2022 年度卒業研究概要

粒子群最適化による C4.5 の事後枝刈り演習問題自動生成手法の提案

大谷 紀子 研究室
1972079　名護 貴裕

1. はじめに

本学で開講している「人工知能とデータマイニング」という講義の内容に，決定木生成アルゴリズム C4.5 がある．C4.5 では，訓練事例を正しく分類できるように決定木を形成した後，過学習を回避するために事後枝刈りをする．

C4.5 に対する理解度を向上させるためにはレポート課題が有効であり，学生が自力で課題に取り組む状況を作り出すためには，学生ごとに異なる問題を課す必要がある．しかし，手作業で学生数分の演習問題を作ることは困難である．以上の問題を受け, 森は C4.5 の決定木形成演習問題を自動生成するシステムを作成している[1].ランダムに生成した複数の演習問題のうち，もっとも学習効果が高い問題を一つだけ出力する．現在講義では，森のシステムを学生数分繰り返し実行することでレポート課題を作成しているが，生成される問題の難易度にばらつきがあるため，学生は不公平感を持つ．また，森のシステムは訓練事例を用いた決定木形成演習問題を生成するのみで，テスト事例を用いた reduced-error pruning による事後枝刈り演習問題を生成していない．そのため，講義を受けている学生の事後枝刈りへの理解が進まないという問題がある．

本研究では，学生の C4.5 に対する理解度，演習問題の不公平感の解消を目的として，事後枝刈りに関する演習問題の自動生成手法を提案する．

2. 提案手法

提案手法ではまず，与えられた N 問の決定木形成演習問題 P_1~P_N それぞれについて，学生の負担量の指標となる計算スコア $S(P_i)$を求める．次に，決定木形成演習問題と事後枝刈り演習問題を合わせた演習問題の難易度が N 問の間で同程度となるように，事後枝刈り演習問題 P'_i における枝刈りの要否判定回数を目標難易度 $D(P'_i)$とし，$S(P_i)$を考慮して決定する．続いて，決定木形成演習問題 P_1～P_N それぞれに対応する事後枝刈り演習問題 P'_1～P'_N を粒子群最適化(PSO; Particle Swarm Optimization）によりひとつずつ生成する．PSO とは鳥や魚が効率よく群れで餌を探す行動を模倣した最適解探索アルゴリズムである．問題に対する解の候補を鳥や魚の位置として表現し，餌場への近さを解の良さとして，鳥や魚をより良い位置に向かって移動させることで最適解を導く．P'_i を生成する際の PSO における解候補 c は，訓練事例の数を n，属性数を m としたとき，$n(m+1)+1$ 次元のベクトルで表現される．第 1 成分はテスト事例の数，第 2 成分以降は各データの内容を表す．解候補 c の評価値 $V(c)$は $f_1(c)$～$f_4(c)$の和で算出される．$f_1(c)$は枝刈り要否判定回数が目標値にどれだけ近いかを表し，枝刈り目標回数 $S(P_i)$と，生成した事後枝刈り演習問題 c における枝刈り回数 $PN(c)$を用いて式(1)によって算出される．

$$f_1(c) = \left| \frac{S(P_i) - PN(c)}{S(P_i)} \right| \qquad (1)$$

$f_2(c)$は与えられた訓練事例に含まれる全属性値をテスト事例でどれだけ使用しているかを表し，決定木形成演習問題 P_i に含まれる属性値とクラスの数 $AC(P_i)$,解候補 c に含まれる属性値とクラスの数 $UseAC(c)$を用いて式(2)によって算出される．

$$f_2(c) = \left| \frac{AC(P_i) - UseAC(c)}{AC(P_i)} \right| \quad (2)$$

$f_3(c)$は枝刈りする場合としない場合の両方を含んでいるかどうかを表しており，枝刈りをする場合としない場合の両方が含まれているときは 0，片方のみのときは 1 とする．$f_4(c)$は誤分類数が代表的クラスのノードと子ノードで同数になることを 1 度だけ含んでいるかどうかを表しており，解候補 c において誤分類数が同数であった回数 $JP(c)$，生成した事後枝刈り演習問題における枝刈り回数 $PN(c)$を用いて式(4)によって算出される．

$$f_4(c) = \left| \frac{1 - JP(c)}{PN(c)} \right| \quad (4)$$

解候補 c の評価が高いほど評価値 $V(c)$は小さい値となり，$V(c)$=0 になった場合，問題 c を出力する．

3. 評価実験

生成された問題を解いたときの事後枝刈りに対する理解度を確認する実験 A，および生成された問題の公平性を確認する実験 B を実施した．実験 A では「人工知能とデータマイニング」を履修済みの学生 11 名を対象とし，事後枝刈りについて説明した後，提案手法により生成された事後枝刈り演習問題を 1 問配布した．被験者には配布した演習問題を解いたのちに，答え合わせを行ったうえで事後枝刈りの理解度についてのアンケートに回答させた．「理解できた」「少し理解できた」「どちらともいえない」「あまり理解できなかった」「理解できなかった」の 5 項目から回答させた結果，「理解できた」が 9 人，「少し理解できた」が 2 人であった．また問題を解くにあたって悩んだ点について，対象事例を数え上げるのに苦労したという意見もあった．実験 B では 3 つの決定木形成演習問題 a,b,c に対して提案手法により生成した事後枝刈り演習問題 A,B,C を用いる．「人工知能とデータマイニング」を履修済の学生 13 人を対象とし，事後枝刈りについて説明した後，A,B,C 間，および a+A,b+B,c+C 間における不公平感について

表 1　不公平感調査における回答人数

項目	A,B,C	a+A,b+B,c+C
感じた	3	4
やや感じた	7	9
どちらともいえない	2	0
あまり感じなかった	1	0
全く感じなかった	0	0

のアンケートに回答させた．なお使用した 3 問の演習問題は，100 問の決定木形成演習問題から計算したスコア $S(P_i)$の最小，中間，最大スコアの問題である．表 1 に A,B,C のみで見比べたとき，a+A,b+B,c+C で見比べたときの不公平感調査の回答人数を示す．また不公平感を抱く理由について，a,b,c 内の計算量によって難易度差や不公平感を生むという意見が多かった．さらに問題を見て感じたことについて，事後枝刈りの際ノードに到達する事例の数が 0 個だと，枝刈りの判断が楽に感じたという意見もあった．

4. 考察

実験 A では全員が事後枝刈りに対しておおむね理解できた結果となったため，提案手法で生成した演習問題は事後枝刈りの理解を深めるための問題として有用であるといえる．しかし実験 B では事後枝刈り演習問題で難易度に差を生むことはできたが，事後枝刈り演習問題の難易度差を用いて決定木形成演習問題の難易度差や不公平感を埋めることはできなかった．決定木形成演習問題の計算量によって不公平感が生まれるという意見が多く，事後枝刈り演習問題で全体の不公平感を埋めることは難しいといえる．しかし事後枝刈り演習問題で難易度差や不公平感を埋めるには，用いた評価値 $f_1(c)$～$f_4(c)$のほかに，事後枝刈りをするときにノードに到達する事例数が 0 個になることや，数え上げの作業量も考慮することが必要である．

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