ネットワーク構造の違いによるK-メディアン探索方法の特性評価*

永田　大†　入月卓也‡　伏見卓恭†　斉藤和巳†‡　池田哲夫†‡　武藤伸明†‡
静岡県立大学大学院経営情報学研究科†　静岡県立大学経営情報学部‡

1. はじめに

複雑ネットワークの分析において，優れたコミュニティ抽出（クラスタリング）技法の探求は，重要な研究課題である．そこで我々は[1]の論文において，K-メディアン問題として定式化したクラスタリングを高速化する，新たな手法を提案した．そこでの評価実験では，対象としたネットワークにより，各種手法の性能差が顕著に現れるケースや，逆に僅かな性能差しか現れないケースがあった．すなわち，ネットワーク構造に依存して各種方法の得手不得手が起こると推察され，このような関係の解明も重要な研究課題と考える．

本研究では，各種クラスタリング手法の特性を精緻に評価するため，平均次数，平均クラスタ係数，平均ノード間距離を変化させたネットワークを，WattsとStrogatzの手法[2]を用いて人工的に生成し，既存の方法と[1]で提案した方法の特性を評価する．

2. K-メディアン問題と解法

2.1 K-メディアン問題

K-メディアン問題とは，クラスタ中心が有限ピボット集合に限定され，オブジェクトとピボット間の類似度に基づいてクラスタリングを行う問題である．

本研究においては，ネットワーク上のノード群をオブジェクト集合Xとし，それら全てがピボット候補集合Yであるという設定のもとで，各ノード間の類似度ρに基づきクラスタリングし，以下の式で表される関数fの値を最大にするピボット集合$Z_K \subset Y, |Z_K| = K$を求める．

$$f(Z_K) = \sum_{x \in X} \max_{y \in Z_K} \{\rho(x,y)\}$$

ここで関数fの値を解品質と定義し，また，類似度ρはノードxとノードy間のネットワーク上での最短距離（最短パス長）$g(x,y)$を求めたとき，$\rho(x,y) = 1/(1+g(x,y))$と定義する．

* The Evaluation of Clustering Methods on Watts-Strogatz Networks
† Graduate School of Administration and Informatics, University of Shizuoka
‡ School of Administration and Informatics, University of Shizuoka

2.2 解法

K-メディアン問題における代表的な解法として，分割改善法 (DI)，貪欲改善法 (GI)，および，局所改善法 (LI) が挙げられる．DIは，各ピボットとの類似度でオブジェクト集合をクラスタ分割し，クラスタ毎に最良ピボットの選択を繰り返す，K-平均法[3]と類似した手法である．GIは，与えられたピボット候補集合から，類似度を最大にするピボットを順次選定，追加するという処理を繰り返し，クラスタリングを行う手法である．LIは，すべてのオブジェクト集合を対象にして，各ピボットを順番に最良ピボットに置き換えてクラスタ集合を求める手法である．

LIを用いれば，他の2手法と比較して，一般に望ましい解品質を安定して求められるが，その計算量は大幅に増大する傾向がある．これに対し我々は，このクラスタリング問題がサブモジュラ性と呼ばれる数理構造を持つことを示すとともに，この構造を利用した遅延評価(LE)と呼ばれる手法の導入により，局所改善クラスタリングを高速化する新たな手法 (LILE) を提案した[1]．

3. 評価実験

3.1 人工ネットワークの生成

人工ネットワークの生成方法を以下に示す．

まず次数$d = \{8,12,16\}$，ノード数10000の円環状のレギュラーグラフを生成する．そして，そのレギュラーグラフと，それをWattsとStrogatzの手法[2]を用いて確率$p = 2^{-n}(0 \le n \le 13, n$は整数$)$で張り替えたグラフを生成し，以降の実験に用いる．なお，張り替えを行ったグラフは連結である．

3.2 特性評価

前節で生成したグラフについて，各種手法を用いてクラスタリングをし，特性を評価する．

詳細には，d, p, Kが与えられたとき，まず次数d，確率pで張り替えたグラフに対して，ピボット数KのクラスタリングをDI，GI，LILEを用いて各5回ずつ実行し，その解品質の平均を求める．このとき，GIは試行回数によらず解品質が一定であることから，これを評価基準とする．

具体的には，GI の解品質を 1 としたときの，各種手法の解品質の平均の値を求める．これを 3 つの手法について，Kを 2 から 30 まで変化させて行い，平均をそれぞれ求める．以下では，これを「解品質(対貪欲法比)」と示すこととする．この一連の処理を，生成した全てのグラフについて行い，特性を評価する．

3.3 実験結果

図 1 から図 3 に実験結果を示す．ここで，参考文献[2]と同様に，レギュラーグラフのときを 1 とした各グラフのクラスタ係数を実線，平均ノード間距離を破線で表す．

DI については，レギュラーグラフ $(p = 0)$ においては，LILE と同等の解品質が得られたが，ランダムグラフ $(p = 1)$ 付近では，同法との差が開き，解品質が低下する傾向が見られた．

LILE においては，どのようなグラフ構造においても，GI と同等かそれ以上の解品質が得られた．特にスモールワールド性（クラスタ係数がレギュラーグラフのように高く，平均距離がランダムグラフのように低い状態）を持つグラフにおいては，良好な解品質が得られる傾向が見られた．なお，この実験の範囲では，次数による顕著な解品質の変動は見られなかった．

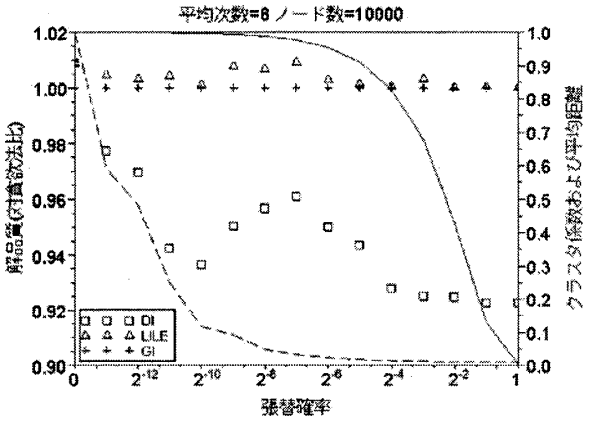

図 1 平均次数=8 ノード数=10000 における実験結果

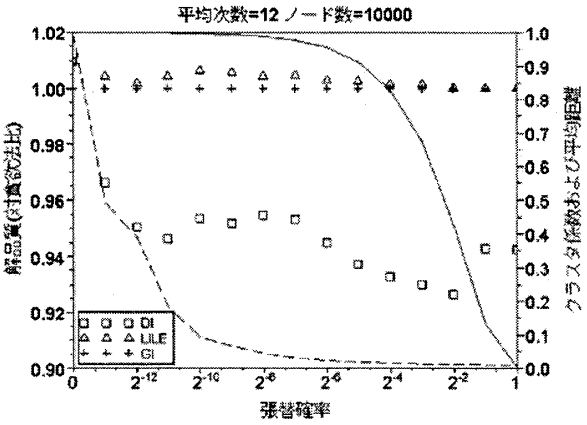

図 2 平均次数=12 ノード数=10000 における実験結果

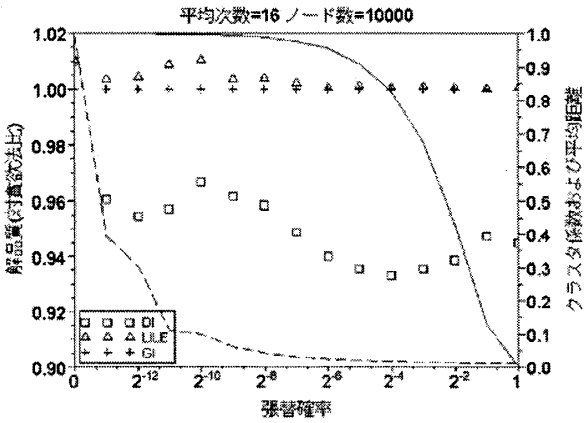

図 3 平均次数=16 ノード数=10000 における実験結果

4. おわりに

本研究では，次数や張り替え確率を細かく変えたネットワークを生成し，各種クラスタリング手法の特性を評価した．今回の実験では，我々の提案手法である LILE は，スモールワールド性を持つグラフにおいて良好な解品質が得られるという結果となった．今後は，何故スモールワールド性を持つグラフでは良好な結果が得られたのか，その要因について研究を行っていく．

参考文献

[1]斉藤和巳，武藤伸明，池田哲夫ほか：遅延評価導入による局所改善クラスタリング法の高速化，情報処理学会論文誌，数理化モデルと応用(TOM)，Vol.3，No.1 (in press).

[2]Watts, D.J. and Strogatz, S.H. : Collective dynamics of 'small-world' networks, Nature 393, pp.440-442 (1998).

[3]Duda, R.O. and Hart, P.E. : Pattern classification and scene analysis, John Wiley & Sons (1973)