ラフ集合に基づく決定ルールの導出とその評価

情 00-2048　西村 文宏
指導教員　　広兼 道幸

１．はじめに

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ことである．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．しかし，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ていた．そこで，計算時間の問題を解決するため，遺伝的アルゴリズム（Genetic Algorithms :GA）とラフ集合を併用した知識獲得手法が提案された．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 を適用する手法が提案された[1].

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した．先の手法では，評価関数の各項目がすべて同じ重み付けで利用されていたため，これらの重み付け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．

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

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．

３．結果と考察

本研究では，評価関数内の各項目の重み付けのかけ方を 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[2]を用

いた．

表 1 は，結果の 1 例を正答率順にソートしたものである．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．「平均」は 10 回試行した結果の平均値，「最大」は 10 回の試行の中で最も高かった値である．「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．

本研究では，評価のための比較対象として，既存の推論システムである C4.5[3]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 より高いか同等の値となっていることが分かる．また，多くのケースの総条件数は，C4.5 より少なくなっているので，より簡潔なルールが導き出せていると言える．

図 1 は，重み付けケース別に全試行結果の正答率を平均し，C4.5 を用いた場合の正答率との差を求め，値の高い順にソートしてグラフ化したものである．

上位の 3 つ「W_1 +4 乗」・「W_3 +5 乗」・「W_3 +3 乗」は，正答率は高いものの，ルール数・総条件数が非常に大きすぎるため，これらのケースを除外して考えると，最も良い結果を示したのは，「W_2 +5 乗」であった．これは条件属性数の重み付けを最も高くした場合である．図 1 より，条件属性数の重み付けを+3 乗（「W_2 +3 乗」）以上にすれば，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，条件の数を小さくするより，条件属性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．

表 1 結果の 1 例

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	平均			
W_1 (val) +4 乗	100%	100%	557.0	21.0	1110.1
W_3 (rule) +5 乗	100%	100%	556.8	21.0	1109.8
W_3 (rule) +3 乗	100%	93%	268.7	19.8	534.9
W_2 (att) +5 乗	90%	70%	20.2	22.0	**38.4**
W_2 (att) +3 乗	80%	66%	17.3	20.4	**33.7**
W_1 (val) +3 乗	80%	61%	16.6	11.1	**32.6**
NORMAL	80%	59%	17.9	9.5	**35.1**
W_1 (val) -5 乗	80%	56%	16.7	10.7	**32.9**
W_1 (val) -3 乗	80%	54%	17.7	9.4	**34.5**
W_3 (rule) -5 乗	80%	52%	17.9	9.7	**34.7**
W_3 (rule) -3 乗	70%	50%	17.7	9.8	**34.5**
W_2 (att) -3 乗	60%	43%	18.4	9.8	**35.9**
W_2 (att) -5 乗	60%	36%	19.9	8.7	**38.9**
(比較) C4.5	60%		17.0	8.0	**47.0**

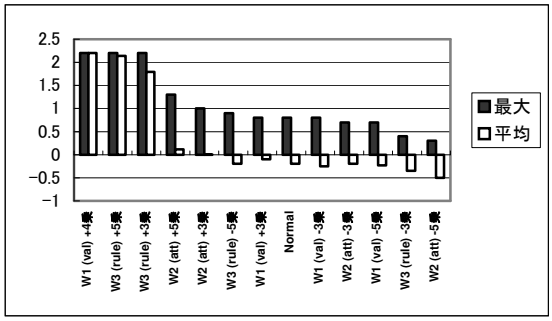

図 1 重み付け別正答率（C4.5 との差）順位

4．おわりに

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，条件属性数の少ない決定表を作成するよう評価関数を調整する方が良いことが分かった．また，得られた決定表のうち，最も良いものは，C4.5 よりも高い正答率が示せている上に，より簡潔な決定表となっているため，知識獲得手法として役立つ可能性が上がったと考えられる．

しかし現段階では，常に最も良い決定表のみを導出する術がないため，今後の課題として，得られた複数の決定表から最も良いものを選択する手法の構築や，常に良い決定表のみを出力できるような評価関数の調整が必要である．また，遺伝子コーディングの方法や遺伝操作の方法についても，より良い方法を検討していく必要がある．

参考文献

[1] 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，年次学術講演会講演概要集，土木学会，pp.879-880，2002.9
[2] 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，日本建築学会近畿支部基礎構造部会，1996.7
[3] J.R.Quinlan 著 古川康一監訳：AI によるデータ解析，トッパン，1995