラスタリングを行った結果を表 1 に示す．最も部分構造の数が少ないセット A を用いたクラスタリングの結果は，クラスタ A1 や A4 のように同一クラスタ内にさまざまなクラスに属する化学物質が含まれてしまった．これは，セット A 内の部分構造のサイズが小さいものが多いため，図 4 のような 2 つの物質間の類似度にサイズの違いが反映されなかったためと考えられる．しかし，セット B，C，D などのように部分構造の数を増やすことで，大きな部分構造を含むか否かという情報によってサイズの違いも反映することができている．結果として，基盤構造のような構造上の違いによって分かれるクラスをほぼ表現しうるクラスタを生成することができている．

また，クラスタリングを行う際にノードと部分構造の関係を利用しているので，そのノードと部分構造に含まれる複数のリンクとノードとの関係を見ていることになる．よって，I と II のように形はほぼ同じなのに一部分のリンクが違うような基盤構造の特徴を持つクラスタを形成できているのである．

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，検討すべき課題も残っている．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 [2] 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

[1] 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).

[2] 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).