2022 年度卒業研究概要

共生進化を用いた複数の C4.5 演習問題における訓練事例集合の自動生成

大谷 紀子 研究室
1972048　今野 直樹

1. はじめに

本学では，情報システム学科所属の学生を対象に,「人工知能とデータマイニング」という講義が開講されており，データマイニングの 1 手法である決定木生成アルゴリズム C4.5 が取り上げられている．決定木は分類規則の木構造による表現方法であり，C4.5 では事例集合を基に決定木を生成する．授業内では，問題文と事例集合から構成される演習問題が 1 つ提示され，全員で解き進めながら教員によって解説される．学生の理解をさらに深めさせるには演習問題を解かせることが有効だが，自力で取り組む状況を作るには学生ごとに異なる演習問題を課すことが必要である. しかし，演習問題を学生の数だけ手作業で作成することは困難である．

以上の問題を受け，森は演習問題を自動で生成するシステムを構築した[1]. 森が構築したシステムは, ランダムに生成した複数の演習問題のうち，学習する上で重要なポイントを最も押さえている演習問題と解答を出力する．しかし，複数の異なる演習問題の生成を想定していないため，複数回実行すると生成された演習問題の難易度に差が生じる．また，生成される演習問題において，分岐する属性や終端ノードの表すクラスが一意に決まるとは限らないので，解答する際の条件を問題文に付けなければならない．現在は，森が構築したシステムで生成した演習問題をレポート課題として学生に課しているため，学生は条件を考慮してレポート課題を解かなければならず，学生ごとの難易度の差により不公平感を抱くこともある．

本研究では，学生が演習問題に対して抱く不公平感の解消を目的とし，C4.5 を題材とした同難易度の異なる演習問題が自動で複数生成される手法を提案する．

2. 提案手法

提案手法では，同難易度の異なる演習問題および解答を指定された数だけ生成する．演習問題の生成には，遺伝的アルゴリズムの 1 手法である共生進化を用いる．共生進化は，部分解を個体とする部分解集団と，部分解の組み合わせを個体とする全体解集団を並行して進化させる点を特徴とする．演習問題生成にあたっては，1 つの演習問題の訓練事例集合を部分解とする．部分解の染色体は 69 個の遺伝子で構成され, 0 番目の遺伝子は訓練事例の数, 1～3 番目の遺伝子は各属性の値の種類数, 4 番目の遺伝子はクラス値の種類数を表す. また, i 個目の訓練事例の各属性を j 個目とするとき，$4i+j$ 番目の遺伝子は訓練事例の各属性値を表し，$4i+4$ 番目の遺伝子は訓練事例のクラス値を表す．全体解の染色体は演習問題数分の部分解へのポインタの列で構成される．演習問題数を i 個とするとき，0～i-1 番目の遺伝子の値は部分解へのポインタを表す.

全体解の評価値は参照する部分解の単問品質の標準偏差とし，部分解の評価値は参照されている全体解の中で最も評価値が高い全体解の評価値とする．全体解に参照されていない部分解の評価値は最も悪い値とする．単問品質は染色体から生成される演習問題の質の良さを表す値であり，部分解 p の単問品質 $q(p)$は式(1)で算出される．

$$q(p) = \frac{\sum_{i=1}^{4} w_i v_i(p)}{\sum_{i=1}^{4} w_i vmax_i(p)} \times \frac{3 - c(p)}{3} \qquad (1)$$

ここで，$v_1(p)$〜$v_4(p)$はそれぞれ p が表す演習問題における非終端ノード数，ノード数，生成される木の高さ，決定木生成時の計算に用いられる対数の数に関する難易度で，w_1〜w_4 はそれぞれ $v_1(p)$〜$v_4(p)$に対する重み，$vmax_1$〜$vmax_4$ はそれぞれ $v_1(p)$〜$v_4(p)$の最大値，$c(p)$は満たされている制約条件の個数を表す．

部分解集団では 3 個の個体がエリート保存によって次世代へと受け継がれる．また，ルーレット選択によって選択された親個体をもとに一様交叉し次世代の個体が生成される．全体解集団では 3 個の個体がエリート保存によって次世代へと受け継がれる．また，ルーレット選択によって選択された親個体をもとに一点交叉し次世代の個体が生成される．

3. 評価実験

提案手法の有用性を示すために実施した評価実験では，「人工知能とデータマイニング」を修得済みの学生 11 名を被験者とした．当科目の受講生は例年およそ 100 名であるため，既存システムと提案手法で 100 個ずつ演習問題と解答を生成し，それぞれの最高評価値問題と最低評価値問題を被験者に提示した上でアンケート調査を実施した．アンケートでは，最高評価値問題と最低評価値問題の難易度に差を感じたかについて 1〜4 の 4 段階で評価させ，評価理由を自由記述形式で回答させた．また，最高評価値問題と最低評価値問題について，計算量や生成される木の高さ，対数計算の複雑さ，難易度，学習への有効度を 1〜5 の 5 段階で評価させた．

得られた評価値の平均を表 1 に示す．難易度の差について問う評価項目に関しての t 検定の結果，既存システムと提案手法の評価値間に有意な差が見られた（$t(10)=1.812$，$p<0.05$）．したがって，演習問題について感じる難易度の差が小さくなったといえる．難易度の差の評価理由として，既存システムで生成した演習問題に対しては「ページ数や枝の多さ等に差があると感じる」などが挙げられた一方，提案手法で生成した演習問題に対しては「最終的な木の深さや計算量にあまり差を感じなかったため」などが挙げられた．

表 1 評価値の平均

評価項目	既存システム		提案手法	
	低評価	高評価	低評価	高評価
難易度の差	3.18		2.45	
計算量	2.64	3.28	3.09	2.64
木の高さ	4.00	4.23	2.27	3.27
対数の複雑さ	2.45	2.55	2.64	2.73
難易度	2.55	3.00	2.55	3.00
有効度	3.73	4.09	3.73	4.27

4. 考察

演習問題について感じる難易度の差が小さくなったため，提案手法によって解答者は演習問題の違いによる不公平感を抱きにくくなったといえる．しかし，生成した演習問題の難易度の差について「かなり差がある」と回答した被験者から「終端ノードに辿り着くまでの計算回数に明確な差があると感じたため」という理由が得られた．提案手法では，単問品質の算出に非終端ノード数を用いることで分岐の回数が考慮されていたが，四則演算の回数自体は考慮されていなかったため，計算回数の差による不公平感が生じたと考えられる．決定木の生成過程における四則演算の回数を単問品質の算出に用いることで，不公平感が抱かれにくい演習問題を生成できると考えられる．

今後，難易度の調整機能を実装することで，学習者の習熟度に合わせた演習問題を提供することができ，C4.5 に対する理解度の向上が期待できる．

参考文献

[1] 森健太郎, “データマイニングの演習問題自動作成”, 武蔵工業大学環境情報学部情報メディア学科卒業論文, 2009.