3F1-02

構造的類似性に基づくグラフクラスタリング

Graph Clustering with Structure Similarity

庄田 良介[*1] 松田 喬[*1] 吉田 哲也[*1] 元田 浩[*1] 鷲尾 隆[*1]
Ryosuke Shoda Takashi Matsuda Tetsuya Yoshida Hiroshi Motoda Takashi Washio

[*1]大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University

This paper presents a graph clustering method with structure similarity. The structure of a graph is converted into a spectrum in terms of the characteristics of connected subgraphs in the graph as in the TFS method and the similarity of graphs are analyzed as the similarity of converted spectrums. The spectrums are then clusterd by k-menas method based on their similarity. Furthermore, the entropy of a cluster is defined to estimate the quality of clusters and used to determine the number of cluster K in K-means automatically. Preliminary experiments with synthesized graphs were conducted and the results are reported.

1. はじめに

近年の計算機ハードウェア性能とネットワーク技術の急速な進歩により，電子的に可読なデータの量も増加の一途をたどっている．それに伴い，膨大な蓄積データから有用な知識を発見することを目的とするデータマイニングの新しい手法が研究開発されており，さまざまな分野で多大な成果をあげている．従来から構造を有しない通常のデータに対してデータを属性とその値のペアで表現し，属性の値と同定したいクラスの関係を決定木 [Quinlan 93] や分類規則 [Michalski 90] で表現するさまざまな手法が研究されてきた．データマイニングでよく使われる相関規則 [Agrawal 94] もこの表現形式に入る．しかし，属性と値のペアの表現形式は複雑な構造データを表現するのには適しておらず，より強力な表現形式が必要となるタスクもある．このため，表現形式を拡張して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 AGM [Inokuchi 00], FSG [Kuramochi 01], GBI [Matsuda 02a] 等様々な研究が行われている．

著者らはアクティブマイニングプロジェクトを通じ，GBI 法を用いて専門家との密接な連携を通じた肝炎データからの特徴的なパターン抽出に取り組んできた [Matsuda 02b]．その際，GBI 法は大規模なグラフ構造データから特徴的な部分グラフを高速に抽出できる反面，大規模なグラフ構造データから非常に大量の部分グラフが抽出されるため，抽出したパターンの評価を専門家に依頼してフィードバックを得にくいという問題にしばしば直面した．この一因として，マイニングを通じて得られた大量の部分グラフを人間がすべて把握し，理解することは非常に困難な作業であることが挙げられる．抽出した部分グラフは構造的に類似するものが多いため，それらを構造的な特徴に基づいてクラスタリングし，各クラスタから少数のサンプルのみを提示すれば，専門家の負担を軽減しながら個々の抽出パターン（部分グラフ）の評価とマイニング対象データの全体的な傾向を把握することができると期待される．そこで，本稿ではグラフを構造的類似性に基づいてクラスタリングする手法を提案し，人工データを通じた定量的評価を報告する．

グラフ構造データの構造的特徴付けに対しては化学構造物質の構造的類似性の観点から TFS [Takahashi 98] が提案されている．また，大規模グラフ構造の高速な類似性判定に対しては [Palmer 02] などの研究が行われている．本稿の手法はグラフの構造的類似性の捉え方に関しては TFS と類似する部分が多いが，部分グラフの特徴の捉え方やその特徴に基づいたクラスタリングの評価方法が異なっている．

連絡先: 大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
e-mail: yoshida@ar.sanken.osaka-u.ac.jp

2. グラフクラスタリング

本稿で提案する手法は，与えられたグラフ集合の各グラフをグラフ中に含まれるすべての連結部分グラフを列挙してそれぞれ数値化し，その数値に基づいてグラフ構造を多次元空間のベクトル（これをグラフスペクトルと呼ぶ）に変換する．次に，変換したグラフスペクトル間の類似性に基づいてクラスタリングを行う（図 1 参照）．以下，2.1 節でグラフスペクトルについて述べ，2.2 節でクラスタリングついて述べる．なお，現状では頂点，辺にラベルがなく，また有向辺を持たないグラフを対象としている．

2.1 グラフスペクトル

グラフ G を，有限集合 $V(G)$ と $E(G) \subseteq V(G) \times V(G)$ に対し，G=(V(G), E(G)) と定義する．ここで，$V(G)$ の要素は頂点 (vertex)， $E(G)$ の要素は辺 (edge) に対応する．グラス G のサイズをその頂点数 $|\ V(G)\ |$ と表し，グラフ G の頂点 x に接続している辺の個数を x の次数として $deg(x)$ と表す．また，G_1 と G_2 をグラフとするとき，H が G の部分グラフ (subgraph) であるとは，$V(G_2) \subset V(G_1)$ かつ $E(G_2) \subset E(G_1)$ となることをいい，これを $G_2 \subset G_1$ と表す．

本稿ではグラフ G に含まれる連結部分グラフに基づいてその構造的特徴をグラフスペクトルとして表現する．グラフ G のグラフスペクトルを下記の手順により生成する．

Step1 すべての連結部分グラフを列挙する．

Step2 部分グラフの特徴を定量化し部分グラフ SG に点数付けする（関数 **SGScore**）．

Step3 部分グラフの点数を次元とし，同じ点数を持つ連結部分グラフの頻度から成るスペクトルを生成する．

例えば，図 1 では点数 4 の連結部分グラフが 1 個，点数 5 の連結部分グラフが 4 個，点数 6 の連結部分グラフが 3 個, ...

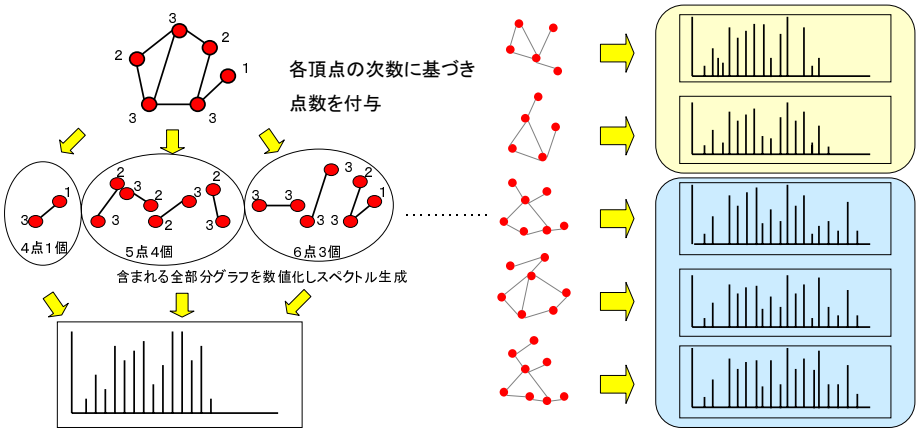

図 1: グラフスペクトルに基づくクラスタリング

含まれるためグラフスペクトルは {0, 0, 1, 4, 3, . . . } となる．上記の手順を図 2 に示す．

Graph_spectrum(graph G)
  GV: graph spectrum of G
  initialize GV to **0**
  forall connected subgraph $SG \in G$
    GV[**SGScore**(SG, G)] := GV[**SGScore**(SG, G)] + 1
  return GV

図 2: グラフスペクトル生成アルゴリズム

SGScore 関数

部分グラフの特徴を点数として定量化する際，グラフの接続関係に着目し，各頂点の次数に基づいて各頂点に点数を付け，その点数に基づいて部分グラフの点数を計算する．その際，各頂点に対する点数付けの方法として下記の 2 種類を考える．

- 部分グラフのみに着目し，部分グラフ内での次数とする．
- 元のグラフ中における部分グラフとその外部との接続関係に着目し，もとのグラフ中での次数とする．

また，各頂点の点数に基づいて部分グラフの点数を計算する際，べき乗を用いると頂点ごとの点数の差異を顕在化することに役立つ．ここでは，部分グラフの点数を 1) 各頂点の次数の和 2) 各頂点の次数の 2 乗和とする方法を考える．部分グラフの SGScore 関数として本稿で考慮した SGScore1, SGScore2, SGScore3 を表 1 にまとめる．表 1 で - に対応する方法は頂点の次数を部分グラフ内での次数とし，その和により点数を計算することに対応するが，部分グラフ内の頂点数と辺数が同じであれば接続関係に依存せず同じ点数になってしまい構造的な特徴を反映できないため考慮していない．

グラフ G に含まれる部分グラフ数は $|V(G)|$ に指数関数的に増加するため，グラフスペクトルの各要素を同じ点数を持つ部分グラフの頻度とすると，サイズの異なるグラフ間ではスペクトルの各要素の値が大きく異なってしまいスペクトルを比較することが困難になる．グラフのサイズの影響を除去するため

表 1: SGScore 関数

	頂点の次数	
	元のグラフ	部分グラフ
和	SGScore1	-
2 乗和	SGScore3	SGScore2

にスペクトルの要素の総和が 1 となるようにグラフ中に含まれる部分グラフの総数で頻度を正規化することを考え，上記の SGScore1, SGScore2, SGScore3 に対して正規化をする場合 (NSGScore1, NSGScore2, SGScore3) としない場合の計 6 種類の SGScore 関数を用いて部分グラフの点数付けを行う．

2.2 クラスタリング

グラフを 2.1 節で述べた手法でグラフスペクトルに変換し，これをクラスタリングすることで，グラフ集合をいくつかのクラスタに分割する．従来より様々なクラスタリング手法が提案されているが，本稿ではグラフスペクトルをそれぞれ多次元ベクトルとみなし，ベクトル間の類似度に基づいて K-means 法を用いてクラスタリングを行う．ベクトル間の類似度にはコサイン類似度等さまざまなものが提案されているが，本稿ではベクトルを各次元ごとに比較して類似度を考え，グラフスペクトル i，j の類似度 S_{ij} を以下により定義する．

$$S_{ij} = \frac{\sum_{m=1}^{M}(1 - \frac{a_{im} - a_{jm}}{\max a_m - \min a_m})}{M} \quad (1)$$

a_{im} : グラフスペクトル a_iの次元 m での値
$\max a_m$: 全グラフスペクトル中の次元 m での最大値
$\min a_m$: 全グラフスペクトル中の次元 m での最小値
M : 全グラフスペクトル中の最大次元数

K-means 法ではクラスタ数 K を事前に指定する必要があるが，K の値をあらかじめ指定することは困難である．本稿ではクラスタ数 K を自動的に決定するために，クラスタ内でのデータの偏りに基づいてクラスタ内のエントロピーを定義し，クラスタリング前後でのエントロピーの変化を求め，この変化が最大となる K によりクラスタ数を決定する．

$|C_k|$ をクラスタ C_k 内のグラフ数として，k(=1,2,···,K) 番目のクラスタ C_k 内でのエントロピー $Ent(C_k)$ を下記のよ

うに定義する．

$$Ent(C_k) = \frac{-\sum_{i=1}^{|C_k|}\sum_{j=1}^{|C_k|}(S_{ij}\log_2 S_{ij} + (1-S_{ij})\log_2(1-S_{ij}))}{|\ C_k\ |^2} \quad (2)$$

$Ent(C_k)$ はクラスタ C_k 内でデータが偏りなく分布するとき，あるいはデータが一点に重なって集中するとき低くなるような指標である．

上記のエントロピーに基づき，クラスタリング前のデータ（初期状態 C） を K 個のクラスタに分割した場合の情報利得比 (Infomation Gain Ratio for Clustering, ($IGRC$)) をクラスタリング前後でのエントロピーの差に基づいて下記により定義する．

$$IGRC = \frac{Ent(C) - \sum_{k=1}^{K}\frac{|C_k|}{|C|}Ent(C_k)}{-\sum_{k=1}^{K}\frac{|C_k|}{|C|}\log_2\frac{|C_k|}{|C|}} \quad (3)$$

本稿では IGRC の最大化を規範として k の値を決定し，K-means 法を適用した．

3. 予備的実験

上記の手法を計算機上に実装し，人工データを用いて下記の 2 種類の評価実験を行った．

3.1 実験で用いたデータ

各実験においては基本となるグラフ（基本グラフと呼ぶ）を用意し，図 3 に示す 3 つの部分グラフをグラフの各頂点に付加して生成したグラフ（派生グラフと呼ぶ）の集合に対し，IGRC が最大となる K の値でクラスタリングを行った．グラフの構造的特徴として頂点数，辺数と形状を考え，基本グラフを

図 3: 付加する部分グラフ

実験 1 頂点数，辺数が異なり，形状は類似なグラフ

実験 2 頂点数，辺数は同じで，形状が非類似なグラフ

とし，図 3 の部分グラフを各頂点に付加する際の対称性を考慮して実験 1 では計 21 個，実験 2 では計 57 個のグラフ集合を生成した．図 4, 5 に実験 1, 実験 2 で使用した基本グラフを示す．ただし，現状では形状の類似性・非類似性の判断は主観的なものにとどまる．

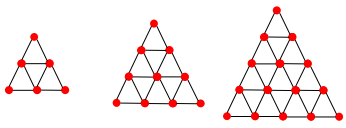

図 4: 実験 1 での基本グラフ

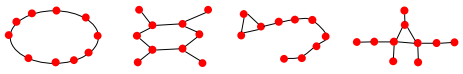

図 5: 実験 2 での基本グラフ

3.2 実験結果

2.1 節での 6 種類の SGScore 関数を用いて，式 (3) の値が最大となる K の値を用いて K-means 法により生成したグラフ集合をクラスタリングした．グラフの構造的特徴が頂点数，辺数および形状にそれぞれ反映されるならば，各実験において派生グラフは基本グラフが同じクラスタに分割されるのが望ましいと考えると，実験 1 では K=3，実験 2 では K=4 が最適と考えられる．各 SGScore 関数を用いてグラフスペクトルを生成した場合に IGRC が最大となった K の値を表 2 に示す．また，NGScore1 を用いた場合に，クラスタ数 K に伴う IGRC の変化を図 6, 7 に示す．

表 2: IGRC に基づく K の値

SGScore 関数	IGRC 最大となる K	
	実験 1	実験 2
SGScore1	9	2
SGScore2	2,3	2
SGScore3	8	17
NSGScore1	3	2
NSGScore2	2	2
NSGScore3	3	2

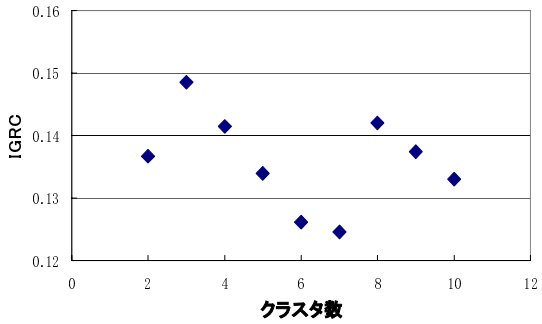

図 6: 実験 1 での IGRC の変化 (NSGScore1 を使用)

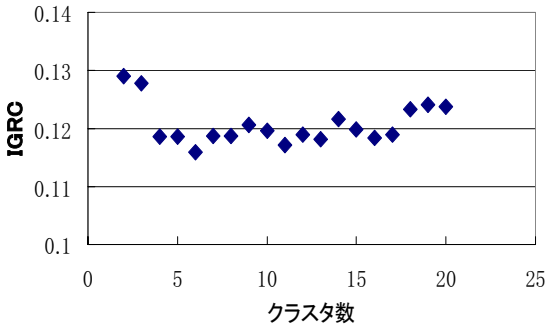

図 7: 実験 2 での IGRC の変化 (NSGScore1 を使用)

表 2 より，同じ点数を持つ部分グラフ数を部分グラフの総数で正規化した場合（NSGScor1, NSGScore2, NSGScor3 ）のほうが良好な結果となった．前述したように，グラフサイズの増加に伴い部分グラフの数は指数関数的に増加するため，同じ

基本グラフから生成した派生グラフのスペクトル間の距離は基本グラフのサイズが大きくなるほどより大きく（スペクトルの各次元での頻度の差が大きく）なるため別のクラスタに分割されやすくなる．頻度を含まれる部分グラフの総数により正規化することにより，この影響を軽減して同じ基本グラフから生成した派生グラフのスペクトル間の距離を近づけるのに役立ったと考えられる．

頂点の点数を部分グラフ内のみでの次数から捉えるSGScore2，NSGScore2は両実験でもK=2でIGRCの値が最大となり，望ましいクラスタリングが行えなかった．これは，もとのグラフを一種の文脈として利用して部分グラフを特徴付ける（ここでは点数の値に対応する）ほうが構造的特徴を反映するのに効果的であることを示唆しているとも考えられる．

実験1, 2ではグラフでの次数に基づいて頂点の点数を考えるNSGScore1, NSGScore3を用いた場合にIGRCが最大となるKの値は同じとなったが，実験2ではSGScore3を用いた場合以外は全てK=2でIGRCが最大となった．NSGScore1を用いてK=2としてクラスタリングした結果を図8に示す．図8に示すように，同じ基本グラフから生成した派生グラフは全て同じクラスタに分割され，クラスタ間にまたがって分割されることはなかったが，図8の左のクラスタ内，右のクラスタ内でのグラフを区別することはできなかった．図8の左のクラスタに分割されたグラフの基本グラフ同士は，環状のグラフの1つの枝を削除し，1つの枝を追加することで変換することができるため，グラフの書き換え操作の観点からは形状が類似するととらえることもできる．このためK=3を最適とみなすこともできるが，残念ながら現状で定義するIGRCの最大化という規範からはこの値を得ることはできなかった．このため，式(1), (2), (3)の定義を今後改良していく必要がある．

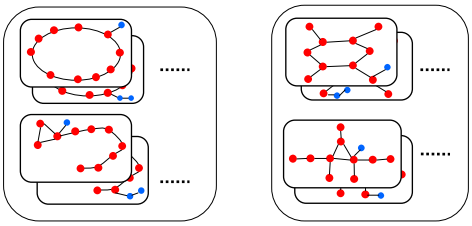

図 8: 実験 2 のクラスタリング結果

4. おわりに

本稿ではマイニングを通じて得られる大量のグラフを専門家が全て把握して理解することが困難であるという観点から，構造的類似性に基づくグラフクラスタリング手法を提案した．提案する手法ではTFS [Takahashi 98]と同様にグラフ中に含まれる連結部分グラフに基づいてその構造的特徴をスペクトルに変換し，変換したスペクトルをK-means法を用いてクラスタリングする．また，K-means法ではクラスタ数Kを事前に指定する必要があるという課題に対し，クラスタのエントロピーに基づく評価式を提案してKの値を自動的に決定することを提案した．人工データを用いた評価実験を通じ，頂点数や辺数に対する構造的な特徴を捉えられるものの，形状に対しては課題が残ることがわかった．このため，スペクトル間の類似度やエントロピーの評価式を改良する必要がある．また，実世界でのグラフ表現には有向グラフや頂点・辺にラベルが付いたグラフが用いられることも多いため，今後はラベルや有向辺を扱えるように拡張する予定である．

謝辞

本研究の一部は文部科学省科学研究費特定領域研究「情報洪水時代におけるアクティブマイニングの実現」(No.13131101, No.13131206) の補助による．

参考文献

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499 (1994).

[Inokuchi 00] Inokuchi, A., Washio, T., and Motoda, H.: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, in *Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 13–23 (2000).

[Kuramochi 01] Kuramochi, M. and G.Karypis, : Frequent Subgraph Discovery, in *Proc. of the 1st IEEE ICDM*, pp. 313–320 (2001).

[Matsuda 02a] Matsuda, T., Motoda, H., Yoshida, T., and Washio, T.: Mining Patterns from Structured Data by Beam-wise Graph-Based Induction, in *Proc. of The Fifth International Conference on Discovery Science*, pp. 422–429 (2002).

[Matsuda 02b] Matsuda, T., Yoshida, T., Motoda, H., and Washio, T.: Beam-wise Graph-Based Induction for Structured Data Mining, in *International Workshop on Active Mining (AM-2002): working notes*, pp. 23–30 (2002).

[Michalski 90] Michalski, R. S.: Learning Flexible Concepts: Fundamental Ideas and a Method Based on Two-Tiered Representaion, *Machine Learning: An Artificial Intelligence Approach*, Vol. 3, pp. 63–102 (1990).

[Palmer 02] Palmer, C., Gibbons, P., and Faloutsos, C.: ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs, in *Proc. of the KDD-2002* (2002).

[Quinlan 93] Quinlan, J. R.: *C4.5:Programs For Machine Learning*, Morgan Kaufmann Publishers (1993).

[Takahashi 98] Takahashi, Y., Ohoka, H., and Ishiyama, Y.: Structural Similarity Analysis based on Topological Fragement Spectra, *Adavances in Molecular Similarity*, Vol. 2, pp. 93–104 (1998).