2025 年度卒業研究概要

共生進化に基づく決定木生成における遺伝子型の改良

大谷 紀子 研究室
2272083　深代 夏帆

1. 背景と目的

共生進化（Symbiotic Evolution）とは遺伝的アルゴリズムの一手法であり，部分解集団と全体解集団を並行して進化させることで最適解を探索する手法である．部分解は解の一部であり，全体解は部分解の組み合わせにより表現される解候補である．共生進化は，部分解集団と全体解集団を並行して進化させることで，局所解への収束を回避した最適解探索を可能とする．

共生進化を用いた先行研究として，決定木生成システム SESAT（Symbiotic Evolution for Simple and Accurate Trees）がある[1]．SESAT は，予測正解率が高く簡素な決定木の生成を目的として提案された．当該システムでは，高さ 1 の部分木を部分解とし，複数の部分木から構成される決定木を全体解とする．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決定木の簡素化に寄与している．SESAT で生成される決定木は, C5.0 で生成された決定木と比較して，ノード数を 70%程度に削減しつつ正解率を同程度に保つことが示されている．

共生進化は原則, 離散最適化問題を対象とする．SESAT において部分解の持つ遺伝子は整数値のみで表され，決定木でのデータ分類のために設定されるアーク間の閾値が取り得る値も限定される．結果として, SESAT で生成される決定木の正解率向上には限界が生じる．

本研究では，共生進化により生成される決定木の正解率向上を目的として，アーク間の閾値として任意の実数を取ることができる手法を提案する．また，提案手法により決定木を生成するシステム RCSESAT （ Real-Coded Symbiotic Evolution for Simple and Accurate Trees）を構築し，提案手法の有用性を検証する．

2. 提案手法

SESAT で用いられる従来の部分木の遺伝子型は, 部分木の子ノード数の上限値を M とした場合, 要素数 M+1 の整数値配列である. 配列の 1 番目には根ノードに割り当てられる属性番号が格納される．残り M 個の要素には，葉ノードに割り当てられるクラス番号，非終端ノードを示す−1，ノードがないことを示す−2 のいずれかが格納される．各アークには遺伝子座に対応する番号が付与され，たとえば 2 番のアークがクラス 5 のノードに接続される場合，配列の 3 番目に 5 が格納される．従来手法では，アーク間の閾値は前述の各アークに割り当てられた番号から算出されるため，取り得る値が限られる．

一方，提案手法における部分木の遺伝子型は，要素数がそれぞれ M+1 である整数値配列と実数値配列から構成される．整数値配列の 1 番目には属性番号が格納される．2 番目以降にはクラス番号または−1 が, 部分木において左に位置するものから順に格納される．上記処理後，配列の余剰要素には−2 が格納される. 実数値配列は各アーク間の閾値および属性値の最小値と最大値を昇順で保持するため，閾値は属性値の範囲内の実数すべてを取り得る．

提案手法における部分解集団は，進化過程では属性番号ごとに分割され，全体解を生成する際にはすべての個体を含む 1 つの集団として扱われる．

また，部分解の交叉手法は一点交叉とする．交叉点にあたるアーク間の閾値は，当該 2 アークの両側にある隣接アークとの閾値を上限および下限とし，一様分布に従う乱数で決定する．突然変異は各部分解に対し等確率で発生させる．突然変異発生時には，対象の部分解が保持する属性番号，クラス番号，閾値のすべてを，それぞれの取り得る値の範囲内でランダムに決定し直す．

全体解 T の適応度$tfit(T)$は，RCSESAT，SESAT のいずれにおいても式(1)により算出する．$acc(T)$ は訓練事例における T の正解率，$bias(T)$は T における葉ノード間での正解数のばらつきを示す正解局在率，α は正解局在率を考慮する度合いを表す定数である．

$$tfit(T) = acc(T) \cdot (1 - \alpha \cdot bias(T)) \qquad (1)$$

3. 評価実験

提案手法の有用性を示すため，UCI 機械学習リポジトリの 4 種類のデータセットを用いて評価実験を実施した．本稿では，連続値属性を 9 個もつデータセット glass，およびカテゴリカル属性を 4 個もつデータセット balance について報告する．なお，balance の属性値はすべて順序尺度である．式(1)における α を 0 から 1 まで 0.2 刻みで変化させ，5-fold クロスバリデーションによる決定木生成を RCSESAT と SESAT で 10 回ずつ繰り返した．両システムにおけるパラメータのうち，部分解集団の個体数，突然変異確率，世代交代数は，評価実験に使用しないデータセット 3 種類を用いた予備実験の結果に基づき決定した．その他の各パラメータには，SESAT のデフォルト値を用いた．

4. 考察

glass および balance に関して，両システムの訓練事例およびテスト事例における α ごとの平均正解率, 平均ノード数をそれぞれ図 1, 図 2 に示す．

glass では，訓練事例とテスト事例のいずれにおいても，ノード数が同程度の場合，RCSESAT の正解率が SESAT の正解率を上回った. RCSESAT は，

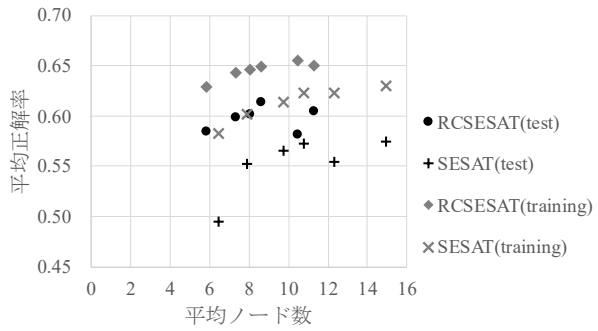

図 1 glass における平均正解率，平均ノード数

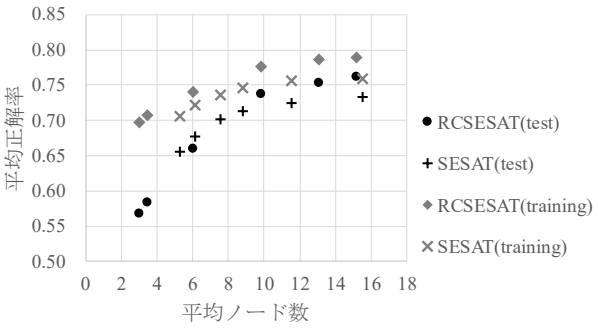

図 2 balance における平均正解率，平均ノード数

提案手法により SESAT と比べて適切な閾値を設定できたため, 高い正解率を示したと考えられる．

一方，balance のテスト事例において，RCSESAT の正解率はノード数が減少すると大幅に低下する傾向にあり，ノード数 6 付近では SESAT の正解率を下回った．しかし，訓練事例においてはノード数が変化しても正解率が安定しており，ノード数 6 付近でも SESAT の正解率を上回った. RCSESAT は, 訓練事例においては SESAT よりも高い正解率を示したため，決定木の最適化に提案手法が寄与しているといえる．しかし，テスト事例においては，ノード数の少ない場合に正解率が大幅に低下したため, RCSESAT が過学習を起こしている可能性が高い. 閾値の詳細な設定が可能であるために，訓練事例に過度に適合したと考えられる．

今後の課題として，整数値を多く含むデータセットにおける過学習抑制手法の検討が挙げられる．

参考文献

[1] 大谷紀子，志村正道, “共生進化に基づく簡素な決定木の生成,” 人工知能学会論文誌, Vol.19, No.5, pp. 399-404, 2004.