大規模ネットワークに適したクラスタリングシステムの開発

西澤 慎紘† 木村 昌臣†
芝浦工業大学工学部情報工学科†

1.研究背景と目的

複雑ネットワークのクラスタ分割手法としてNewman らによるモジュラリティが最大になるよう各ノードのクラスタへの割り当てを決める手法が近年注目されている。さらに Reichardt[1]らはモジュラリティの考えを応用し、エッジで繋がっているノード同士は同じクラスタへ、繋がっていないノード同士は違うクラスタへ分類されると評価が良くなるよう設計した評価関数を物理学におけるスピングラス系の Potts モデルハミルトニアンとして定義し、シミュレーテッド・アニーリング法(SA)による最適化を提案している。しかしSA は最適解への収束が遅いため、本研究ではより早い収束が期待される遺伝的アルゴリズム(GA)を用いることにする。加藤ら[2]は GA を用いた個体群のクラスタリング手法として、染色体を多値で表現し、それに対する遺伝的操作を提案しているが、その手法ではクラスタの分割数はパラメータとして予め与える必要がある。そこで本研究では、得られるべきクラスタの数は解析後に初めて与えられるべきと考え、クラスタ数を前提としない拡張された GA の提案を行う。

2.研究内容

SA は解空間を最初は広く探索し、時間とともに探索範囲を狭めていく手法であるが、評価値の変化が激しい解空間では局所解に陥りやすく、また探索の初期ではランダムな探索が多くなるため収束も遅い。そこで、局所解に陥りにくい多点探索である GA を SA に替えてこの問題に適用することにした。以下に本研究が対象とする問題に適用可能な拡張した GA の詳細を述べる。拡張 GA は「パラメータの設定」「初期集団の設定」「交叉」「近傍探索」「淘汰および突然変異」の各処理からなる。

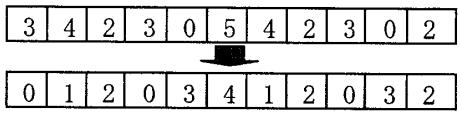

図1 遺伝子の置き換え

The development of the clustering system suitable for the large-scale network
†Masahiro Nisizawa, Masaomi Kimura
†Shibaura Institute of Technology

2.1.染色体表現

GA を本研究に適用すると、加藤らによる手法と同様に分割状態を表す染色体を表現できる。この手法では、染色体の遺伝子座はノード番号と対応し、遺伝子はクラスタ番号と対応する。なお、最大分割数はノード数となる。このように多値で遺伝子を表現すると一つの分割状態に対して複数の表現が表れるため、それを統一するクラスタ番号標準化処理が必要となる。クラスタ番号標準化の具体的な処理は図 1 のように行う。先頭の遺伝子座の対立遺伝子と同一の対立遺伝子をすべて 0 に置き換え、次にその隣の遺伝子座の対立遺伝子が 0 に置き換え済みでなかったら、その対立遺伝子をすべて 1 に置き換える、という処理をすべての対立遺伝子が置き換えられるまで繰り返す。なお、この処理は初期集団を生成した後と遺伝的操作で染色体に変更があるごとに毎回行う。加藤らは、ランダムに選ばれた遺伝子座を始点として同様の処理を行っているが、この方法ではクラスタ番号とノードの対応関係がゆるく、異なるクラスタのノードが次世代で不用意に同一のクラスタとしてラベル付けされてしまうおそれがある。そこで本研究では、遺伝子座の若い方から順に対立遺伝子を置き換える処理のみを行うことにした。

2.2.交叉

交叉は以下の手順で行う。

① 交叉を行う前に両親のペアに偏りが生じないよう集団内の染色体の順番をバラバラにする。
② 両親のペアを決定し、一点交叉を適用する。この際、通常のGAでは交叉に用いた両親は消滅するが、本研究では適応度の高い染色体が消滅する可能性を排除するため、両親をそのまま集団内に残すことにする。よって、一回の交叉で二つの子孫が生成されるので、集団数は交叉を行うたびに二つずつ増えることになる。

2.3.近傍探索処理

本研究では遺伝子を多値で表現しているため、通常の交叉のみでは収束に膨大な時間が掛かる。これは、従来の GA の 0,1 での遺伝子表現よりも組み合わせ数が非常に多いためと考えられる。

そこで本研究では、近傍探索処理を追加し、収束時間を早めることを目指した。具体的には、あるノードのクラスタ割り当てを変更する際に、エッジでつながっているノードと同じクラスタへ割り当てるようにする。手順はまず、遺伝子座をランダムに決定する。その遺伝子座に対応するノードをaとすると、aとエッジで繋がっているノードをランダムに決定し、決定したノードが含まれるクラスタ番号へaを割り振る。図2のネットワークを例として、この処理の追加の効果を図3に示す。図3より、近傍探索を追加すると近傍探索をしない処理よりも少ない世代数で収束しており、この処理が有効であることが分かる。

2.4.淘汰

交叉により増えた染色体を元の集団と同数にまで減らす処理を行う。淘汰の仕方は、集団内の染色体を適応度順にソートして、元の集団と同数になるまで適応度の高いものから順に次世代に残すものとする。

通常GAでは淘汰とは別に突然変異を起こす処理を行うが、本研究で提案する方法では、遺伝子配列が全く同じ染色体ができた場合、片方を消滅させ集団内からランダムに一つ染色体を選択し、その染色体に突然変異を起こさせる処理を追加している。この処理により、集団内は常に多様性を維持できる。突然変異の方法は、突然変異を起こす遺伝子座をランダムに決定し、その遺伝子座に対して(0～ノード数－1)までの一様整数乱数を用いて対立遺伝子を決定する。

3.実験

拡張GAとSAを比較するため、GAでは1世代を1ステップ、SAでは1回の遷移を1ステップとし、同じステップ数での収束の度合いを計測する実験を行った。実験には、10個のノードが一つのまとまりになり、数珠つなぎに繋がっているネットワーク(図2)を作成して使用する。ノード数は100、500、1000の三種類とした。

4.結果・考察

実験の結果、拡張GAで収束できる世代数では、SAは収束し切らないことが示された。図4は100ノードの結果であり500、1000ノードでも同様の結果が得られた。このことから、拡張GAはSAよりも少ないステップ数で収束に至ることが分かる。1ステップ当たりの処理時間はSAの方が早いため、単純にステップ数での収束度合いを比較して優劣を判じることはできないが、拡張GAの1ステップ当たりの処理時間を早める工夫を施すことで拡張GAの有効性を高められると考えられる。

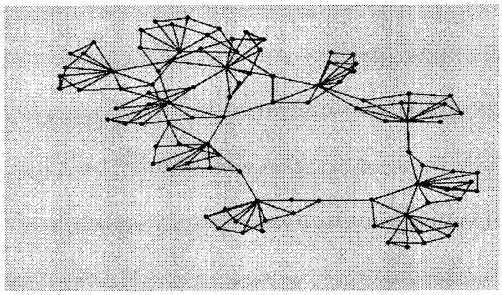

図2　本研究で利用したネットワーク(ノード数100)

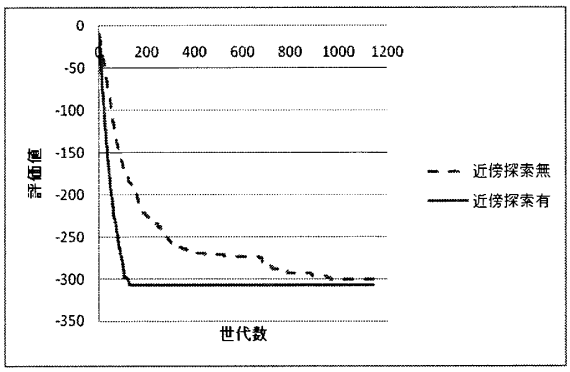

図3　近傍探索の有無による評価値の変化

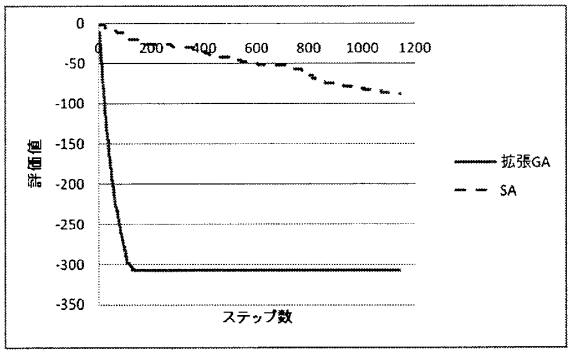

図4　拡張GAとSAの比較実験結果(ノード数100)

5.まとめ

本研究では、クラスタの分割数を事前に定めずにネットワークをクラスタリングする拡張したGAの手法を提案した。SAとの比較では、拡張GAはSAよりも少ないステップ数で収束に至ることが示された。今後の課題として、1ステップ当たりの処理時間を早めることが挙げられ、その方法としてGAの並列性の高さを利用し、分散処理を行うことを考えている。

参考文献

1) Joerg Reichardt, Stefan Bornhold : Statistical mechanics of community detection, Physical Review E, vol. 74, 016110, pp.1-14 (2006).
2) 加藤常員,小沢一雅:遺伝的アルゴリズムを用いた非階層的クラスタリング,情報処理学会論文誌, Vol.37, No.11, pp.1950-1959 (1996).