

対象グラフ集合の特性を反映した構造類似性の提案

New Structure Similarity in Accordance with the Feature of Target Graph Sets

和田 貴久[▼]
稲積 宏誠[▲]

大野 博之[◆]

Takahisa WADA
Hiroshige INAZUMI

Hiroyuki OONO

蓄積されるデータの多様化や複雑な構造を持つデータの増加に伴い、構造データを取り扱うための有用な DB システムやデータマイニング手法の開発は、データの有効活用のために重要となってきた。本論文では、対象とするグラフ集合より特徴的な部分構造を用いることで定義できる構造類似性を提案する。この類似性は、各グラフのノードと部分構造の関係で表される構造分布行列の比較と重み付け計算より定義される。また、この類似性の特性について、評価および応用について考察する。

The graph data with the complex structure increases more and more along with the diversification of the accumulated data. Therefore, effectively to leverage data, the development of a useful DB system and the data mining technique to handle structural data is imperative. In this paper, we propose new structure similarity in accordance with the feature of target graph sets, and discusses a graph clustering method based on its criterion. This similarity is defined by the comparisons of the processions be represented as the relation between a node in each graph and a substructure. We consider the evaluation of the similarity characteristic.

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。中でも、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、比較的新しい分野であり、精力的に研究が進められ、多くの有用なアルゴリズムが提案されている [1]。我々は、特に、対象とするグ

ラフ集合に共通する部分グラフ抽出のためのアルゴリズムである Graph-Based Induction (GBI) 法 [2, 3] に注目して、化学物質の特性を分析するためのツールの開発や、GBI 法により得られた部分グラフ情報を用いた分類問題やクラスタリングへの応用について検討を行ってきた。GBI 法は、ノードの置き換えと Greedy 探索による見落としが存在していたが、Chunkingless Graph-Based Induction (CI-GBI) 法 [4, 5] の提案によって、従来の GBI 法の欠点が完全に補われ、その応用範囲も広がったと考えられる。

本稿では、グラフ構造情報のより有効な分析を実現するために、グラフにおける構造上の類似性に注目し、分析対象とするグラフ集合全体の持つ構造上の特徴に基づくグラフ間の構造類似性を検討する。構造的な類似性の考え方には、与えられたグラフ集合のグラフ毎に、グラフ中に含まれる連結部分グラフを列挙し、それを数値化したものを利用するという考え方が多く用いられている [6]。すでに、グラフ構造の構造的特徴付けとして化学物質の構造的類似性の観点から TFS (Topological Fragment Spectra) [7] や、対象とするグラフ集合をそこから得られる部分グラフにより特徴付け、それらの部分グラフ間の包含関係により定義される部分グラフの半順序構造に注目する特徴付け [8] などを用いた類似性の尺度がある。本稿での考え方は、対象とするグラフ全体をあらゆる特徴として、部分グラフ間の包含関係に基づく部分グラフ集合を用いるのではなく、包含関係を有する部分グラフが、対象とするグラフ中にどのように分布しているかを評価する。すなわち、グラフを構成する一つ一つのノードが、どのような部分グラフにどの程度含まれているかを情報として保持し、それを用いてグラフ間の構造的な類似性を評価するという考え方である。つまり、グラフ構造データから CI-GBI 法を用いて部分グラフの抽出を行い、それを用いたグラフ間類似度の計算方法を提案する。グラフ構造は汎用的なデータ構造であり、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。ただし、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、本稿で提案した構造類似性の特性の検討を行う。

2. 部分構造情報に基づく構造類似性

2.1 CI-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、CI-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グラフを高速に抽出することができる。

CI-GBI 法は、グラフ内に存在する 2 つのノードとそれらを結ぶリンクから構成されるノードペアをすべて特定し、その中から出現頻度の高いノードペアを予め設定するビーム幅だけ疑似チャンクを行う。疑似チャンクとは、GBI 法で行なわれるチャンク (ノードペアを表現する疑似ノードを生成し、完全に置き換える操作) とは異なり、ノードペアのノードとリンクの情報はそのまま保持する形で追加する操作である。この疑似ノードも含めて頻度の高いノードペアの抽出を繰り返し実行する。これにより、部分的に重なる部分グラフなどのすべての部分グラフを抽出する

▼ 学生会員 青山学院大学大学院理工学研究科博士前期課程 t-wada@ina-lab.it.aoyama.ac.jp

◆ 非会員 青山学院大学理工学部情報テクノロジー学科 oono@ina-lab.it.aoyama.ac.jp

▲ 非会員 青山学院大学理工学部情報テクノロジー学科 hiro@ina-lab.it.aoyama.ac.jp

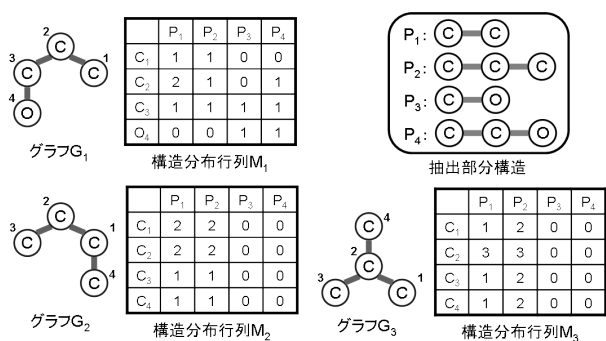


図1 構造分布行列の例

Fig. 1 Example of Structural Distribution Matrices

ことができることが可能である。

しかし、本稿で提案する構造類似性を意味のあるものにするためには、部分グラフの全探索を実現するよりも、グラフ間の特徴の違いを反映するノードペアを抽出することが重要となる。なぜならば、抽出部分グラフの増大は計算量の増大につながるため、なるべく効率的にグラフ間の特徴の違いを表現することのできる部分グラフを抽出しておきたいためである。そこで、擬似チャンクの候補になっているノードペアを含有数、つまりそれを含むグラフの数に応じてグループ分けを行い、このグループ数はビーム幅と同一とした上で、グループ毎から1つずつノードペアを擬似チャンク対象とする。あるいは、頻度順に一定間隔を置いたうえで、ビーム幅に相当するノードペアを擬似チャンク対象とする方法や、領域知識に基づく制御戦略を導入することも有効であると考えられる。

以上の方法から、分析対象であるグラフの各ノードには、擬似ノード生成過程から、抽出された部分グラフの構成要素となっているかという情報が保持されている。これらが、構造類似性を評価するうえでの基本情報となる。

2.2 グラフの構造分布行列表現

CI-GBI法により抽出された部分グラフを用いることによって、各グラフは、それを構成するノードと部分グラフとの関係に基づいて表現することができる。対象とするグラフ集合から抽出された部分グラフ集合を P 、抽出された部分グラフ数を J 、ノード n_i^k を含む部分グラフ $p_j \in P, j = 1, 2, \dots, J$ 、の数を $m_{ij}^k = m^k(n_i^k, p_j)$ とおくと、任意のグラフ G_k は以下に示す行列 M_k で表現することができる。

$$M_k = \begin{bmatrix} m_{11}^k & m_{12}^k & \cdots & m_{1J}^k \\ m_{21}^k & m_{22}^k & & \\ \vdots & & \ddots & \vdots \\ m_{I_k 1}^k & & \cdots & m_{I_k J}^k \end{bmatrix} \quad (1)$$

この行列 M_k を構造分布行列と呼び、 G_k の構造上の特徴が表現されているとみなす。これは、CL-GBI法の実行過程の情報から容易に作成することができる。

図1に構造分布行列の例を示す。これは、対象とするグラフ

集合全体を化学物質としたときに、抽出された部分グラフが $p_1: C-C, p_2: C-C-C, p_3: C-O, p_4: C-C-O$ であったときの、グラフ G_1, G_2, G_3 の特徴表現である。

2.3 グラフ間類似度の算出法

本稿では、まず、任意の2つのグラフ中に含まれる共通ラベルを持つノード集合ごとに、各ノード間の類似性を定義し、次に、そのノード間の類似性を用いてグラフ間の構造類似性を定義する。その際、各ノード間の類似性は、関連している部分グラフの共通数の多いノードペアから評価していくこととする。そのため、対象とするグラフ間で同一ラベルを持つノード数が異なる場合には、余分のノードは評価対象としない。たとえば、比較対象となる一方のグラフにのみ、あるノードラベルを持つノードが多数存在しても、その多くのノードは構造類似性の尺度には、直接は反映しない考え方である。しかし、一方にしか含まれないノードは、共通するノードと関連を持つ部分グラフの情報によって間接的に反映される。

まず、任意のグラフ対に対して、構造類似性を表す尺度であるノード間類似度を定義する。今、比較対象となるグラフを G_1, G_2 とする。簡単のために、ノードラベルを1種類、ノード数については、 G_1 のほうが G_2 より多いとする。また、対象とするグラフ集合全体から抽出されている部分グラフの数を J 、それぞれの構造分布行列を M_1, M_2 、さらに、部分グラフ p_i を構成するノード数を $size(p_i)$ とする。このとき、グラフ G_1 のノード x とグラフ G_2 のノード y のノード間類似値 r_{xy}^{12} およびノード間相異値 d_{xy}^{12} を以下のように定義する。

$$r_{xy}^{12} = \sum_{j=1}^J \alpha_j \min(m_{xj}^1, m_{yj}^2) \quad (2)$$

$$1 \leq \alpha_j \leq size(p_j)$$

$$d_{xy}^{12} = \sum_{j=1}^J \beta_j |m_{xj}^1 - m_{yj}^2| \quad (3)$$

$$1 \leq \beta_j \leq \alpha_j$$

$$m_{ij}^1 = m^1(n_i^1, p_j) \in M_1$$

$$i = 1, 2, \dots, I_1$$

$$m_{ij}^2 = m^2(n_i^2, p_j) \in M_2$$

$$i = 1, 2, \dots, I_2, \quad I_2 \leq I_1$$

これらを用いて、ノード間類似度 s_{xy}^{12} を以下のように定義する。

$$s_{xy}^{12} = \frac{r_{xy}^{12}}{r_{xy}^{12} + d_{xy}^{12}} \quad (4)$$

ここで、定義されたノード間類似度がより高くなるようなノードペアを選び出す。そのための準備として、2つのグラフに含まれる部分グラフのうち最も大きな共通部分グラフを用意し、そ

の部分グラフと関連を持つノードを各グラフより取り出す．そして、それらのノード群の中でノード間類似度を計算し、その値がより高くなるノードペアを見つける．続いて、残りのノード群の中でノード間類似度が高くなるノードペアを見つける．これは、比較するグラフに共通する最も大きな部分グラフに関連するノード同士はより高い類似度を得る可能性が高いと考えられるからである．最終的に I_2 個のノードペアを選び、それをグラフ間類似度を評価するための比較ノードペアとして確定する．その結果、グラフ G_1 と G_2 から選択されたノードペア $(x_1, y_1), (x_2, y_2), \dots, (x_{I_2}, y_{I_2})$ に対して、グラフ G_1 と G_2 のグラフ間類似値 R^{12} 、グラフ間相異値 D^{12} およびグラフ間類似度 S^{12} を以下のように定義する．

$$S^{12} = \frac{R^{12}}{R^{12} + D^{12}} \quad (5)$$

$$R^{12} = \sum_{i=1}^{I_2} r^{12}_{x_i y_i} \quad (6)$$

$$D^{12} = \sum_{i=1}^{I_2} d^{12}_{x_i y_i} \quad (7)$$

このようにして定義したグラフ間類似度は、次のような考え方に基づいている．

1. 各ノードの類似性を評価する際には、同一の部分グラフをもつ個数を加味して評価する．したがって、部分グラフが対象とするグラフ中にどのように分布しているかという点についても、類似性評価に加味されることになる．
2. 部分グラフのサイズを類似性評価に加味することができることとしている．これは、大きな部分グラフを共有するかどうかという点で、類似性の評価に大きく影響するであろうとの考え方による．ただし、類似値については、 $1 \leq \alpha_j \leq size(p_j)$ 、相異値については、 $1 \leq \beta_j \leq \alpha_j$ とする．これは、部分グラフのサイズの反映については、類似値が相異値を下回らないという条件を置いたことによる．

3. 構造類似性の特性の考察

本稿で提案するグラフ間類似度では、特に、ノード数が大きく異なる場合や、一方のグラフのみにノードラベルが存在し、他方にはそのラベルが存在しない場合に、適切な類似度評価が行えているかどうかについての検討が必要となる．

そこで、ここでも化学物質を例として、図 2 に示す 4 つのグラフを用いて考えてみることにする．ここでは、ノード間類似値および相異値の重みは $\alpha_j = size(p_j), \beta_j = 1$ とし、類似部分を高く評価することとする．リンクラベルは 1 種類、ノードラベルは、C と O の 2 種類とし、グラフ G_1 とグラフ G_3 がノード数が 4 で一方に O が含まれ、グラフ G_2 とグラフ G_4 がノード数 3 で一方に O が含まれている．また、抽出されている部分構造も図中にあるとおり、少数ラベルである O についても、部分グラフ

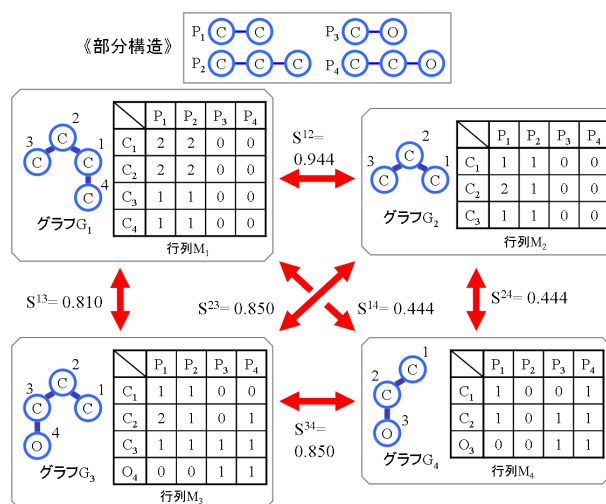


図 2 構造分布行列とグラフ間類似度

Fig. 2 Structural Distribution Matrices and Graph Similarity

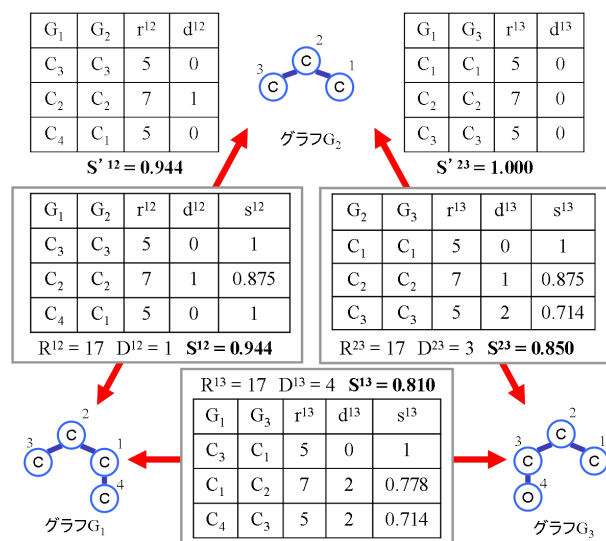


図 3 グラフ間類似度の抽出部分グラフによる影響

Fig. 3 Graph Similarity Influenced by Selected Subgraphs

としての頻度は基準以上であり、評価対象となることを意味している．4 つのグラフ間のグラフ間類似度では、同様の条件ながら、 $S^{13} \geq S^{24}$ となっている．まず、グラフを構成するノード数が少ないほど、類似度を評価する際のノード数の差やラベルの違いによる影響が反映されることが推測される．図 3 は、ノード間類似度とグラフ間類似度をまとめた．図 3 を見ると、 G_2 は、 G_1 、 G_3 に同様に含まれているにもかかわらず、 $S^{12} \geq S^{23}$ となっていることがわかる．このことは、ノードラベルの違いが、抽出された部分グラフに反映される場合には、ノードラベルの違いにより類似度を下げるように機能することがわかる．しかし、部分グラフである P_3 や P_4 が一定頻度に満たず、採択されない場合

には、それらが、相異値として反映されなくなるため、 $S^{123} = 1$ 、 $S^{12} = 0.944$ となり、 $S^{12} \leq S^{123}$ と逆転することになる。このように、ある共通部分を共有したうえで、異なるラベルが付加されているような場合には、その付加ラベルを含む部分グラフがどのような頻度であるかによって、グラフ間類似度が制御されることになる。よって、評価対象となるグラフ集合のもつ全体としての性質が、グラフ間類似度に反映されていることがわかる。

したがって、本稿で提案したグラフ間類似度は、対象とする2つのグラフ間でのノード数の違いやラベルの偏りについて、次のような特性を示すものと考えられる。

1. ノード数が大きく異なる場合：少ないノード数を持つグラフを基準にノード間類似度が評価されることになる。しかし、多くのノードをもつグラフにのみ存在する部分グラフが、対象とするグラフ中に一定頻度以上存在し、対象とするグラフ集合のなかで採択されている場合には、一部のノードではその影響を受けて相異値の増大を生むこととなり、相対的に類似度を低下させることとなる。しかし、多くのノードをもつグラフにのみ存在する部分グラフが、対象とするグラフ集合にほとんど存在しない場合には、その部分グラフの影響が相異値として反映されないため、類似度を低下させる要因とはならない。
2. 一方のグラフのみにノードラベルが存在し、他方にはそのラベルが存在しない場合：該当するノードラベルについてのノード間類似度はまったく評価されない。しかし、該当するノードラベルをもつグラフにのみ存在する部分グラフが、対象とするグラフ集合中に一定頻度以上存在しており、その部分グラフが、対象とするグラフに共通するノードラベルとの組み合わせで構成されている場合には、一部のノードで相異値の増大を生むこととなり、相対的に類似度を低下させることとなる。しかし、多くのノードをもつグラフにのみ存在する部分グラフが、対象とするグラフ集合にほとんど存在しない場合には、その部分グラフの影響が相異値として反映されないため、類似度を低下させる要因とはならない。

4. まとめ

本稿で提案した構造類似性を反映したグラフ間類似度の定義は、対象とするグラフ集合の特徴を反映するものとして定義された。これは、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グラフ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る。ただし、これを実現するための部分グラフ抽出については、同型の部分グラフを漏れなく抽出することができるCL-GBI法を用いるのが適当であると考えた。なぜならば、部分グラフの戦略的な抽出が可能のため、必要に応じて抽出の粒度を変えられ、計算量の調整を行うことができるからである。ただし、対象とするグラフ集合の性質と分析目的に応じて、必要とする部分グラフのもつ要件を明確にする必要がある。また、本稿で

提案した類似性は、化学物質のクラスタリングへ応用されているが、今後は、更なる領域への応用を検討していきたい。

[文献]

- [1] Kuramochi, M. and Karypis, G.: An Efficient Algorithm for Discovering Frequent Subgraphs, *IEEE Trans. Knowledge and Data Engineering*, Vol.16, No.9, pp.1038-1051 (2004)
- [2] 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, *人工知能学会誌*, Vol.16, No.4, pp.363-374 (2001)
- [3] Matsuda, T., Motoda, H., Yoshida, T. and Washio, T.: Mining Patterns from Structured Data by Beam-wise Graph-Based Induction, *Proc. of DS2002*, pp.422-429 (2002)
- [4] Nguyen, P., Ohara, K., Motoda, H. and Washio, T.: Cl-GBI: A novel strategy to extract typical patterns from graph data, *SIG-KBS-A403*, pp.105-110 (2004)
- [5] Nguyen, P., Ohara, K., Motoda, H. and Washio, T.: Cl-GBI: A Novel Approach for Extracting Typical Patterns from Graph-Structured Data., *Proc. of PAKDD2005*, pp.639.649 (2005)
- [6] Palmer, C., Gibbons, P. and Faloutsos, C.: ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs, *Proc. of the KDD-2002* (2002)
- [7] Takahashi, Y., Ohoka, H. and Ishiyama, Y.: Structural Similarity Analysis based on Topological Fragment Spectra, *Advances in Molecular Similarity*, Vol.2, pp.93-104 (1998)
- [8] 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 - 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)

和田 貴久 Takahisa WADA

2006年青山学院大学理工学部情報テクノロジー学科卒業。現在、同大学大学院修士課程在学中。人工知能学会学生会員。日本データベース学会学生会員。

大野 博之 Hiroyuki OONO

2002年青山学院大学大学院理工学研究科経営工学専攻修了。同年日本電気株式会社入社。2005年より青山学院大学理工学部情報テクノロジー学科助手現在に至る。情報処理学会会員。

稲積 宏誠 Hiroshige INAZUMI

1984年早稲田大学大学院理工学研究科(博士前期課程)修了。同年早稲田大学情報科学研究教育センター助手。1987年相模工業大学(現湘南工科大学)工学部情報工学科専任講師。1993年青山学院大学理工学部経営工学科助教授, 2003年青山学院大学理工学部情報テクノロジー学科教授現在に至る。工学博士。電子情報通信学会, 情報処理学会, IEEE, ACM 各会員。