社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (10)
2006／6／15

複数の最小木を考慮した確率的進化計算による遺伝子データ・クラスタリング

波平 光洋† 名嘉村盛和† 岡崎 威生† シバスンタランスハルナン††

† 琉球大学 工学部 情報工学科
〒 903-0213 沖縄県中頭郡西原町字千原 1 番地
†† 有限会社アクシオヘリックス
〒 135-8073 東京都江東区青梅２丁目４５番 タイム２４ビル３階
E-mail: †namihei@ads.ie.u-ryukyu.ac.jp, ††{morikazu,okazaki}@ie.u-ryukyu.ac.jp,
†††ssivasundaram@axiohelix.com

あらまし 本研究では複数の最小木を考慮した確率的進化計算によるクラスタリング法を提案する．提案手法は最小木に基づくクラスタリングを応用したものである．最小木を用いてクラスタリングを行う場合，解空間が過度に縮小されるために良いクラスタリング結果を得られない場合がある．そこで，複数の最小木を基とした分析を行うことで解空間を拡張し，クラスタリング精度の向上を目指す．また，大腸菌の反応データを用いて提案手法の有効性の確認を確認する．

キーワード クラスタリング，最小木，確率的進化手法 (StocE 法)，K-means 法

Stochastic Evolutionary Computation based on Multiple Minimum Spanning Trees for Gene Data Clustering

K.Namihira†, M.NAKAMURA†, T.OKAZAKI†, and S.Suharnan††

† Department of Infomation Engineering, University of Ryukyus
1 Senbaru, Nisihara, Nakagami, Okinawa, 903-0213 Japan
†† AxioHellix Co.
Time24buliding 3 floor, Ome, Koto, Tokyo, 135-8073 Japan
E-mail: †namihei@ads.ie.u-ryukyu.ac.jp, ††{morikazu,okazaki}@ie.u-ryukyu.ac.jp,
†††ssivasundaram@axiohelix.com

Abstract This paper considers gene function analysis of coli bacteria and presents an algorithm using stochastic evolutionary computation based on multiple minimum spanning trees. The idea of using the minimum spanning tree is to reduce drastically the search space. However, we often can lose good solutions because of the reduction. Therefore, we try to overcome this weak point to use multiple minimum spanning trees. The stochastic evolutionary computation is also effective for this approach. Experimental evaluation shows efficiency of our method. The possibility that a high quality clustering is obtained can be improbed by considering multiple minimum spanning trees. This resarch aims to obtain effective clisters in actual data of coli bacteria.

Key words Clustering, Minimum spaninig tree, Stochastic Evolutionary Computation (StocE), K-means

1. はじめに

近年，遺伝子の機能及び遺伝子間の相互関係を解明する研究が注目されており，機能解析を行ったデータに基づいた新薬の設計などが盛んに行われている．遺伝子の機能解析とは遺伝子の反応データ等から遺伝子の分類を行い，未知の遺伝子機能を解析することである．本研究では遺伝子機能解析データとして大腸菌を用いる．大腸菌は生物の生命活動に共通して，基本的な役割を果たす遺伝子の多くを含んでおり，遺伝子解析において重要視されている生物の１つである [1]．

最小木によるクラスタリング [2] は，クラスタリングの対象となるデータを最小木を用いて表現することで解空間の縮小を

目指したものであったが，あまりに解空間を縮小することで良質な解を求めることができない可能性がある．複数の最小木に基づくクラスタリングは，準最小木を含めた複数の最小木を基にクラスタリングを行うことで欠点の解消を試みたものである．

本稿では実際に遺伝子機能解析の手段として得られたデータに対し，提案手法を用いることで，現場での提案手法の有効性を確認することを目的とする．

2. 基礎概念

2.1 遺伝子機能解析

遺伝子機能解析の目的は遺伝子の反応データ等から遺伝子の分類を行い，未知の遺伝子機能を解析することである．本研究では大腸菌の機能解析データを用いる．今回用いるデータは，それぞれ違う遺伝子を一つずつノックアウトした大腸菌を様々な環境の培地で培養し，その呼吸量を計測したものである．このデータを時間毎の呼吸量で分類することで，どの遺伝子がどのような環境に適応する際に必要とされているのかを推測し，未知の遺伝子機能を解析することを目指している．

2.2 遺伝子データ・クラスタリング

遺伝子データ・クラスタリングは，遺伝子の発現状態や，反応データ等を時系列のデータとして数値化し，データのパターンを基に遺伝子の分類を行う手法である．遺伝子をクラスタリング（分類）することによって，未知の遺伝子の機能及び遺伝子間の相互関係を推定することができる．

遺伝子機能解析のために得られた，膨大なデータを処理する方法として，クラスタリング手法が広く利用される．クラスタリングとは統計解析の分野などで使用されているデータ解析手法であり，データの集合を類似度でグループ分けを行う．この操作で得られるグループのことをクラスタと呼ぶ．本節では，代表的なクラスタリング手法の 1 つである K-means 法について説明する．

ここでは非階層型クラスタリングの中で代表的な K-means 法についての説明を行う [3]．非階層型は段階的には動作せずに，一度に全ての点をグループに分けていく．その為，階層型よりもデータ量の増加の際の計算時間があまり増加しないという特徴を持つ．ただし，非階層型の場合処理を行う前に何グループに分けるか指定する必要がある．代表的な非階層型クラスタリングの一つである K-means 法のアルゴリズムを以下に示す (図 1)．

K-means 法のアルゴリズム

Step1: 分類するグループの数を k とする．データの中からランダムに k 個の点を選択する．

Step2: 各データを，k 個の点のうち最も近い点の要素としてグループ化する．

Step3: 前回のグループと今回得られたグループが同一だった場合クラスタリングを終了する．同一ではない場合各グループの中心を k 個求め，再度 Step2 を行う．

K-means 法は初期点に左右されるといった欠点もあるが，計算速度をあまり必要とせずに精度の高いクラスタリングを行える優れた手法である．そのため，商用のソフトウェア等に広く使用されている．

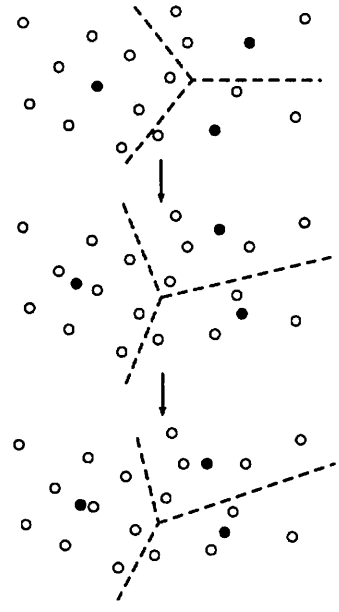

図 1 K-means 法の処理の様子

2.3 確率的進化手法

確率的進化手法（StocE：Stochastic Evolution）は，組み合わせ最適化問題を解くための新しい効率的な方法として開発された手法である [4]．この手法はヒューリスティックアルゴリズム（最適化アルゴリズム）の多くにおいて採用されているランダムウォーク手法に同意せず，完全にランダムウォークを排除したことを特徴として主張している．StocE 手法は多くの問題に対し，速くかつ有効性の高い解を求めることができる優秀なアルゴリズムである．

確率的進化手法の基本的なアルゴリズムを以下に示す．

確率的進化手法のアルゴリズム

Step1: 評価値改善の期待値 R，改善の許容値 P を設定し，初期解を求める．

Step2: 前回の解の評価値を $Cost(S)$ とし，近傍から得られた解の評価値を $Cost(S')$ とする．

Step3: $Cost(S) - Cost(S')$ が $[-P, 0]$ の範囲よりランダムに生成された整数よりも大ならば遷移を採択し，現在の解とする．

Step4: 現在の解と前回の解の評価値が同じ場合，局所的最小解に陥った可能性があるので P の値を増加させ，局所解からの脱出をはかる．そうでなければ P を初期値へ戻す．

Step5: 現在までの最良解と現在の解を比較し，現在の解の方が優れていた場合，最良解の値を更新し，現在までのループ回数を 0 とし再び R 回のループを行う．現在の解が悪かった場合，次の遷移へと移る．ループ回数が R 回に到達していたならば終了する．

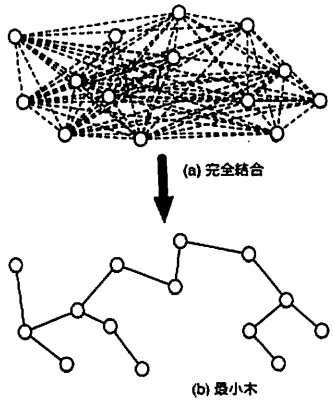

図 2　最小木の構成

P の値は問題によって異なるが，多くの場合，期待値 R は 10 から 20 の時に良い結果をもたらすことが知られている．これよりも大きい場合，多くの問題において探索の後半において多大な時間を浪費する可能性が高い．

StocE は通常は良い解を選択し，局所解に陥った時にのみ悪い解を選択するため，速く，質の良い解を求めることができる．しかし，その性質上常に選択されることのない解が多く存在し，大域的最良解を求めることができない可能性をもつという欠点が指摘されている．ただ，この問題は理論的な考えによるもので，実際の問題において StocE は他のアルゴリズムより少ない処理時間で，優れた解を生成することが示されている．

3. 複数の最小木を考慮した確率的進化計算によるクラスタリング

提案手法は，以下に述べる最小木に基づくクラスタリングをベースとしている．

3.1 最小木に基づくクラスタリング

クラスタリングにおいて，遺伝子発現データ間のユークリッド距離を辺の重みとした完全結合グラフを入力データとして考える．この時，非階層型クラスタリングは，完全結合グラフから辺を取り除くことによって k 個の部分グラフに分割する問題ととらえることができる．その際の目的としては，各部分グラフ内の辺の重みから計算される評価関数が最小となるものが望まれる．しかし，このアプローチでは解空間が莫大になるために何らかの工夫が必要となる．その点に対し，最小木に基づくクラスタリングでは全てのデータ間の関係を最小木で表現する (図 2)．最小木を構成した後に $k-1$ 本の辺を取り除くことで，k 個のクラスタに分割する (図 3)．最小木から取り除く辺は最適化アルゴリズムを用いて設定し，最適な分割パターンを探索する．この手法では最小木を用いて全てのデータを表現することで，複雑な関係の表現が容易になるという利点を持つ．また，最小木を構成し，それを用いて解を探索するため，解空間が縮小され，計算時間も短縮される．

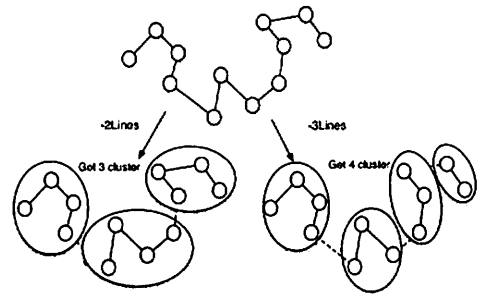

図 3　クラスタの分割例

最小木に基づくクラスタリングでは，クラスタリングは一つの最小木に基づいて動作する．しかし，1 つの最小木では真の解が含まれる空間まで削られる可能性があるという欠点がある。複数の最小木を考慮したクラスタリングでは，一つの最小木ではなく複数の準最小木を含めた最小木に対してクラスタリング分割パターンを求めることで，上記の欠点の解消を目指す．

3.2 提案手法の概要

複数の最小木を考慮したクラスタリングでは，複数の（準）最小木を生成し，それぞれの最小木に対して分割パターンを求め，クラスタリングを行う．複数の最小木の生成には様々な手法が考えられるが，本研究では一定の許容範囲を定め，許容範囲内の差は無視をして辺の長さの比較を行い，複数の準最小木を生成する．なお，本研究では以後，最小木という表現は準最小木も含むものとする．複数の最小木を用いてクラスタリングを行うことで，効果的な分割パターンを得られる確率を増加させている．

3.3 最小木の生成

本研究では，最小木の生成に Prim のアルゴリズムを使用している [5]．Prim のアルゴリズムを用いて辺の長さを比較する際に，許容範囲内の差を無視し，複数の最小木を生成する．最小木を生成するアルゴリズムを以下に示す．

最小木の生成アルゴリズム

Step1: 最もデータ間の距離が短い辺の長さを E_{best}，許容値を P とする．全ての辺 E のなかで ($|E| \leq |E_{best}| \times (1+P)$) の条件を満たす辺をランダムに一つ選択し，その 2 点と辺を最小木の 1 辺とする．

Step2: 最小木の中の点とまだ最小木に含まれていない点を結ぶ辺に対し，Step1 の操作を行う．

Step3: 全ての点が最小木に含まれるまで Step2 を繰り返す．

Step4: 最小木が生成される．

3.4 分割計算

求めた最小木の全てに対し，最適な分割パターンを探索する．提案手法では探索アルゴリズムとして確率的進化手法を用い，探索を行う．

3.4.1 解表現

n 本の辺で構成される最小木から $R-1$ 本の辺を取り除けば，k 個の部分グラフに分割可能である．従って，解表現は，n

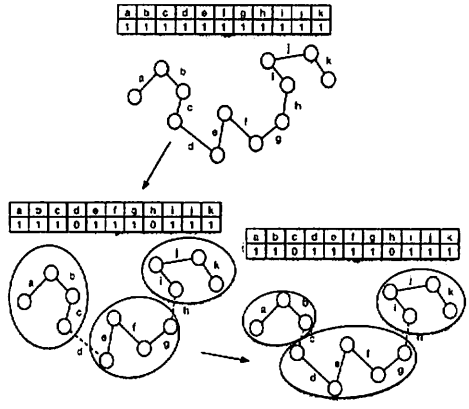

図 4 解の表現

ビットの2進数で $k-1$ 個のゼロを含むものとして定義できる．解の近傍は，ゼロの場所を入れ換えてできる n ビット2進ベクトルの集合となる (図 4)．

3.4.2 評　　価

評価関数は，各クラスタに含まれる点同士のユークリッド距離を求め，その総和の平均を用いる．この関数を最小化することが，分割計算の目的となる．分割計算の評価は次の式を用いる．

$$\sum_{l=1}^{k}\Big(\sum_{x_m,x_n\subseteq C_l,m\neq n}\frac{d(x_m,x_n)}{S}\Big) \qquad (1)$$

$d()$:x_m,x_n のユークリッド距離 C_l:クラスタ l S:辺の総数

3.4.3 分割計算のアルゴリズム

分割計算に用いる提案手法のアルゴリズムを以下に示す (図 5)．以下の例では k 個のクラスタに分割するものとする．

分割計算のアルゴリズム

Step1: 期待値 R，改善許容値 P を設定する．R 回のループを開始する．

Step2: $k-1$ 個の辺をランダムに取り除き，得られた解の評価値を求める．

Step3: ランダムに近傍解を求め，その評価値を得る．

Step4: (前回の評価値-近傍の評価値) を求め，$[-P,0]$ の範囲よりランダムに得た整数よりも大きい場合は近傍解を採択し，現在の解とする．そうでない場合，次の近傍の探索に移る．

Step5: 前回の解と現在の解の値が同じ場合，局所解に陥った可能性があるので改善許容値 P を増加させる．同じでなければ P の値を初期に戻す．

Step6: 現在までの最良解と採択された解を比較し，採択された解が良い場合は最良解を更新しループ回数を 0 にして再び R 回のループを行う．悪い場合にはループ回数を増やし，Step3 へ戻る．

Step7: ループ回数が R に達した場合，分割計算を終了する．

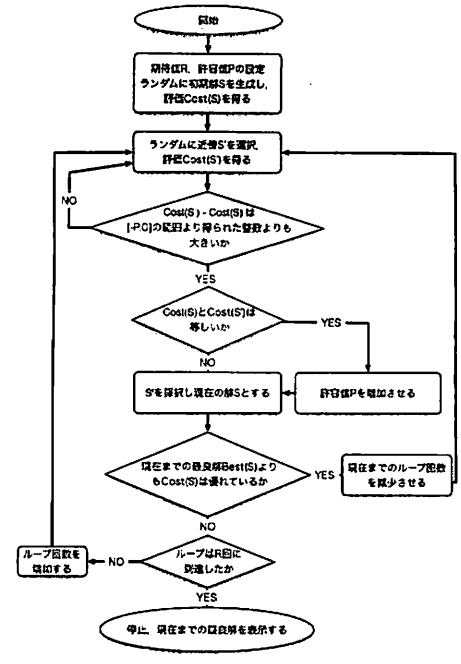

図 5 分割計算の流れ図

4. 実装と検証

4.1 実験環境

開発環境・実験環境ともに，同一のマシンを使用した．

```
OS : Mac OS X 10.3.9
Kernel    : Darwin 7.9.0
CPU model name : PowerPC G4
CPU GHz : 1.00
Memory : 768 MB
gcc : 3.3 20030304 (Apple Computer, Inc. build 1495)
```

C 言語を用いてプログラムの実装を行った．データファイルより時系列データを読み込み，提案手法を用いてクラスタリングを行うプログラムを作成した．また，同様の手順で K-means 法を用いたクラスタリングプログラムも作成し，提案手法との比較を行った．

4.2 実　　験

実験で用いられるデータとして各時間における大腸菌の呼吸量の変動を観察したものを使用する．106 個の異なる遺伝子を取り除いた大腸菌の反応を観察，分類することで取り除かれた遺伝子がどのような働きをしていたのかを推測する．

実験では，K-means 手法と提案手法の性能比較を行った．実験を行った際の提案手法のパラメータは以下の通りである．

- 期待値 $R = 20$
- 許容値 $P = 500$

K-means 手法を用いたものと提案手法の各誤差率での性能

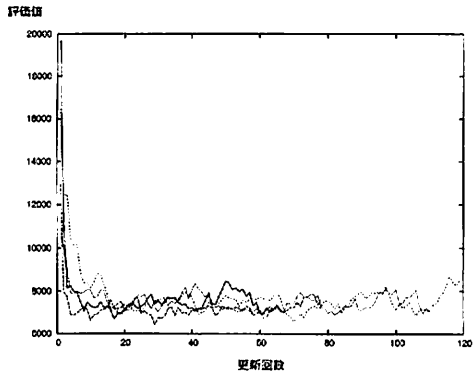

図 6 解の改善

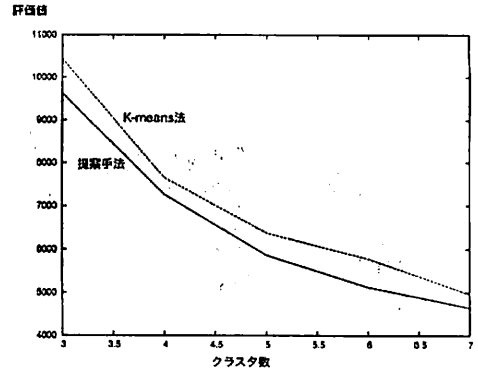

図 8 K-means 法と提案手法の平均比較

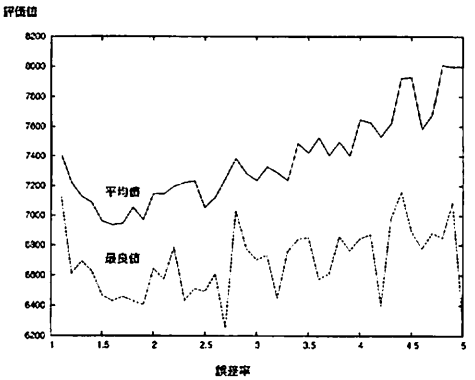

図 7 各誤差率における評価値の平均と最良値

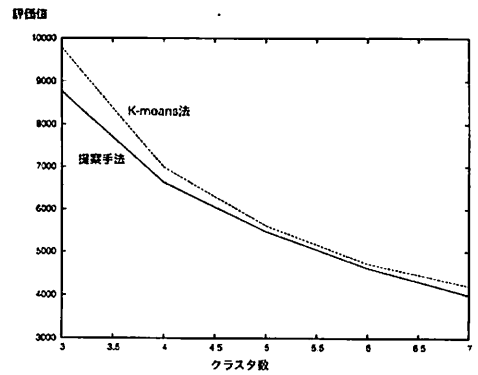

図 9 K-means 法と提案手法の最良解比較

比較をグラフで示す．図 6 は，各誤差率での解の改善の様子を示し，図 7 は，各誤差率での提案手法の解の平均，最良解を示す．図 8，図 9 は K-means 法と提案手法の解の平均，最良解の比較の結果を示している．図 6 の横軸は解の世代，縦軸が評価値となっている．図 7 では横軸が誤差率，縦軸が評価値となる．図 8，図 9 は横軸にクラスタ数，縦軸が評価値を示す．それぞれのデータは４００回ずつ実行した結果の平均である．

各アルゴリズムの平均動作時間は以下の通りである。

K-means 7.41 sec

提案手法 53.21 sec

4.3 考　　察

提案手法において設定する必要のある誤差率は，ある程度高くなければ良い性能を示さないが，高すぎても悪いという結果となった．実験で計測した結果では誤差率は 1.5〜2.7 の間が良いという結果であった (図 7)．

提案手法はどのクラスタ数においても K-means 手法より良い性能を示した (図 8，図 9)．しかしながら，提案手法は K-means に対し 7 倍ほどの時間が必要となっている．これは最小木の構成に時間がかかる為と考えられる．その為，最小木の構成アルゴリズムを変更することで，計算時間を短縮できる可能性がある．提案手法でのクラスタリング結果を示す (図 10)．

5. ま と め

本稿では，実際に遺伝子機能解析の手段として得られたデータに対し，複数の最小木を考慮した確率的進化計算を用いたクラスタリング手法を提案，実装，検証を行うことで，機能解析の現場において提案手法の有効性を確認した．

今後の課題として，

- 他の遺伝子に対しての有効性の確認
- 並列化による動作の効率化

が挙げられる．

文　　献

[1] 森　浩禎，磯野　克巳，「大腸菌におけるゲノム機能の体系的解析」，戦略的基礎研究推進事業「ゲノム構造と機能」公開シンポジウム,2001.

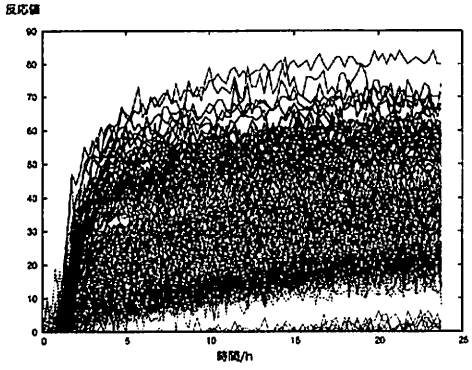

図 10　クラスタリング結果：クラスタ数 4

[2] 鹿川 大輔, 「複数の最小木を考慮したインタラクティブ計算による遺伝子発現データ・クラスタリング」,Proceedings of the 2004 IEICE Society Conference, 2004.

[3] P.S.Bradley, U.M.Fayyad, Refining Initial Points for K-Means Clustering, in Proceedings of the 15th International Conference on Machine Learning, pp.91-99 (1998)

[4] Sadiq M.Sait, Habib Youssef,「組合せ最適化アルゴリズムの最新手法-基礎から工学応用まで-」, 丸善, 2002.

[5] R.C.Prim, Shortest connection networks and some generalizations, Bell Sys. Tech.journal, pages 1389-1401, 1957.