遺伝的アルゴリズムによる記号列知識の獲得

1 U − 5

久保長徳 西野順二 小高知宏 小倉久和

(福井大学工学部)

1 はじめに

DNA 配列中のスプライス部位のような記号列パターンの学習においては,学習結果として対象部位の特徴を表現した正規表現や文法等のような知識が獲得できることが望ましい. しかし,例えばニューラルネットワークのような方法を用いた場合,学習結果はシナプス結合の重みやニューロン細胞の閾値といったパラメータの組合せとして蓄えられ,そこから対象に関する知識を抽出することは困難である.

ニューラルネットワークに対して遺伝的アルゴリズム (以下 GA) は解を表現する記号列である遺伝子を,選択,淘汰,交叉,突然変異等の遺伝的操作を繰り返すことで獲得する方法であり,それらの表現を実験者が決定できるため GA では,獲得された遺伝子から解に関する知識を理解することは容易であると思われる.

そこで本研究ではあらかじめ用意した学習セットに対しての記号知識の獲得を簡単な知識 (遺伝子) 表現を用いて試みた. その結果,単純な知識表現では獲得された知識がそれ以上の構造をもつことはなく,有用な知識になりえないという結論を得た. そこで今回はある程度の構造を表現できるような知識表現を用いての実験を試みている. 本稿ではその知識表現と獲得方法の概要について報告する.

```
01010##0#          010100000
                   010100001
                   010100100
don't care char    010100101
# = 0 or 1        010101000
                   010101001
                   010101100
                   010101101
```

図 1: ドントケア表現

Acquiring Symbol Sequence Knowledge Using Genetic Algorithm,
Takenori Kubo, Junji Nishino, Tomohiro Odaka, Hisakazu Ogura
Faculty of Engineering, Fukui University
3-9-1 Bunkyou, Fukui 910, Japan

2 知識表現の改善

本研究ではこれまで知識表現として「ドントケア表現」と名付けた図 1 のような表現方法を用いてきた. 図の左上の文字列がある記号列の特徴を表す知識の記述である. そして文字列中の'#' はドントケア文字とよばれ,この位置にはどんな記号も許されることを示している. この図では記号として 0,1 をもちいているので,ドントケア文字はそのどちらも受け付ける. その結果,この知識が表す記号列を図の左に記している.

この方法によってある程度の判別は可能である. しかしこの方法では知識として獲得できるものは記号列中の特定の位置の選択可能な記号の集合の列,いわゆるマスクパターンのようなもので,記号列中の前後関係などのより複雑な構造の表現は不可能である.

そこでより複雑な構造の表現を可能とするために,知識表現として文法を用いる. 文法といっても GA で獲得するために非常に簡略化したものであり,以下のような書換ルールの集合としている. 書換ルールは書換前後の 2 つの記号列 a,b のペアであり,それは a を b に書換えることを意味する. 例えば,A → xA という書換ルールを記号列 A に適用すると xA に書換えられる. この記号列 xA の A にさらにルールを適用することでさらに xxA へと書換えられる. 同様にこのルールを繰返し適用することで,xxxA,xxxxA といった記号列が生成できる.

これらの知識は GA によって生成されるため,文法として完全なものにならない可能性が高いが,完全な文法を生成するような,遺伝的操作は困難かつ手間がかかると思われる. また判別時に利用されることがなくても,遺伝的操作によって優秀な遺伝子を生成するような書換ルールの存在も考えられる. そのためこの表現によろ知識は適当な初期記号から記号列が生成できる書換ルールの組を持っているかを調べることで,記号列の正負を判別する.

3 知識獲得

知識獲得は GA を用いて以下のように行なう.

1. 初期遺伝子を無作為に生成する.

2. 遺伝子の評価を行なう.

3. 交叉, 突然変異を行なう.

4. 選択を行なう

5. 2～4 を繰返す

また知識表現に関わる評価, 交叉, 突然変異は以下のようにして行なう.

3.1 評価

知識の評価は学習セットの判別結果を用いる. 正例負例からなる学習セットに含まれる記号列の判別を行ない. その結果, 学習セット中の正例負例の評価として用いる.

また遺伝子がルールの集合ということで, 遺伝子の評価自体を小さなGA と考えることができる. そこで判別する際にルールの得点をつけて突然変異におけるルールの削除などで利用頻度の低いものほど削除されやすくするといったこも考えられる.

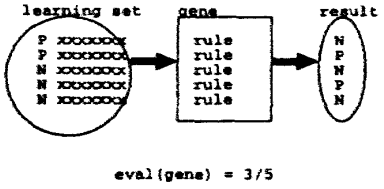

図 2: 遺伝子の評価

3.2 交叉

交叉では書換ルールの入替, 書換ルール同士の交叉を行なう. 書換ルールの入替は交叉を行う遺伝子が持つ同数のルールを交換する.

またルール間の交叉は, 交叉を行なう遺伝子間でいくつかのルールを選び, そのルール間で記号列 a,b を交換する.

3.3 突然変異

突然変異ではルールの追加とルールの削除を行なう. ルールの追加では無作為に生成したルールを遺伝子に追加する.

また削除では遺伝子の持つルールからのいくつかを選択しそれらを削除を行なう. 削除がうまく働くことで遺伝子の肥大化の防止が期待できる.

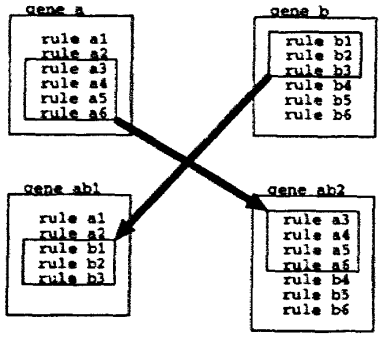

図 3: 交叉

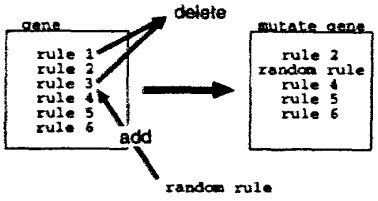

図 4: 突然変異

4 結び

GA による記号知識の獲得に関して, ある程度複雑な構造を表現できるような知識表現とそのGA における操作について報告した.

今後は実際に実験を行ない, その結果について考察を行なってゆく.

参考文献

[1] 北野 宏明, 遺伝的アルゴリズム 1,2, 産業図書株式会社,1993,1995.

[2] 嵩忠雄, 都倉信樹, 谷口健一, 形式言語理論, 電子情報通信学会,1988.