多目的GAによるクラスタリングの初期化アルゴリズムの検討

千田　智治

1　はじめに

Web 上の行動分析を行なうと，Web サイト内でのユーザへの適切な情報提示，ユーザの行動を考慮した Web 広告配信システムなどが行なえる．膨大なユーザの行動データを分析するには，クラスタリング手法が必要となる．クラスタリングとは，類似したデータ同士をいくつかのグループに分割する，データマイニングの一種である．

クラスタリングでは，K-means や凝集法などのアルゴリズムがよく用いられているが，多目的 GA を用いた多目的クラスタリング (Multiobjective clustering with automatic k-determination:MOCK) もある．MOCK では，K-means 法や凝集法などと比較して高いクラスタリング性能を示す [1]．

しかしながら，MOCK の初期化時における最小全域木生成での計算コストが問題となる．計算コストを減少させるアルゴリズムを考案したが，近傍の値によっては最小全域木が生成できないという問題が残る．本報告では，最小全域木生成において更に改良を加え，どの近傍の値でも最小全域木を作成できるアルゴリズムを考案した．

2　多目的クラスタリング (MOCK)

MOCK は，2004 年に J.Handl と J.Knowles が提唱したアルゴリズムである．MOCK アルゴリズムは，多目的遺伝的アルゴリズムを用いて大域的な評価指標と局所的な評価指標を同時に最適化を行なう．

2.1　評価関数

MOCK では，クラスタのコンパクト性に関する評価関数である Overall Deviation とデータ点の連結性性に関する評価関数 Connectivity を用いる．この 2 つの評価関数を同時に最適化することでクラスタリングを行なう．この評価関数より，解集合内のクラスタ数の多様性を維持したまま解の質を高めることができる．

2.1.1　Overall Deviation

Overall Deviation は大域的な評価指標である．Overall Devitaion は各クラスタの中心点 μ_k から同じクラスタ内の各データ i までの距離の総和で定義される．この指標を最小化することにより，コンパクトなクラスタが生成される．この評価指標の数式を以下に載せる．

$$Dev(C) = \sum_{C_k \in C} \sum_{i \in C_k} \delta(i, \mu_k) \tag{1}$$

2.1.2　Connectivity

Connectivity は局所的な評価指標として用いられる．Connectivity はあるデータ i の 1〜L 番目までの近傍 (その近さ順を j とする) が，データ i と異なるクラスタに存在する場合，$1/j$ のペナルティを与えるという指標であり，全てのデータについてのペナルティ値の総和が Connectivity 値となる．Connectivity も最小化され，それにより近傍のデータ同士がより高い確率で同じクラスタに存在することになる．この評価指標の数式を以下に載せる．

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_i(j)},$$
$$where\ x_{r,s} = \begin{cases} \frac{1}{j} & if\ \nexists C_k : r, s \in C_k \\ 0 & otherwize \end{cases} \tag{2}$$

2.2　グラフベースの個体表現

MOCK ではグラフベースの個体表現を行ない，任意のクラスタ数を持つ個体を同時に表現可能である．また，クラスタ数が未知の問題において，最適なクラスタ数を特定する機能を持っている．Fig. 1 に，個体表現を示す．

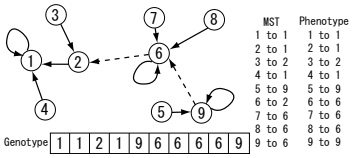

Fig.1　グラフベースの個体表現 (出典：自作)

1 つのノードは 1 つのデータに対応している．両端のデータが同じクラスタに含まれるものをエッジと呼ぶ．グラフベースの遺伝子表現は初期化や突然変異などのオペレータが適用しにくいが，同時に複数のクラスタ数をもつ個体を表現できる．

この個体表現では，各データを有向グラフの各ノードとして捉え，ノード間のエッジは，両端のデータが同じであるクラスタに属することを意味する．遺伝子表現の際，どのエッジを対応させるかを決定するかを調べるために矢印の向きが意味を持つ有向グラフを用いている．

2.3　最小全域木

MOCK の初期化には，与えられたグラフより最小全域木 (Minimum Spanning Tree:MST) を生成する．最小全域木とは，エッジが最小の数で生成できるグラフのことをいう．最小全域木を Fig. 2 に示す．

与えられたデータセットに対して完全な MST を作成する．初期個体の i 番目の個体は $(i-1)$ 番以上に長いエッジを取り除いて作成される．このように，グラフベースの個体表現と MST を用いた初期化アルゴリズムより，任意のクラスタ数を持つ質の高い初期個体を生成することが可能となる．

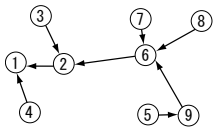

Fig.2　最小全域木 (出典：自作)

2.4　問題点

MOCK では，大規模データへの適用の際には，初期化時の最小全域木生成にかかる計算コストが問題となる [2)]．また，計算コストが軽減するアルゴリズムの考案を行なったが，近傍の値によっては初期化時に最小全域木生成が行なえないという問題があった．

3　初期化アルゴリズム

クラスタリングを行なう際，初期化時にすでにノード間の距離が短い最小全域木を作成できれば，初期化時の最小全域木によりデータ分析は依存するため，その後のクラスタも効率よく行なえる．現行のアルゴリズムと比較して現在考案したアルゴリズムを以下に示す．

3.1　現行のアルゴリズム

Fig. 3 に示すように 0〜9 までの 10 個のノードが存在すると仮定する．

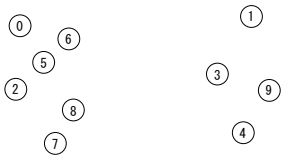

Fig.3　10 個のノードの設定 (出典：自作)

このツリーのルートを 0，近傍を 3 とする．ルートから最小全域木を作成する際，ルートの近傍は 2，5，6 の 3 つとなる．この 3 つの中で，ルートの 0 と最も短いエッジとなるノードを，次の最小全域木のノードとする．0 と最も近い 5 がルートと繋がる次のノードとなる．これを Fig. 4 に示す．

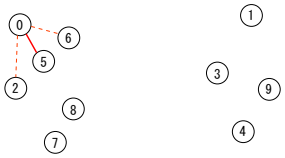

Fig.4　最小全域木の次ノードの決定 (出典：自作)

次ノードを見つけてはその近傍をとり，その近傍の中で最も短いエッジを持つノードを選択する．これを繰り返したとき，近傍の値の関係で全てのノードを探索することができない場合がある．Fig. 3 のように左側と右側に離れたデータがある場合，ルート 0 からツリーを生成してもノード 7 で生成がとまってしまう．Fig. 5 に示すとおり，この状態では初期化時の最小全域木は生成できない．

3.2　考案した初期化アルゴリズム

新しく考案した初期化アルゴリズムでは，初期化時に最小全域木が作成できなくなることはない．Fig. 6 に示すように，最初に全ノードの近傍を求める．

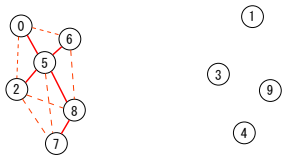

Fig.5　最小全域木が生成できない場合 (出典：自作)

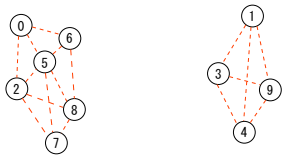

Fig.6　全ノードの近傍 (出典：自作)

次に，最短エッジから順にノードを結んでいく．このとき，ループにならないように注意する．全てのエッジを見ていくと Fig. 7 に示すように，左側と右側に二つのツリーが生成される．

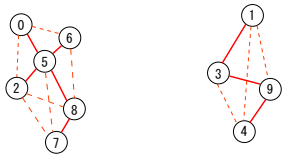

Fig.7　全ノードの近傍 (出典：自作)

最後に，この二つのツリーをそれぞれクラスタとみなして，最短エッジで結合する．これを Fig. 8 に示す．これで，初期化時に最小全域木が生成できなくなることはない．また，最短エッジを順に選択してツリーを生成し，類似性の高いノード同士をグループ化しているため，精度の高いクラスタリングが行なえる．

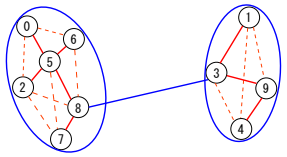

Fig.8　最小全域木の生成 (出典：自作)

4　まとめ

クラスタリングとは，類似するデータ同士をいくつかのグループに分割するデータマイニングの一種であり，その評価指標として，K-means 法や凝集法などがある．K-means 法や凝集法の欠点を補った最適なクラスタ手法に，多目的 GA を用いた MOCK がある．また，MOCK では大規模データを扱う際，初期化時において MST が生成できないという問題がある．そのため初期化時の際に，どのような近傍の値でも MST を生成できるアルゴリズムを考案した．

参考文献

1)　Julia Handl and Joshua Knowles. Multiobjective clustering with automatic determination of the number of clusters, Technical Report No. TR-COMPSYSBIO-2004-02, UMIST, Department of Chemistry, August 2004.

2)　多目的クラスタリングにおける最小全域木を用いた初期化の問題点とその解決策としての引き戻し方法の検討
http://mikilab.doshisha.ac.jp/dia/research/report/2006/0712/005/report20060712005.html