建築基礎被害調査事例からのラフ集合に基づく決定ルールの導出とその評価

関西大学大学院 総合情報学研究科 学生会員 ○西村 文宏
関西大学 総合情報学部 正 会 員 広兼 道幸
関西大学 総合情報学部 正 会 員 古田 均

1. はじめに

知識獲得に関する研究の目的は，知識ベースシステムを構築する際に必要となる知識，すなわち様々な属性間の関連や重要度を発見することにあるといえる．現在の知識獲得研究における最大の関心事は，データベースなどに蓄えられた膨大なデータから，如何に知識を抽出するかという事である．ラフ集合論は，近年このような知識獲得の分野で注目されている手法のひとつである．そこで，著者らは斜面崩壊危険度診断事例からルール型知識を抽出するためにラフ集合の適用を試みた[1)]．しかし，ここではデータサイズの増加に伴い計算時間が指数関数的に必要となる問題が残された．そこで，計算時間の問題を解決するため，個々の属性値を 1 つの遺伝子座と考え，遺伝的アルゴリズムとラフ集合を併用した知識獲得手法を提案した[2)]．しかし，この手法においてもデータサイズの増加に伴い，準最適解への収束が困難になるという問題が示された．そこで，これらの問題を解決するための方法として，ルールの絞り込みを行い，残されたルール群に対して GA を適用する手法を提案した[3)]．ところが，この手法では未知の事例に対しての正答率が高くないという問題が示された．そこで本研究では，未知の事例に対してより多く正答できる知識を獲得することを目的として，GA における遺伝子の評価関数に，項目ごとの重みを付加する方法を提案する．評価関数の各項目の重みを様々に変化させることで，正答率の高い極小のルール群を見つけ出す方法を検討した．

2. システムの概要

事例が決定表の形で与えられると，まずは，ラフ集合を用いてルール群の導出を行う．最小となるルール群の導出は組み合わせ最適化問題となるので，ルールをあらかじめ絞り込むことで計算量を削減する．残ったルール群に対して，GA を用いて最適化を行う．GA では，ルーレット選択を基本として，エリート戦略を併用する．その際の遺伝子の評価方法として，式(1)に示す評価関数を用いる．最終的に，評価値が全く同じ値となる簡約化された複数の決定表が得られる．

$$F = \frac{BF}{R_{val}^{W_1} \cdot R_{attr}^{W_2} \cdot R_{rule}^{W_3}} \cdot 0.2^{N_{nc}} \tag{1}$$

GA を用いて最適化を行う過程で利用する遺伝子の評価は式(1)で行い，完全性・無矛盾性・取り除いた要素の数・取り除いた条件属性の数・減少したルール数を考慮したものとし，導出される評価値の高さでそれぞれの個体を評価する．

ここで，BF は基準となる適応度であり，一律して 200 という値を用いている．R_{val}，R_{attr}，R_{rule} はそれぞれ，決定表中の，条件の数(*val*)，条件属性数(*attr*)，ルール数(*rule*)が，元の決定表でのそれぞれの数に対してどれだけの割合にまで減っているかを示す．これらの割合が低いほど簡潔な決定表であると考えることができる．そのため，式(1)では，これらの値が低いほど，適応度(F)が高く評価される．また，N_{nc} は，その個体によって表現される決定表（ルール群）を，元の決定表の各事例に当てはめたとき，どのルールにも当てはまらなかった事例の数である．N_{nc} は決定表の不完全性を示す指標であり，N_{nc} が 1 以上であれば，ペナルティが与えられる．

本研究では，この R_{val}，R_{attr}，R_{rule} の 3 つの条件に対する重み W_1，W_2，W_3 を様々に変化させることで，未知の事例に対する正答率を高めることができるかどうかを検証した．

Keywords: 建築被害，知識獲得，ラフ集合，遺伝的アルゴリズム
連絡先: 〒569-1095 高槻市霊仙寺町 2-1-1 関西大学総合情報学部 TEL 0726-90-2402 FAX 0726-90-2402

3. 建築基礎被害調査事例への適用

本研究では，重みのかけ方を12通り用意した．また，正答率の評価は，k-fold cross-validation法を採用した．すなわち，50件の事例からランダムに選んだ40件を学習用データ，残りの10件をテストデータとした検証用データを複数用意し，それぞれのデータで正答率を求めた．また本研究では，兵庫県南部地震による建築基礎被害調査事例のデータ[4]を検証に用いた．

1ケースにつき10回実行し，1回の実行につき20～30の決定表（決定アルゴリズム）が得られる．「最大」とは，そのすべての決定アルゴリズムの中で最も高かった正答率であり，「最大平均」とは，10回の実行それぞれで最大値を求め，その10個の最大値を平均したものである．表1は，C4.5での正答率が20%～80%の事例の平均結果である．重み付けケースごとに，正答率の最大値と平均値，決定表の平均ルール数，平均条件属性数，および平均総条件数を求めた．「Normal」はW_1・W_2・W_3とも同じ割合の重み付けをした場合，「+5乗」はその項目の重みを最も高くした場合，「-5乗」は最も低くした場合である．

表1 C4.5での正答率が20%～80%の事例の平均結果

重み付けケース	正答率		ルール数平均	使用条件属性数平均	総条件数平均
	最大	最大平均			
W_3 +4乗	99%	99%	553.1	21.1	1134.9
W_1 +4乗	99%	99%	554.2	21.1	1136.8
W_3 +5乗	99%	99%	554.2	21.2	1137.0
W_3 +3乗	99%	96%	273.0	19.8	560.2
W_2 +3乗	76%	59%	17.9	21.1	34.7
W_2 +4乗	74%	57%	19.0	21.8	36.5
W_2 +5乗	73%	56%	19.7	21.9	37.8
W_3 -5乗	71%	53%	17.7	11.8	34.5
W_2 -5乗	70%	51%	20.2	10.4	39.6
W_1 -5乗	69%	53%	17.6	11.8	34.5
W_1 -4乗	69%	52%	17.6	11.7	34.5
W_2 -3乗	69%	52%	18.0	11.3	35.0
W_3 -4乗	69%	51%	17.6	11.4	34.4
Normal	69%	51%	17.8	11.2	34.8
W_3 -3乗	67%	52%	18.2	11.0	35.3
W_2 -4乗	67%	50%	19.4	10.5	37.9
W_1 +3乗	66%	55%	17.4	12.0	34.1
W_1 -3乗	66%	50%	18.2	11.2	35.5
(比較) C4.5	50%		14.0	6.4	34.4

本研究では，評価のための比較対象として，C4.5を用いた．表1の最下段はC4.5での結果である．最大正答率は，どのケースもC4.5より高い値となっていることが分かる．

上位の4ケースは，正答率は高いがルール数や総条件数が非常に多すぎるため，これらのケースは除外して考える．すると，最も高い正答率を示したのは「W_2 +3乗」であった．また，それに続く2位，3位は「W_2 +4乗」と「W_2 +5乗」であった．いずれも，使用条件属性数を少なくする点を重視しないケースである．従って，使用条件属性数を少なくする点を重視しない場合，つまり，総条件数やルール数の方を重視した場合の方が，良い結果を示していると言える．

4. まとめ

本研究では，GAの評価関数の各項目に重み付けを加えることで，未知の事例に対する，正答率の高い極小のルール群を見つけ出す方法を検討した．その結果，評価関数では，使用条件属性数を重視しなければ，すなわち，総条件数とルール数を重視すれば，C4.5よりも良い結果が得られる可能性が高いことが分かった．この結果から，未知の事例の正答率を高めるためには，条件属性数を小さくするより，総条件数とルール数を小さくする方向で検討する方が良いと考えることができる．今後，多目的最適化の問題に置き換え，さらに検討が必要である．

参考文献

1) 広兼道幸・古田均・中井真司・三雲是宏：斜面の崩壊危険度診断事例からのラフ集合を用いたルール型知識の抽出方法，土木学会論文集，No,582/III-41，pp.285-294，1997.
2) 原川浩一・広兼道幸・古田均：遺伝的アルゴリズムを用いた斜面崩壊危険度診断事例からの知識獲得手法，システム最適化に関するシンポジウム，土木学会，pp.79-84，1999.
3) 西村文宏・広兼道幸・古田均・原川浩一：斜面崩壊危険度診断事例からの支持度と条件数に基づく決定アルゴリズムの導出，土木学会，pp.879-880，2002.9
4) 兵庫県南部地震建築基礎被害調査委員会：兵庫県南部地震による建築基礎の被害調査事例報告書，日本建築学会，1996.7