ノード類似度分析による解析目的に応じたマルチグラフ構造推定

山本 康高†, 細見 格†, 久寿居 大†
NEC 共通基盤ソフトウェア研究所†

1 はじめに

Webや人脈など，世の中の様々なデータは，データ間の相互関係をエッジで表すことによりグラフとして表現できる．このときデータ間には潜在的に多種多様な相互関係があるため，グラフはマルチグラフとなる．

グラフ解析はグラフ化したデータの解析手法一般を指す．グラフ解析の結果は，解析の仕方以上に，グラフの構造，すなわちデータ間のどのような相互関係をエッジとしてグラフ化するかによって大きく左右される．したがって，解析目的に適した相互関係の種類を適切に選ぶことが重要となる．しかしながら，解析前に，その種類を明らかにすることはしばしば困難である．

そこで筆者らは，解析目的に適したグラフ構造を推定する機構を取り入れたグラフ解析の枠組みを検討している．本枠組みを構造推定型のグラフ解析と呼ぶ．本稿では，本枠組みに必要となるグラフ構造の推定方法について述べ，その基礎的方法であるノード類似度分析について説明する．

2 グラフ構造の推定

グラフ構造は，全ノード間に存在する様々な相互関係の強度を定量化したものと表現できる．ここで各相互関係の強度をエッジ強度と呼ぶ．

グラフ解析においては，解析目的に適したグラフ構造を解析することによって意味のある結果が得られる．しかしながら，そのようなグラフ構造を決定することは難しい．その理由は，解析目的に対するグラフ構造の適切さを図る基準が確立されておらず，グラフ構造を推定するためのモデルが構築しづらい点にある．望ましい解析結果が得られるエッジ強度を求めるというだけでは，本来データが有する相互関係を無視することになる．

そこで本稿では，データ間の相互関係が多種多様であることを配慮し，各種類の相互関係のエッジ強度に対して，その種類毎の重要度（以降，種別重要度）を乗算し補正したエッジ強度をグラフ構造と捉える．こうすることで，グラフ構造の推定を，「重視すべき相互関係の種類を種別重要度として定量化する問題」として扱う．すなわち，解析目的に適した種別重要度を求めることが課題となる．

筆者らは，このような種別重要度を求めるためのモデルとして下記の基準が重要と考えている．

精度： 解析者にとって妥当な解析結果が得られるグラフ構造を推定できること
透明度：精度を勘案し，必要な種類の相互関係によってグラフ構造を推定できること

これらは一般的なモデル構築において重視されるaccuracyとinterpretabilityに相当する[1]．本来，精度は解析目的を形式的に記述し，その目的と解析結果が一致しているかどうかを客観的に判断することが望ましい．しかしながら，解析目的は解析者ですら形式的な記述が困難であるため，解析者による主観的な妥当性の判断を解析目的の代替情報として用いる必要が生じる．ただし，精度のみを追求すると，解析対象となるグラフ構造に本来重要でない相互関係も含まれることが多く，結局，解析者は何を解析しているかが理解できなくなる．そのため，透明度の基準が重要となる．透明度は必要となるエッジの種類数や各エッジの重要性などが関わり，解析対象となるグラフ構造に対する解析者の理解を深めるために必須の基準である．

グラフ構造の推定に関わる従来技術としてリンク予測がある[2]．リンク予測は，既知である一部のグラフ構造を手がかりに未知の部分を予測することを課題とする[3]．リンク予測は，ノードの特徴量とエッジの有無との関係をモデル化するものであり，予測の精度を高める帰納的な推論が重要となる．リンク予測は，正解となるグラフ構造が得られること，透明度という観点を考慮していないこと，など本稿とは課題が異なる．

3 構造推定型のグラフ解析

構造推定型のグラフ解析の流れ図を図1に示す．

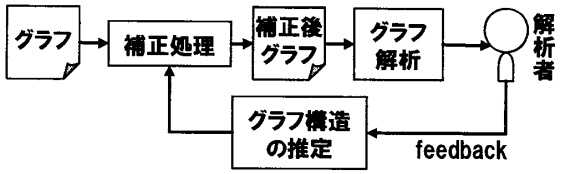

図1 構造推定型のグラフ解析の流れ図

グラフは，データ間の様々な相互関係が抽出されたものであり，相互関係の種類毎に全ノード間のエッジ強度が求められているものとする．

補正処理は，エッジ強度に対して種別重要度を乗算し，エッジ強度を補正する．この結果がグラフ構造となる．種別重要度の初期値は全てのエッジの種類に対して1を与えるものとする．

グラフ解析では，得られたグラフ構造に基づいて解析結果を出力する．グラフ解析は，解析結果

Estimation of Multi-Graph Structure using Node Similarity Analysis for Appropriate Data Analysis
†Kosuke YAMAMOTO, Itaru HOSOMI, and Dai KUSUI
Common Platform Software Research Laboratories, NEC Corporation

がグラフ構造に依存し，ノード単位で解析結果を出力するものとする．例えば，PageRank のようなランキングやノードの分類を行うクラスタリングなどである．

解析者は，解析しているグラフ構造の特徴および解析結果を見て，解析結果の妥当性を判断する．解析結果が妥当であると判断した場合にはグラフ解析を終える．そうでない場合には，妥当と判断できる一部のノード間の解析結果の関係性に対して確信度を与える．これが，形式化できない解析目的をシステムに伝えるための代替情報となる．

グラフ構造の推定は，確信度が付与されたノード間に張られているエッジの強度や種類に基づいて種別重要度を求める．種別重要度の導出は 2 節に述べたように精度および透明度を基準とする．その後，再度補正処理に移り，上記プロセスを繰り返す．

筆者らは，適切なグラフ解析を行うためには，このような枠組みが必要であり，これを実現するために，大きく分けて下記の検討課題を解決する必要があると考えている．

- マルチグラフを生成するための相互関係の抽出技術
- 解析者の解析目的を推定する技術
- 解析目的に適したグラフ構造を推定する技術

この中でも 3 つ目の技術に関わるノード類似度分析について 4 節で述べる．

4 ノード類似度分析

グラフ構造の推定のための第一歩として，透明度を重視した種別重要度の導出方法であるノード類似度分析について述べる．ここでの透明度は，必要となるエッジの種類を重要度として定量化したものと捉える．

ノード類似度分析は，解析者が妥当な解析結果と判定したノードに共通するエッジの種類の種別重要度を定量化するものである．これは，解析結果が妥当であったノード群には，その結果を導き出した共通するエッジがあるという考えに基づいている．そこで，本手法では，確信度の高いノード間にのみ出現しやすいエッジの種類は解析者が解析したい情報をうまく構造化していることを仮定し，確率モデルにより種別重要度を求める．

4.1 マルチグラフ

N_V個のノードがあり，N_C種類のエッジからなるマルチグラフを考える．エッジベクトル$\boldsymbol{e}^{ab}=\{e^{ab}_1, e^{ab}_2, \ldots, e^{ab}_{Nc}\}$は，a番目とb番目のノード間のエッジを表し，i種類目のエッジの強度がe^{ab}_iとなる．全てのノード間にエッジベクトルがあるため，エッジベクトルの数はN_E $(=N_V(N_V-1)/2)$ 個となる．

4.2 解析者による確信度の付与

解析者は，妥当と判断できる各ノードの解析結果の関係性に対して確信度r^{ab}を付与する．例としては，ノードのランキングにおける上下関係，クラスタリングにおける所属クラスタの同一性などがある．r^{ab}のaとbの組み合わせの和集合を$\boldsymbol{R}$とし，その要素数をN_Rで表す．

4.3 各種エッジの寄与率の推定

確信度が入力されたノードペアに基づいて，種別重要度を求める．種別重要度は，前述の仮定に基づき定量化する．すなわち各種エッジがN_R個のノードペアを何回繋いでいるかを求め，それが各種エッジの生起確率に対してどの程度偏っているかが種別重要度となる．

i種類目のエッジの生起確率p_iを(1)式により求める．δはe^{ab}_iが0なら0，非0なら1を返す関数である．

$$p_i = \sum_{a=1}^{N_V} \sum_{b>a} \delta\left(e_i^{ab}\right) \Big/ N_E \tag{1}$$

次に，N_R 個中何個のノードペアがいずれかのエッジで繋がれているかを(2)式で計算する．(2)式では確信度の高いノードペアを繋ぐエッジの種類を重視するために，確信度を重み係数として乗算している．

$$a_i = \sum_{(a,b)\in \boldsymbol{R}} e^{ab}_i r^{ab} \tag{2}$$

各種エッジが p_i で生起する二項分布に従うと仮定し，各種エッジの a_i の値が平均 $u_i=p_i\times N_R$，分散 $s_i=p_i\times(1-p_i)\times N_R$ の正規分布に従うものと近似する．また，(3)式に示す z_i は平均 0，分散 1 の標準正規分布に従う．

$$z_i = \left(a_i - u_i\right) \Big/ \sqrt{s_i} \tag{3}$$

a_i（≧0）が u_i よりも統計的にみて大きい場合，i 番目のエッジの種類は，集合 $\boldsymbol{R}$ において有意に発生しやすいエッジの種類であるといえる．そこで標準正規分布を$[-\infty:z_i]$の区間で積分した値を種別重要度 g_i とする．なお，a_i が 0 である場合は，$g_i=0$ とする．g_i は[0:1]の値をとり，本仮説に基づく種別重要度を定量化したものとなる．

5 まとめと今後の課題

解析目的に適したグラフ構造を推定する機構を取り入れたグラフ解析の枠組みである構造推定型のグラフ解析について述べ，グラフ構造の推定方法としてノード類似度分析について説明した．

予備的実験において，解析者が確信度を与えたノードペアにのみ生起しやすいエッジの種類の種別重要度が高くなることを確認している．ただし，現状は全てのエッジの種類を独立として扱っているため，今後エッジの組み合わせの重要性などを推定できるようにしていく必要がある．また，透明度の基準を整理し指標化すること，ならびに精度の基準の考慮等を行いグラフ構造の推定方式を確立していく予定である．

参考文献

[1] J. Casillas, O. Cordon, Francisco Herrera et al., "Interpretability Issues in Fuzzy Modeling," SPRINGER, 2003

[2] L. Gettor, C. P. Diehl, "Link Mining: A Survey," ACM SIGKDD Explorations Newsletter, Vol.7, No.2 pp. 3-12, 2005

[3] 鹿島，"ネットワーク構造予測"，人工知能学会論文誌，Vol.22, No.3, pp.344-351, 2007