5E−7

自己相似型ネットワーク計算機FINを用いた遺伝的アルゴリズムによる概念クラスタリングについて

高岡 健　辰巳 昭治　韓 明黙　北村 泰彦　奥本 隆昭
大阪市立大学

1　はじめに

観測によって得られた数多くの対象からなる集合を、意味のある集団に分類する手法が数多く提案されている。非数値データをも含めた集合を対象とした分類方法として、CLUSTER/2[4]が提案されている。本稿ではCLUSTER/2において多くの計算量が要求され評価基準を満たす素カバーを階層的に構成する手続きの高速化処理について考える。

基本的には、事象集合の評価基準値を漸次変更して高い評価基準値を持つ集合へ移行することによって階層的構造を持つ分類木が構成されることになる。この手続きは、組合わせ最適化問題を考えることとなる。

本稿では、我々が提案している自己相似型ネットワークをもつマルチプロセッサシステム[2]を用い、最適組合わせを解く手法として有効視されている遺伝的アルゴリズム(GA)[1][3]によるクラスタリングについて考察する。

2　FINによるGA

FINは複数の処理要素(PE)部分及びPEとPEとを連結する通信リングとで構成される。PEは、任意の数のレジスタを持ち算術論理演算, 比較, 条件分岐, 送受信操作の基本命令が実行できる。各レジスタは十分なbit長を有し、また各PEは基準クロックに同調して動作する。またPEは、十分な局所メモリーを持つ。通信リンクはPEごとの入出力ポートと接続されており、各PEはこのポートに対して入出力命令を送出することにより隣接するPEとデータのやり取りを行なうことができる。また各PEは接続されているすべての通信リンクに対して、データの授受を並行して行うことができる。また、FINは各レベル毎に共有メモリを持つ。共有メモリには、そのメモリを共有するPEによりアクセスすることができ、また1つ上、1つ下の階層の共有メモリと内容を交換することができる。FINの形態を図1に示す。

以下に、FINのネットワーク形態に関する用語で以下の説明に必要なものを定義しておく。

レベルlのブロック　レベル0のブロックは1つのPEからなる。レベル$l(l>0)$のブロックは、3個のレベル$l-1$ブロックの完全結合で構成される。ここで云うブロックの結合とは、あるブロックの中の1つのPEと他のブロックの中の1つのPEとの結合を意味する。レベルlのブロック内の3個のレベル$l-1$ブロックの内、他のレベルlブロックと結合するものをレベルlのコーナブロックという。

ブロックリーダ　レベルlブロックのうち1つのPEをそのブロックのブロックリーダと呼ぶ。レベルlのブロックリーダは、他のレベルlブロックと結合していないコーナPEである。

Conceptual Clustering using Genetic Algorithm on Fractal Geometory-based Interconnection Network FIN
TAKAOKA,TATSUMI,HAN,KITAMURA,OKUMOTO
Osaka City University

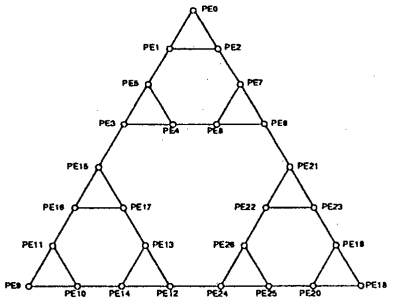

図1　レベル3のFIN

並列GAを行う際の計算機モデルにFINを用いることは、FINの持つ拡張性とフラクタル形態を利用することを可能とする。特に、いくつかの部分問題に分割するような問題では、それにあわせて解くような使い方ができるという利点がある。

次に、FINによる遺伝的アルゴリズムの実行方法について述べる。FIN上でGAを行うにあたって、次の点に留意する。

1. 問題特有の処理に対する自由度を大きく採るようにする。
2. 並列化による処理の高速化をめざす。
3. FINのフラクタル形態を利用した制御を行う。

FINによるGAでは、集団はFINの結合形態により階層的に自己相似型の副ブロックに分かれている。各個体はFINのレベル0ブロック(1つのPE)に相当し、各PEのメモリに自分自身の遺伝子を持つ。つまり、FINのレベルをlとすると、集団の個体数は3^lとなる。

GAにおける最も重要な遺伝操作である交叉は、通信リンクを通してPE間で行う。また、複数の交叉を同時に行う制御として、FINのフラクタル形態を利用した階層的交叉を行う。

レベルlの交叉　各レベル$l(l\geq 1)$のブロック内のレベル$l-1$の副ブロックのコーナPEを親とする交叉を行い、新しくできた子孫と置き換える。

この方法によると、大きさレベルLのFINでは、レベルlの交叉を3^{L+1-l}個同時に行うことができる。また、初期解の生成や評価値の計算も$\frac{1}{3^L}$になる。特に評価値の計算と交叉は毎世代行わなければならない処理であり、並列に行われることにより、シーケンシャルなシステムに比べ全体の探索時間を大幅に短縮することができる。

3　FINによる概念クラスタリング

概念クラスタリングとは、与えられた対象や事象の集合から、人間が行うような意味的解釈を持つグループに分類する問題である。こ

のことにより分類された集団が持つ意味がその集団の持つ概念を記述する。

分類の対象は、複数の概念属性値を持つ事象であり、概念属性は名称変数、線形変数、構造変数で表現される。

このような概念属性値を持つもののクラス分類は、組み合わせ的側面を持ち、また分類する組み合わせ(属性数)が大きいような対象の分類は非常に困難な問題となる。よって、並列処理向きの近似解法であるGAを使って解くことが考えられる。

FINによる概念クラスタリング

FINによるクラスタリングでは、クラスタリングは階層的にトップダウンに行われる。各階層では、あらかじめ決められた評価基準を達成するまでGAが繰り返される。また、ある階層で決められた世代以内に目標とする評価を達成できなかったときには、1つ上の階層に戻り再度クラスタリングをやり直すことにする。各階層でのクラスタリングの評価は、高い階層ではクラスタの記述が実際のデータとよく一致することよりもクラスタ記述が単純であることの方が重要であり、階層が下がると共にクラスタ記述と実際のデータとの一致度が重要となる。FINを用いた概念クラスタリングでは、FINの結合形態に沿って分割数を3とすることにする。

クラスタリングの質の評価

ここでは、クラスタリングの質の評価に次の評価基準を導入した。(A) クラスタの記述が実際のデータとよく一致すること。(B) クラスタ記述が単純であること。(C) 分割されたクラスタ内の対象の数が、大きく異ならないこと。

4 クラスタ分割向き遺伝操作

コーディング

1つの遺伝子は、1個あるいは複数個の遺伝因子から成るものとする。ひとつの遺伝因子は、すべての事象を1つあるいはそれ以上の属性値を用いた、2つあるいは3つのクラスタへの完全な分割を記述する。二つ以上の遺伝因子からなる遺伝子の場合には、各遺伝因子の示すクラスタ分割の組み合わせの中から、事象の数とクラスタ記述との一致度を基にした評価基準により表現型が現れるものとする。1つの遺伝因子は一定長からなり、遺伝子の長さは遺伝子に含まれる遺伝因子の数によって変わる。遺伝子の1つめの要素で遺伝因子の数を表し、遺伝子に含まれる遺伝因子数をk個、1つの遺伝因子の長さをlとすると、遺伝子の長さは$k \times l + 1$となる。以下に、属性の種類によるコーデイング方法を述べる。

名称変数 各名称に対して、その名称の所属するクラス番号の列で表現する。

線形変数 各クラスの値の範囲とそれを分割するしきい値で表記する。この分割による1つめのクラスは、値域の始めから1番目のしきい値までに含まれる事象の集合となる。以下同様にn番目のクラスは、$(n-1)$番目のしきい値からn番目のしきい値までとする。

構造変数 ここでは構造変数のための構造を、木構造に限定する。各ノードから出る枝に対して1から順番に数を割り当てる。コード化は、根からたどる枝の番号列で表現する。

遺伝操作

遺伝操作は、交叉, 突然変異, 転置を用いる。交叉は、二つの親となる遺伝子の持つ遺伝因子を組み合わせる。

突然変異は、遺伝因子に記述されているクラスの割当てを変える。名称変数に対する突然変異は、各名称の要素にたいして表現されているクラス番号をランダムに他のクラス番号に変更する。線形変数に対しては、クラスタ割当てを行うしきい値をランダムに変更する。構造変数に対しては、各クラスに分割される木のブロック番号を変更する。

転置は、遺伝因子に記述されているクラスの順番を変える。線形変数に対する転値は、値域の始めの部分から順番にクラス番号が$1, 2, 3, \cdots$であるものを、$\cdots, 3, 2, 1$という順番に変更する。構造変数に対する転置は、分割される木のブロックの各クラスへの割当てをランダムに変更する。

評価

クラスタの記述が実際のデータを記述するに適切であるか否かを判断する尺度として、クラスタの希薄度を考える。[4] クラスタを形成している複合体をαとしたとき、事象全体の数を$t(\alpha)$、未観測事象の数を$s(\alpha)$とすると、希薄度は$r(\alpha) = \frac{s(\alpha)}{t(\alpha)}$で表される。希薄度が大きいということは、すなわちクラスタ内に未観測事象が多いということを意味しており、クラスタの記述と実際のデータがあまりよく一致していないことを示す。あるクラスタ分割に対する希薄度は属性数をnとするとき、

$$\begin{aligned}
\text{希薄度} &= (\sum^{n}[\text{分割に対する各属性の希薄度}])/n \\
\text{名称変数の場合} &= \frac{[\text{分割した後取り扱う名称の数}]}{[\text{全名称数}]} \times 100 \\
\text{線形変数の場合} &= \frac{[\text{分割した後取り扱う値域}]}{[\text{全値域}]} \times 100 \\
\text{構造変数の場合} &= (1 - \frac{[\text{構造化で統合される事象数}]-1}{[\text{全構造化事象数}]}) \times 100
\end{aligned}$$

とする。クラスタ記述の単純さをクラスタ分割を行うのに必要な属性値の数であらわし、

$$\text{クラスタの単純さ} = (1 - \frac{[\text{クラスタ分割に必要な属性数}]}{[\text{全属性数}]}) \times 100$$

とする。また、分割されるクラスタに属する事象の数が各クラスタで同じ数になるほうが良いと考え、

$$\text{事象散布度} = (\frac{\sum^{3}_{i=1}|[\text{第}\ i\ \text{クラスタの事象数}] - ([\text{全事象数}]/3)|}{(4 \times [\text{全事象数}])/3}) \times 100$$

とする。よって

$$\text{評価値} = \alpha_l(100 - [\text{希薄度}]) + \beta_l[\text{クラスタの単純さ}] + \gamma_l[\text{事象散布度}]$$

とする。ここで、$\alpha_l, \beta_l, \gamma_l$は、$l$階層における各評価基準間のトレードオフパラメータであり、階層的に$\alpha_i < \alpha_{i-1}, \beta_j > \beta_{j-1}, \gamma_k > \gamma_{k-1}$となるような経験的な値である。

参考文献

[1] Lawrence Davis Ed., : " Genetic Algolithms and Simulated Annealing", Morgan Kaufmann Publishers,1987.

[2] 菅谷 光啓 前場 隆史 辰巳 昭治 阿部 健一, : " 自己相似型ネットワークを有するマルチプロセッサシステム上での並列アルゴリズム", 電子情報通信学会論文誌, D-1 Vol.j73-D-1 No.11 847-855 Nob 1990.

[3] 高岡 健, 辰巳 昭治, 奥本 隆明, 北村 泰彦, : " 自己相似型ネットワーク計算機FIN-1による遺伝的アルゴリズムについて", 情報処理学会第43回全国大会 7B-6,Oct 1991.

[4] R.S.Michalsky Ed., 電総研 AI研究 Group Trans., : "教示学習と知的 CAI" , 知識獲得と学習シリーズ3 , 共立出版 , 1987.