論 文

Technical Papers

共生進化に基づく簡素な決定木の生成

Generating the Simple Decision Tree with Symbiotic Evolution

大谷 紀子
Noriko Otani
武蔵工業大学 環境情報学部
Faculty of Environmental and Information Studies, Musashi Institute of Technology
otani@yc.musashi-tech.ac.jp, http://www.yc.musashi-tech.ac.jp/~otani/

志村 正道
Masamichi Shimura
（同　上）
shimura@yc.musashi-tech.ac.jp

keywords: decision tree, genetic algorithm, symbiotic evolution

Summary

In representing classification rules by decision trees, simplicity of tree structure is as important as predictive accuracy especially in consideration of the comprehensibility to a human, the memory capacity and the time required to classify. Trees tend to be complex when they get high accuracy. This paper proposes a novel method for generating accurate and simple decision trees based on symbiotic evolution. It is distinctive of symbiotic evolution that two different populations are evolved in parallel through genetic algorithms. In our method one's individuals are partial trees of height 1, and the other's individuals are whole trees represented by the combinations of the former individuals. Generally, overfitting to training examples prevents getting high predictive accuracy. In order to circumvent this difficulty, individuals are evaluated with not only the accuracy in training examples but also the correct answer biased rate indicating the dispersion of the correct answers in the terminal nodes. Based on our method we developed a system called SESAT for generating decision trees. Our experimental results show that SESAT compares favorably with other systems on several datasets in the UCI repository. SESAT has the ability to generate more simple trees than C5.0 without sacrificing predictive accuracy.

1. はじめに

決定木は分類規則の木構造による表現技法である．決定木の非終端ノードには属性の種類，アークには属性値，終端ノードにはクラスが割り当てられており，事例の各属性値に従って根ノードから終端ノードまで決定木を辿ることで，事例の属するクラスが判定される．属するクラスが既知である訓練事例の分類結果を指標として，木の形状および値の割り当てを決定し，未知事例分類のための決定木を生成する．

情報量に基づく様々な決定木生成アルゴリズムが提案されており，ID3[Quinlan 86]，CART[Breiman 84]，C4.5[Quinlan 93] 等が代表的な分類システムとして挙げられる．これらのシステムは，分類誤り率による枝刈りやブースティングアルゴリズム等の利用により，未知事例分類において高い正解率を得ている．しかし，正解率を重視して生成された決定木は，複雑でノード数が多いという傾向がある．記憶容量や分類処理の速度，分類規則の解釈の容易さの点から，簡素さは決定木の評価において重要な要因といえる．高正解率を保持しつつ木を簡素にすることは，決定木生成での一目標となっている．

最適化問題における解の探索手法の１つに遺伝的アルゴリズム (Genetic Algorithm; GA)[Goldberg 89] がある．与えられた問題に対する解候補を個体として表現し，個体の集団において評価，選択，生殖というサイクルを繰り返すことで，解として最適と思われる個体を生成する方法である．GA は，構造が不明で広大な解空間を持つ問題において，大域的探索を効率よく行なうことができるため，多様な分野に応用されている．

Moriarty らは GA の一手法として共生進化 (symbiotic evolution) を提案し，ニューラルネットワークの隠れ層を学習するシステム SANE を構築した [Moriarty 95, Moriarty 96, Moriarty 98]．共生進化の特徴は，部分解と全体解をそれぞれ個体とする２つの集団を並行して進化させる点にある．全体解は部分解の組み合わせにより表現する．SANE では隠れ層のニューロンを部分解，ニューロンの組み合わせであるネットワーク構成子 (network blueprint) を全体解として集団を形成し，両者を並行進化させている．このとき，全体解の評価に基づいて部分解を評価し，その評価値に従って進化した部分解を全体解に反映する．両者を相互に関係付けながら進化させることで集団内の個体の多様性が維持され，局所解への収束を回避した効率的な最適解探索を可能としている．

本論文では，未知事例を正確に分類でき，かつできる

だけ簡素な決定木の生成を目的として，共生進化に基づく新しい決定木生成法を提案する．決定木は高さ 1 の部分木を複数組み合わせたものといえるので，高さ 1 の部分木を部分解，部分木の組み合わせで表現される決定木を全体解として，共生進化により決定木を生成する．進化過程では，部分木の結合関係の保持，決定木のノード数増加の抑制など，対象が決定木であるために必要となる処理を行なう．進化の基準となる決定木の評価は，未知事例の正解率向上を考慮した適応度関数により決定する．提案手法に基づく決定木生成システム SESAT(Symbiotic Evolution for Simple and Accurate Trees) を構築し，ベンチマークデータによる実験結果を示す．

以下，2 章で SESAT における決定木生成手法について説明する．3 章では，提案手法の有効性を検証するために，UCI リポジトリのデータを用いて行なった評価実験の結果を示す．SESAT に加え，属性，属性値，クラスをランダムに割り当てて決定木を生成するシステム，一般的な GA で決定木を個体とする集団のみを進化させるシステム，および C4.5 の後継である C5.0 でも実験を行なう．評価実験の結果を踏まえて 4 章で提案手法に関する考察を行ない，5 章で結論を述べる．

2. 決定木生成における共生進化

ニューラルネットワークでは，入力層からの情報が隠れ層にある複数のニューロンを介して出力層に伝達される．隠れ層の各ニューロンの動作は互いに独立であるため，隠れ層を構成するニューロンが決まるとネットワーク全体の動作は一意に定まる．SANE では全体解を隠れ層のニューロンの組み合わせで表現しているが，その組み合わせ方には制約がないため，遺伝操作により構造上不備のあるネットワークが生成されることはない．

一方，決定木は複数の部分木の組み合わせであるが，決定木に含まれる部分木が指定されても，決定木の構造は一意に定まらない．各部分木の結合関係により決定木の構造は変化する．部分木の数に過不足が生じたり，冗長で無意味な決定木が生成される場合もある．従って，決定木生成に共生進化を適用する際には，部分解の単なる組み合わせで表現された全体解を一般的な遺伝操作で進化させることは望ましくない．部分解の結合関係を保持した形で全体解を表現し，解候補として適切な決定木のみを集団の個体とすることが必要となる．以下，SESAT における決定木生成方法の詳細について説明する．遺伝子表現および集団の構成に関しては，[大谷 04] において検討した結果を反映している．

2·1 決定木の制約

GA により木構造の解を探索するという点で，本手法は遺伝的プログラミング (Genetic Programming; GP) に類する．GP では学習過程で木のノード数が急速に増大するブロート現象が発生しやすい [Angeline 98]．ブロート現象の原因は，実行されても正解率に関与しない冗長な部分木 (意味論的イントロン)，および実行されない無意味な部分木 (構文的イントロン) の存在にある．イントロン (intron) とは DNA 中で遺伝情報を持たない部分を指す．SESAT においては，隣り合う同一の部分木[*1]が前者，どの事例も到達しない部分木が後者にあたる．イントロンは最適解探索に有益な場合もあるが，実行時間，解の複雑さ，過学習の点を踏まえると，木の複雑化は可能な限り回避すべきと思われる．SESAT では，複雑な木の生成を避けるため，解候補として生成される木に対して次の 4 つの制約を課す．

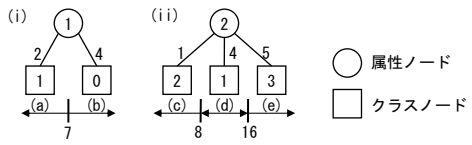


図 1 sprig の例

制約 1： 木の高さは H 以下である．

制約 2： 非終端ノードは 2〜M 個の子ノードを持つ．

制約 3： 隣り合う部分木は必ず異なる．

制約 4： 全ノードにいずれかの訓練事例が到達する．

制約 3 が意味論的イントロン，制約 4 が構文的イントロンへの対処である．H および M の値，さらに後出のパラメータ p_m，N_g，N_{sp}，N_{tp} の値は，データや目的に応じてあらかじめユーザが指定する．SESAT の学習過程で保持する決定木は，すべて上記 4 制約を満たす．

2·2 sprig

SESAT の部分解 sprig を図 1 に示すような高さ 1 の部分木で表し，根ノードを属性ノード，葉ノードをクラスノードと呼ぶ．事例に出現する属性数が A，クラス数が C であるとき，属性ノードには 1〜A の属性番号，クラスノードには 1〜C のクラス番号もしくは 0 の値が入る．sprig が決定木に組み込まれる際，クラス番号が 0 のクラスノードには別の sprig の属性ノードが接続され，そのノードは非終端ノードとなる．

各アークは 1〜M の属性値番号によりラベル付けされている．事例の属性値に従って属性ノードからクラスノードへと走査するときは，アーク間に設定された閾値と属性値の大小関係に従って辿るアークを選択する．閾値としては，訓練事例の属性値の範囲を 1〜M に対応付け，ラベル平均に相当する値を設定する．例えば，子ノード数の最大値 M が 5，訓練事例における属性値の範囲が属性 1 で 5〜9，属性 2 で 2〜18 であるとき，図 1(i) の閾値はラベル平均 3 に相当する 7，(ii) の閾値はラベル平

*1 隣り合う部分木とは，共に同一の親を持ち，かつ隣り合っているノードを根ノードとする部分木を指す．例えば，図 1(i) の部分木が 2 つ並んでおり，同じノードを親としている場合，この 2 つの部分木は隣り合う同一の部分木という．

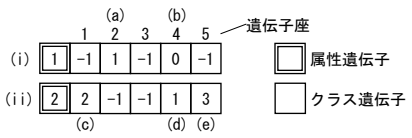


図 2 sprig の染色体の例 ($M = 5$ の場合)

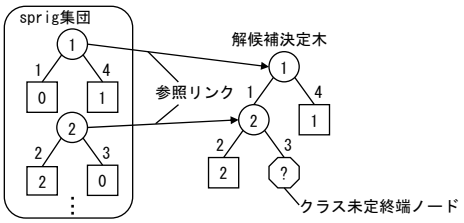


図 3 決定木構成子

均 2.5，4.5 に相当する 8，16 となる．

sprig を表す染色体は，1 個の属性遺伝子と M 個のクラス遺伝子からなる．図 1 の sprig を表す染色体の例を図 2 に示す．属性遺伝子が属性ノードに対応し，クラス遺伝子がクラスノードに対応する．クラス遺伝子が -1 のとき，sprig は対応するノードを持たない．図 2 のクラス遺伝子 (a)〜(e) が，それぞれ図 1 のクラスノード (a)〜(e) に対応する．-1 以外のクラス遺伝子の位置を表す遺伝子座 1〜M がアークのラベルとなる．

sprig の染色体を新たに生成する場合は，制約 2 および制約 3 を満たす範囲で各遺伝子をランダムに設定する．N_{sp} 個の sprig を生成し，初期集団とする．

2·3 決定木構成子

SESAT における全体解を決定木構成子と呼ぶ．決定木構成子は，図 3 に示すような解候補決定木と参照リンクから構成されている．解候補決定木は，sprig 集団から選択した sprig のノードの値およびアークのラベルを参照して形成される．参照リンクは解候補決定木の生成時に参照した sprig からのリンクであり，解候補決定木の非終端ノードを終点とする．決定木構成子の生成手順を図 4 に示す．

クラス未定終端ノードとは，制約 1 を満たすために暫定的に作られた終端ノードであり，Step.4 において訓練正解率が最も高くなるようにクラスが割り当てられる．これにより，同一の sprig を参照する非終端ノードでも子ノードの種類が異なる可能性が生じる．Step.5 と Step.6 では，それぞれ制約 4 および制約 3 が満たされるように決定木構成子を変形する．子ノードが 1 つとなって制約 2 が満たされなくなった場合は Step.7 で修正する．

決定木構成子集団の個体数として指定された N_{tp} 個の決定木構成子を生成し，初期集団とする．複数の決定木構成子から参照される sprig や，どの決定木構成子からも参照されない sprig も存在し得る．

Step.1	sprig 集団から無作為に選択した sprig の値に基づいて部分木を生成し，解候補決定木とする．
Step.2	下位の部分木が不定の非終端ノードに対して，無作為に選んだ sprig による部分木を付加する．レベルが H の非終端ノードはクラス未定終端ノードとする．
Step.3	すべての非終端ノードについて下位の部分木が定まるまで，Step.2 を繰り返す．
Step.4	クラス未定終端ノードのクラスを決定する．
Step.5	訓練事例の到達しない部分木を削除する．
Step.6	隣り合う同一部分木を統合する．
Step.7	子ノード数が 1 つの非終端ノードを削除し，当該ノード以下の部分木を接続する．

図 4 決定木構成子生成手順

2·4 適　応　度

C4.5 では，正解率を高める指標として情報利得比を，過学習を回避するための指標として分類誤り率を採用している．情報利得比に基づいて決定木を生成した後，各ノードにおける分類誤り率により枝刈りを行なう．一方，決定木構成子は解候補決定木の生成後に評価するため，決定木構成子の適応度では，生成後の決定木の正解率を計ると同時に，木全体としてどの程度過学習が起こっているかを見極める必要がある．また，子個体が生成されるたびに適応度を算出するため，処理時間の面から各データでの 1 回の走査結果だけで算出できる適応度が望ましい．

より多くの訓練事例を正しく分類するよう，訓練事例に特化した決定木を生成した場合，各訓練事例が互いに異なる終端ノードで正解と判断されるような決定木になりがちである．このように過学習が起こると，終端ノードごとの正解事例数の散らばりが大きくなり，訓練事例における正解率は高くなるが，高い予測正解率は望めない．終端ノードごとの正解数の散らばりが小さいほど過学習の可能性が低いと考えられる．以上の考察より，正解数の散らばりを表す正解局在率を定義し，正解率が高く正解局在率が低いときに高い値を示す適応度によって，決定木構成子を評価する．

決定木構成子 T の正解局在率 $bias(T)$ は次式に従って算出する．ここで，全正解事例数を c，n 個の終端ノードにおける正解事例数をそれぞれ c_1〜c_n とする．

$$bias(T) = \begin{cases} \dfrac{-\displaystyle\sum_{i=1}^{n} \frac{c_i}{c} \log_2 \frac{c_i}{c}}{-\log_2 \dfrac{1}{c}} & (c \neq 0) \\ 1.0 & (c = 0) \end{cases} \quad (1)$$

式 (1) は，正解事例が最も散らばっているときを基準として，T の正解数がどの程度散らばっているかを表している．すべての正解事例が 1 つの終端ノードに到達したときに 0，互いに異なる終端ノードに到達したときに 1 となり，終端ノードごとの正解事例数の散らばりが大きいほど大きな値を取る．

訓練事例の正解率 $acc(T)$ と正解局在率 $bias(T)$ から，決定木構成子の適応度 $tfit(T)$ を次式により算出する．α は正解局在率を考慮する度合を示す定数である．

$$tfit(T) = acc(T) \cdot (1 - \alpha \cdot bias(T)) \cdot 100 \quad (2)$$

sprig は参照リンク先の部分木により評価する．sprig S を参照する部分木のうち，最も適応度の高い決定木構成子に属する部分木を S の最良部分木と呼ぶ．最良部分木の属する決定木構成子の適応度を S の適応度とする．制約を満たすための変形により，sprig が表す木と参照リンク先の部分木が異なる場合があるが，sprig の適応度算出の際に，sprig が最良部分木を表現するよう sprig の遺伝子に変更を加える．これにより，適応度の高い決定木構成子に参照される sprig が高い評価を受け，その性質が集団内に広まる可能性が高くなる．

2·5 世 代 交 代

sprig 集団の世代交代では，[Moriarty 96] と同様にして，N_{tp} 個の個体のうち上位半数をそのまま次世代に残す．下位半数の個体は，上位四半数から選んだ 2 つの個体を親として交叉を行ない，生成された 2 つの子のいずれかと，2 つの親のいずれかで置き換える．交叉により終端ノード数が 2 個未満になった場合は，ランダムに位置番号と遺伝子の値を設定し，制約 2 を満たすようにする．子が制約 3 を満たさない場合は，問題箇所の遺伝子を他の遺伝子に置き換える．すべての個体の遺伝子に対して確率 p_m で突然変異を発生させ，次世代の個体とする．

決定木構成子集団の世代交代モデルとしては，[佐藤 97] で提案されている MGG(Minimal Generation Gap) モデルを採用する．MGG モデルは，局所解収束の回避と進化的停滞の抑制を意図して考案されたモデルである．集団からランダムに非復元抽出された 2 個体を親として子を生成し，親と子の個体のうち，最良個体およびルーレット選択で選ばれた 1 個体の計 2 個体を次世代に残す．

子として生成されるのは，以下の 4 種類の木 C_1〜C_4 を解候補決定木とする決定木構成子である．ここで，親個体の解候補決定木を P_1，P_2 とし，P_1 と P_2 からランダムに選択したノードを n_1，n_2 とする．

C_1： P_1 の n_1 以下の部分木を P_2 の n_2 以下の部分木で置き換えた木

C_2： P_2 の n_2 以下の部分木を P_1 の n_1 以下の部分木で置き換えた木

C_3： P_1 の各ノードに sprig の遺伝子を反映した木

C_4： P_2 の各ノードに sprig の遺伝子を反映した木

C_1，C_2 は交叉により生成された木であり，C_3，C_4 は解候補決定木の各部分木と参照リンク元の sprig の表す木が同じになるよう変更を加えた木である．変更処理は根ノードから順に行なう．子ノードの減少，あるいは非終端ノードから終端ノードへの変更が発生した場合には不要な部分木を削除する．子ノードの増加，あるいは終端ノードから非終端ノードへの変更が発生した場合には新たに部分木を追加する．

Step.1	sprig の進化
Step.2	決定木構成子集団から親を選択
Step.3	子を生成
Step.4	子の解候補決定木を評価
Step.5	sprig の評価
Step.6	次世代に残す個体を選択

図 5 一世代の処理手順

子の生成後，C_1〜C_4 の全非終端ノードに対して確率 p_m で突然変異を発生させる．突然変異では，当該ノードの参照リンク元の sprig を変更し，以下の部分木を作り変える．P_1，P_2，C_1〜C_4 のうち，最良個体およびルーレットで選択された 2 個体を次世代に残す．

一世代の処理の流れを図 5 に示す．初期集団を生成した後，図 5 の処理を N_g 回繰り返し，最良個体の解候補決定木を SESAT の出力解とする．

3. 評 価 実 験

本章では，提案手法の有効性を確認するための評価実験について説明する．共生進化が最適木の探索に有効であることを示すため，SESAT と同様の木を生成できる次の 2 つのシステムを用意した．

ESAT：一般的な GA に基づいて解を探索する決定木生成システム．SESAT から sprig に関する処理を除き，解候補決定木を個体とする集団のみを保有する．C_1 と C_2 の子を生成して進化させる．初期集団の解候補決定木は，ランダムに各値を決定して生成する．

RSAT：ランダムサーチにより解を探索する決定木生成システム．ESAT の初期集団形成時と同様，ランダムに解候補決定木を生成し，評価するルーチンを繰り返す．

SESAT，ESAT，RSAT および C4.5[Quinlan 93] の後継である C5.0 において，UCI 機械学習リポジトリのデータ [UCI] を用いて実験を行なった．C5.0 の各パラメータはデフォルトの設定とした．使用した 12 種類のデータの事例数，属性数，クラス数を表 1 に示す．属性値には実数値，整数値，数値以外の値があるが，数値以外の属性値には 1 から順に整数の番号を割り当て，low, middle, high など順序付けが可能な場合は番号の大小と属性値の順序が一致するようにした．

実験は，表 1 に示した事例数のうち，10 分の 9 を訓練事例，10 分の 1 をテスト事例とする 10-fold クロスバリデーションとし，訓練事例における平均適応度，テスト事例における平均正解率，決定木の平均ノード数を調査する．SESAT，ESAT，RSAT では，各事例による試行を 10 回繰り返した平均を取る．

SESAT と ESAT で設定したパラメータを表 2 に示す．

表 1 実験用データ

データ名	事例数	属性数	クラス数
aust	690	14	2
balance	624	4	3
breast	683	9	2
bupa	345	6	2
glass	214	9	6
heart-c	297	13	2
iris	150	4	3
monks1	124	6	2
monks2	170	6	2
monks3	123	6	2
pima	768	8	2
post	87	8	3

表 2 パラメータ

パラメータ	値
決定木構成子の適応度の係数 α	0.2
突然変異確率 p_m	0.01
sprig 集団の個体数 N_{sp}	400
決定木構成子集団の個体数 N_{tp}	1000
世代交代回数 N_g	50000
木の高さの上限値 H	10
sprig のクラスノード数の上限値 M	5

H と M は C5.0 で生成された決定木[*2] よりもノード数が多い決定木が生成可能となるように設定した．その他のパラメータには，表 1 のデータの一部による予備実験で安定して良い結果が得られた値を用いた．表 2 に示した値を大幅に変えない限り，予測正解率や木のノード数は大きく変化しないことが予備実験の結果により示されている．RSAT では世代交代を行なわないため，解候補決定木の生成・評価ルーチンは，SESAT および ESAT における決定木構成子の選択回数と同じ回数繰り返した．

SESAT，ESAT，RSAT において，訓練事例で学習を行なった際の平均適応度の一部を表 3 に示す．SESAT と ESAT の平均適応度は，どのデータにおいても近い値となったが，RSAT はいずれのデータでも他より低い適応度となった．図 6 は iris の各世代における最良個体の適応度推移である．SESAT と ESAT の 1 世代目の最良個体は，ランダムに生成した N_{tp} 個の初期集団個体のうちの 1 つなので，RSAT については生成ルーチンを N_{tp} 回繰り返した時点を 1 世代目として表示している．いずれのシステムにおいても，8000 世代以降は最良個体の適応度が変化していない．このグラフから，SESAT と ESAT では着実な学習が行なわれていることが読み取れる．GA による探索のランダムサーチに対する優位性はしばしば議論されるが，以上の結果は GA が本手法の決定木生成においても有効であることを裏付けるものである．

各システムのテスト事例における平均正解率と木の平均ノード数を表 4 に示す．平均正解率はデータごとに多少のばらつきがあるが，全データの母平均の差の検定では SESAT，ESAT，C5.0 の間に有意差は見られなかった．また，RSAT は他と比較してテスト事例における正解率が低いとの結果が得られた．平均ノード数については，RSAT，SESAT，ESAT，C5.0 の順で少ないという検定結果が得られた．平均学習時間は SESAT，ESAT，RSAT でそれぞれ 14 秒，9 秒，88 秒となり，1 秒程度で決定木を生成する C5.0 を大幅に上回った．

*2 aust による決定木のうちの 1 つを除いて，非終端ノードの子ノード数は 4 以下となり，pima による決定木のうちの 3 つを除いて，高さは 10 以下となった．

表 3 訓練事例における平均適応度 (括弧内は標準偏差)

データ	SESAT	ESAT	RSAT
breast	94.9 (0.2)	94.9 (0.2)	92.9 (0.4)
heart-c	81.8 (0.6)	82.7 (0.8)	76.5 (1.2)
monks2	72.4 (2.2)	71.5 (3.6)	66.4 (1.4)
全平均	80.7	81.0	75.8

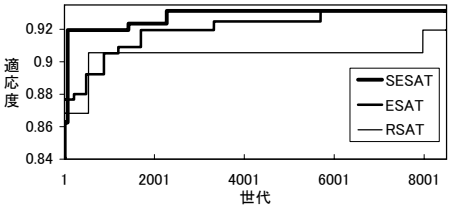


図 6 適応度の推移

4. 考　　察

本章では，3 章の実験結果に基づいて，決定木生成における共生進化の適用可能性，適応度の定義の有効性，共生進化と一般的な GA の違いについて考察する．

SESAT で生成される決定木では，ラベル平均に対応した属性値を閾値とする範囲のみがアークを辿る条件となる．一方，C5.0 による決定木では，個々の訓練事例の属性値から求めた値を閾値とする範囲や，属性値の等値関係をも条件とすることができる．このように SESAT と C5.0 とでは生成可能な決定木が異なるが，両者の予測正解率は同程度となっている．SESAT において，学習時間は C5.0 と比べて大幅に劣るものの，予測正解率を C5.0 と同程度に保ちつつ，より簡素な木が生成されたことから，決定木生成における共生進化の適用可能性に注目したい．

次に，適応度の定義の妥当性について考える．SESAT と ESAT では，RSAT より訓練事例における適応度とテスト事例分類の正解率がともに高く，ノード数の多い決定木が生成された．式 (2) で定義された適応度に基づいて進化した決定木が過学習をしているならば，訓練事例における適応度と木のノード数がともに大きくなり，テスト事例における正解率は低くなるはずである．また，ノード数のみから考えると，他よりノード数の少ない決定木を生成した RSAT は，過学習の可能性が低いと予測できる．ところが，SESAT と ESAT は，RSAT よりもノード数の多い木を生成したにもかかわらず，テスト事例における正解率が高いことから，式 (2) によって過学習回避に有効な適応度が定義できたと考えられる．

最後に，共生進化と一般的な GA で生成される決定木

表 4 テスト事例における平均正解率と平均ノード数 (括弧内は標準偏差)

データ	SESAT		ESAT		RSAT		C5.0	
	正解率 [%]	ノード数	正解率 [%]	ノード数	正解率 [%]	ノード数	正解率 [%]	ノード数
aust	84.9 (4.1)	16.4 (3.3)	85.0 (4.0)	15.4 (6.7)	85.2 (3.8)	7.1 (2.5)	84.5 (5.1)	27.0 (9.9)
balance	78.7 (4.9)	26.5 (5.1)	80.1 (4.8)	40.4 (5.6)	68.8 (7.3)	9.7 (6.6)	77.3 (5.0)	84.0 (9.9)
breast	95.9 (2.3)	16.1 (2.9)	95.8 (2.2)	19.1 (5.6)	95.2 (2.4)	7.8 (3.6)	96.3 (2.3)	18.2 (4.9)
bupa	63.1 (8.0)	18.4 (3.8)	62.4 (9.3)	26.1 (8.0)	61.7 (8.5)	9.8 (4.1)	64.1 (7.0)	50.2 (10.3)
glass	63.1 (11.6)	20.0 (3.0)	65.3 (11.5)	31.3 (5.9)	54.4 (12.4)	15.0 (5.6)	69.5 (11.7)	47.4 (6.1)
heart-c	77.4 (7.6)	21.6 (3.5)	77.6 (8.3)	31.6 (6.8)	71.4 (9.6)	9.4 (4.2)	77.6 (10.3)	36.8 (6.0)
iris	95.6 (5.1)	6.9 (1.7)	95.5 (5.4)	6.4 (0.8)	92.4 (8.8)	5.0 (1.5)	92.7 (6.6)	8.0 (1.4)
monks1	99.6 (2.5)	14.3 (0.8)	98.8 (3.6)	14.2 (1.3)	76.3 (11.3)	11.0 (4.4)	80.8 (11.1)	21.1 (5.1)
monks2	67.6 (12.8)	21.5 (4.8)	64.8 (13.2)	24.2 (10.9)	63.4 (12.7)	9.8 (4.7)	56.5 (11.8)	21.1 (13.8)
monks3	92.5 (5.2)	7.0 (2.9)	92.6 (5.3)	6.4 (2.5)	92.3 (6.8)	6.5 (3.7)	93.3 (5.3)	12.0 (0.0)
pima	73.3 (6.4)	20.3 (4.1)	73.0 (7.1)	26.3 (6.7)	73.6 (6.1)	6.9 (2.9)	74.2 (7.6)	50.2 (13.4)
post	72.3 (17.7)	9.8 (2.3)	74.1 (18.7)	9.5 (2.2)	68.6 (18.2)	7.2 (2.3)	68.8 (17.9)	4.6 (3.9)
全平均	80.3	16.6	80.4	20.9	75.3	8.8	78.0	31.7

の違いに着目する．SESAT と ESAT により生成された決定木を比較すると，SESAT の方が簡素な決定木となっている．解候補決定木の交叉では，有効なビルディングブロックを取り入れられる反面，木が大きくなってしまうが，SESAT では sprig の遺伝子を反映するという子の生成をも行なっている．適応度の高い sprig の遺伝子を解候補決定木に反映することで，ノード数を増加させることなく，有効な解の一部であるビルディングブロックを取り入れることが可能となる．部分解と全体解を並行して進化させる共生進化の特長が，SESAT により生成された決定木の簡素さに表れたと考えられる．

以上の考察より，正解率と正解局在率から求めた適応度によって解候補を評価し，さらに探索手法として共生進化を利用する手法は，簡素な決定木生成に有効な一手法といえる．

5. お わ り に

本研究では，予測正解率が高く簡素な決定木の生成を目的として，共生進化による新しい決定木構築法を提案した．高さ 1 の部分木，および部分木の組み合わせにより構成される決定木をそれぞれ個体とする 2 集団を保持し，正解率と正解局在率により算出した適応度に従って 2 集団を並行して進化させる方法である．提案手法による決定木生成システム SESAT を構築し，UCI リポジトリのデータを用いて評価実験を行なった結果，SESAT は C5.0 と同程度の予測正解率を持ち，ノード数が平均 70%程度の木を生成することが示された．学習時間に関しては C5.0 の方が優れているが，学習時間が問題にならず，簡素な決定木の生成を目的とする場合には，共生進化に基づく手法が有効であると考えられる．

◇ 参 考 文 献 ◇

[Angeline 98] Angeline, P.: Subtree crossover causes bloat, in *Proc. 3rd Genetic Programming Conf.*, pp. 745–752 (1998)

[Breiman 84] Breiman, L., Friedman, J., Olshen, R., and Stone, C.: *Classification and Regression Trees*, Wadsworth & Brooks (1984)

[Goldberg 89] Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989)

[Moriarty 95] Moriarty, D. and Miikkulainen, R.: Efficient Learning from Delayed Rewards through Symbiotic Evolution, in *Proc. 12th Intl. Conf. on Machine Learning*, pp. 396–404 (1995)

[Moriarty 96] Moriarty, D. and Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, *Machine Learning*, Vol. 22, pp. 11–32 (1996)

[Moriarty 98] Moriarty, D. and Miikkulainen, R.: Hierarchical Evolution of Neural Networks, in *Proc. IEEE World Congress on Computational Intelligence*, pp. 428–433 (1998)

[大谷 04] 大谷 紀子, 志村 正道：決定木生成のための共生進化に関する考察, 電子情報通信学会技術報告, Vol. 103, No. 725, pp. 101–106 (2004)

[Quinlan 86] Quinlan, J.: Induction of Decision Trees, *Machine Learning*, Vol. 1, No. 1, pp. 139–159 (1986)

[Quinlan 93] Quinlan, J.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993)

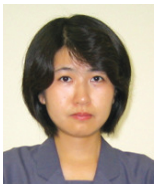
[佐藤 97] 佐藤 浩, 小野 功, 小林 重信：遺伝的アルゴリズムにおける世代交代モデルの提案と評価, 人工知能学会誌, Vol. 12, No. 5, pp. 734–744 (1997)

[UCI] http://www.ics.uci.edu/~mlearn/MLRepository.html

〔担当委員：沼尾正行〕

2004 年 3 月 4 日 受理

著 者 紹 介



大谷 紀子(正会員)

1993 年東京工業大学工学部情報工学科卒業．1995 年同大学院理工学研究科情報工学専攻修士課程修了．同年キヤノン (株) 入社．同社情報メディア研究所にて情報検索の研究に従事．2000 年東京理科大学理工学部経営工学科助手．2002 年武蔵工業大学環境情報学部情報メディア学科講師．情報処理学会，電子情報通信学会各会員．

志村 正道(正会員)

1960 年東京大学工学部応用物理学科卒業．1965 年同大学院博士課程修了．工学博士．同年大阪大学基礎工学部助教授．1976 年東京工業大学工学部助教授，同教授，1997 年東京理科大学理工学部教授，2001 年武蔵工業大学環境情報学部教授．この間パターン認識，人工知能，学習機械などの研究に従事．著書に機械知能論（昭晃堂），人工知能（森北出版）などがある．電子情報通信学会，情報処理学会，ACM，IEEE，AAAI 各会員．