4N-9

多目的GAによるクラスタリングの検討 -初期化アルゴリズムの検討-

廣安 知之 † 三木 光範 † 千田 智治 †
† 同志社大学工学部

1 はじめに

Web 上にはいくつかの広告配信に関する手法が存在し，サイト内の関連コンテンツを自動でリンクして繋げていくコンテンツマッチング広告や，検索エンジンの検索結果によって，ユーザの興味のある情報を表示する広告などがある．このようなユーザの嗜好に合わせた広告を配信するには，Web 上でユーザの行動を分析する必要がある．このような分析を行なう為には Web マイニングが重要であり，本研究では，その中でもクラスタリングに注目する．

クラスタリングとは，類似するデータをグループ化するデータマイニングの一種であり，ユーザの行動別やサイト別に Web データを分類することができる．クラスタリングには，同一クラスタ内のデータは互いに類似する均質性と，異なるクラスタ内のデータは互いに異なる分離性の，2 つの評価指標がある．この 2 つの評価指標を考慮して，与えられたデータをどのように分割し，いくつにクラスタ数を分割するかを決定する．クラスタリングには一般的に K-means 法や凝集法が用いられるが，これらの手法には欠点が存在する．本研究では，他の多くの手法と比較して高いクラスタリング性能を示し，かつクラスタ数自動決定メカニズムを持つ多目的クラスタリング (Multiobjective clustering with automatic k-determination:MOCK) に着目している．

MOCK では初期化時にデータ間をエッジで結び最小全域木 (Minimum Spanning Tree:MST) を生成する．しかし MST の計算コスト及び使用するメモリ量は大きく，大規模データへの適用は容易ではない．よって本論では，この問題を解決して，少ない計算コストとメモリ量で MST を生成するアルゴリズムを提案する．

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 [1] であり，多目的遺伝的アルゴリズムを用いて，精度の高いクラスタリングが行なえる．MOCK は，あらかじめクラスタ数を指定することなく最適なクラスタを発見することが可能であり，図 1 のようにグラフベースの個体表現を行なうことで，任意のクラスタ数を持つ個体を同時に表現可能である．

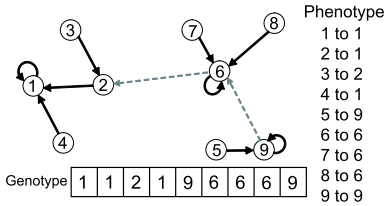

図 1: グラフベースの個体表現

各ノードはそれぞれデータに対応しており，エッジは両端のデータが同じクラスタに含まれることを意味する．また，各遺伝子と各ノードを 1 対 1 に対応させるため，各ノードから出るエッジが 1 本に限定された有向グラフを用いる．

MOCK では，類似するデータ同士は近接するように表現されており，初期化時にデータ間をエッジで結び全域木を生成する．初期化時に生成した全域木のエッジを削除することで任意の数のクラスタが生成できる．このような操作を行なうことで，類似したデータ同士でエッジが繋がり，均質性を持つ初期解が生成できる．

しかしながら，大規模データへの適用の際に，初期化時の MST 生成にかかる計算コスト及びメモリ量が問題となる [2]．本研究ではこれらを考慮したアルゴリズムを用いて MOCK の初期化を行なった．

3 提案アルゴリズム

Handl らの MOCK の MST 生成アルゴリズムは Prim 法 [3] を用いたアルゴリズムである．Prim 法とは，あるエッジを最初にランダムに選択し，そのエッジに近接するエッジの中で最短なものを選択して段階的に 1 つの全域木を拡大していく手法である．Prim 法で使用メモリを削減するには，初期化時に保存する近傍を減らせば良いが，メモリを減らすと MST が生成できない場合がある．そこで提案アルゴリズムは，Kruskal 法 [4] と最短距離法を用いて，データサイズや近傍数に関わらずメモリ使用量を抑えて MST を生成可能とした．Kruskal 法は，閉路にならないようにグラフに含まれる全てのエッジを結ぶエッジの中で最短エッジを選択していき，部分的に多数の全域木を拡

Examination on Clustering of Multiobjective GA
- Examination on Initial Algorithm -
† Mitsunori MIKI(mmiki@mail.doshisha.ac.jp)
† Tomoyuki HIROYASU(tomo@is.doshisha.ac.jp)
† Tomoharu SENDA(tsenda@mikilab.doshisha.ac.jp)
Department of Knowledge Engineering and Computer Science, Doshisha University (†)
Undergraduate Student Doshisha University (††)
1-3 Miyakodani, Tatara, Kyotanabe, Kyoto 610-0321, Japan

大しながら最終的に 1 つの全域木を生成する手法である．提案手法のアルゴリズムは以下のようになる．

3.1 MST の生成手順

0〜9 までの 10 個のノードをセットする．最初に全ノードにおける近傍との距離を求め，Kruskal 法により最短エッジを順に選択しノード間を結ぶ (図 2)．

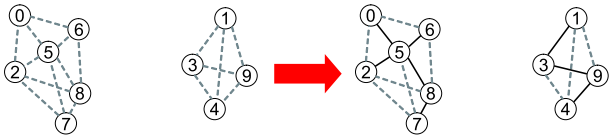

図 2: Kruskal 法を用いた全域木の生成

エッジを結び終わると左右に 2 つの全域木が生成できる．この 2 つの全域木をそれぞれクラスタとみなして，最短距離法より最短エッジでクラスタ間を結合する (図 3)．最後に有向グラフにすることで全ノードを用いた MST が生成できる．

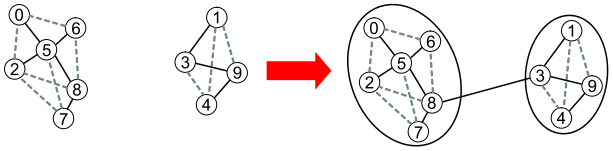

図 3: 最短距離法を用いた MST の生成

4 実験

4.1 実験内容

MOCK の初期化時における MST 生成ではメモリの問題があるため，メモリを考慮したアルゴリズムを考案する必要がある．実験では，メモリ量の比較とデータ量の増加に伴う MST 生成処理の成否を検討する．

実験結果には，データサイズが 1000，5000，9990 の Square1 というテストデータセット [1] を用いた．比較手法には，Prim 法を用いた手法，Prim 法を改良してメモリ制限をした手法，Kruskal 法と最短距離法を用いた提案手法の 3 つを用いた．Prim 法では，近傍数はデータ数と等しく，改良した Prim 法と提案手法は，共に近傍数を 25 とする．

4.2 実験結果

3 つの手法を用いて，データサイズ 5000 の時にメモリ領域がどれだけ使用されているかを検討した．その実験結果を，図 4 に載せる．

図 4 より，Prim 法では，他の 2 つの手法と比較して膨大なメモリを使用していることがわかる．また，改良した Prim 法と提案手法は，共に近傍数が 25 であるため保持するデータが少なくメモリ使用量を抑えている．メモリ量の問題より，実験結果を得ることができない場合があるため，データサイズ 1000，5000，9990 の時に 3 つの手法が実験結果を得ることができるかを表 1 に示す．

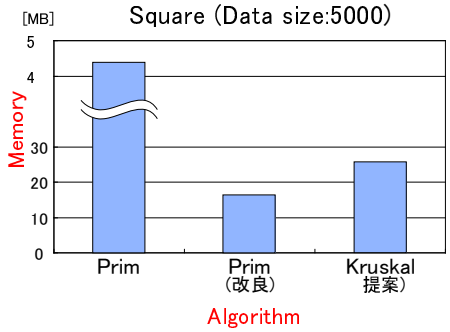

図 4: メモリ使用量の比較

表 1: 実験結果の比較

	1000	5000	9990
Prim	○	○	×
Prim(改良)	○	○	○
Kruskal(提案)	○	○	○

Prim 法では，データサイズが増加するにつれてメモリ制約の為に実験結果を得ることができなかった．

4.3 考察

Prim 法が膨大なメモリ量を使用するのは，近傍がノード数と等しい為，保持するデータが多くなりメモリ量を大幅に使用しているためである．また，メモリ制限をした提案手法を用いることで，データの規模が増大しても MST の生成が行なえると思われる．

5 まとめと今後の展望

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る．クラスタリングでは一般的に K-means 法や凝集法が用いられるが，それらには欠点がある．本研究では他の手法と比較して高いクラスタリング性能を示す多目的 GA を用いた MOCK に注目した．しかし，MOCK ではデータサイズや近傍数に依存して，初期化時に MST が生成できない問題がある．そのため初期化時にメモリ領域を考慮した MST 生成アルゴリズムを考案した．今後は，効率的に初期化時の MST 生成ができるように，Prim 法などを更に改良したアルゴリズムを検討する．

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] 三木 光範，廣安 知之，真武 信和．多目的遺伝的アルゴリズムによるデータクラスタリング，第 16 回設計工学・システム部門講演会講演論文集 pp.317_320，2006.

[3] R. C. Prim. Shortest connection networks and some generalizations. Bell System Technical Journal，36:1389-1401，1957.

[4] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. of the American Mathematical Society，7:48-50，1956.