―――論　文―――――――――――――――――――――――――――――――

ラフ集合と遺伝的アルゴリズムを併用した極小決定アルゴリズムの導出方法

広兼　道幸[†]　小西　日出幸[††]　宮本　文穂[†††]　西村　文宏[††††]

Extraction Method of Minimum Decision Algorithms Using Rough Sets and Genetic Algorithms

Michiyuki HIROKANE [†], Hideyuki KONISHI [††], Ayaho MIYAMOTO [†††], and Fumihiro NISHIMURA [††††]

あらまし　近年の計算機の性能向上に伴い，蓄積され利用可能なデータの量は増加の一途をたどっている．一方，計算機はこれらの蓄積された資源を有効に活用して，より高度で知的な仕事が要求されている．土木分野においても，技術者の経験などによって培われた知識の再利用は重要課題であり，そのためには知識の獲得方法，あるいは獲得知識の明示的な表現方法などの確立が必要である．本論文では，ラフ集合を用いて事例から決定アルゴリズムを導出する過程に遺伝的アルゴリズムを適用し，比較的少ない計算量で，簡潔かつ有用な決定アルゴリズム導出する方法を提案する．また，実際に架設現場での事故事例のデータから決定アルゴリズムを導出し，k-fold cross validation法によって認識率等の検証を行った．

キーワード　データマイニング，ラフ集合，遺伝的アルゴリズム，決定規則，事故事例

1. はじめに

近年の計算機の性能向上と普及に伴い，収集・蓄積され利用可能なデータの量も爆発的に増加している．一方，これらの資源を利用して，人間の意思決定や知識の再利用など，より高度で知的な仕事が要求されている．土木分野においても，このようなシステムの代表であるエキスパートシステムに関する研究が，1986年以来，様々な問題を対象として取り組まれてきた[1],[2]．我々が知的活動を行うとき，規則や常識などを状況に応じて使い分け問題に適用しているが，どのような情報を利用すべきかという知識は，主に経験によって培われていると考えられる．計算機にこのような知的な振る舞いをさせようとしたとき，この経験的知識の獲得・表現が特に重要な意味を持つと考えられる．人間の知的活動を特徴づける知識を獲得し，計算機で利用しようとする過程を知識獲得と言う．しかし，経験的知識はそれを利用する本人ですら明確に認識していないことも多く，たとえ明確に認識できたとしても表現することは非常に困難な場合が多い．エキスパートシステムの研究を通しても，このような知識獲得の難しさは指摘されている[3],[4]ところである．そのため，例えば診断記録や過去の事例など，すでに存在するデータの山から自動的あるいは半自動的に知識を獲得する機械学習的な手法に関心が移されてきた[5]．さらに最近では，計算機上に蓄積された膨大なデータの中から，一見しただけでは伺い知ることのできない有用な知識を抽出しようとするデータマイニングが注目されている[6],[7]．データマイニングは，大規模なデータを対象とする機械学習的な知識獲得手法の1つであり，わが国でも1998年度から3年間，文部省科学研究費補助金特定研究「巨大学術社会情報からの知識発見に関する基礎研究」が行われる[8]など，知識発見に関する研究は様々な分野において大きな注目を集めているところ

† 関西大学総合情報学部，大阪府
Faculty of Informatics, Kansai University, 2-1-1 Ryozenji, Takatsuki, Osaka, 569-1095 Japan
†† 日本橋梁株式会社，兵庫県
Japan Bridge Corporation, Harima, Kako, Hyogo 675-0164 Japan
††† 山口大学工学部，山口県
Faculty of Engineering, Yamaguchi University, 2-16-1, Tokiwadai, Ube, 755-8611 Japan
†††† 関西大学大学院総合情報学研究科，大阪府
Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji, Takatsuki, Osaka, 569-1095 Japan

である．

土木分野においても，この傾向は例外でなく，技術者の経験によって培われた知識，あるいは過去の事例などは，知識の継承・共有といった観点からも非常に重要なことと考えられている．このような状況の中で，土木学会では，知的情報処理に関する様々な委員会が活動を続け，年次学術講演会においても，土木の様々な側面から知的情報処理を議論する場所として，1995年より共通セッションが設けられてきた．この共通セッション内の話題の傾向として，当初は最適化問題や分類問題の解決手法として注目されている遺伝的アルゴリズムの適用[9],[10],[11]やニューラルネットワークの適用[12],[13]に関する研究が多く見られたが，ここ数年をみると，データマイニングに関する話題[14],[15],[16],[17]が多く見受けられ，知識の継承・共有の重要性を裏付けるものとなっている．著者らは，知識獲得技術の１つとして注目されている，Pawlakによって提唱されたラフ集合[18]を適用し，専門家の診断結果に内在する経験的知識を抽出する方法を提案してきた[19],[20],[21]．しかし，ラフ集合を用いた知識獲得手法では数多くの規則が抽出され，実際にそれらの抽出した知識をエキスパートシステム等で利用するためには，事例をより簡潔かつ高精度で推論可能な規則の集まり（決定アルゴリズム）を抽出する仕組みが必要と考えられる．

そこで本論文では，ラフ集合を用いて事例から決定アルゴリズムを導出する過程に遺伝的アルゴリズムを適用し，データ量に依存しない比較的少ない計算量で，簡潔かつ有用な決定アルゴリズムを導出する手法を提案する．また，提案する手法の有効性を検証するために，実際に橋梁の架設現場で発生した事故事例をデータとして用いた．最後に，導出した決定アルゴリズムの簡潔さや有用性を示すため，簡潔な知識を導出するために良く利用されているC4.5[22]による結果との比較を実施した．比較においてはルール数や属性数に基づく知識の簡潔さのみならず，導出した知識の認識率による評価も実施した．認識率の評価はデータをランダムにk分割し，$k-1$組を学習用データとして決定アルゴリズムの導出を試みた．導出した決定アルゴリズムを用いて，残りの1組を評価用データとして用いて認識率を求めた．これをk回繰り返し，その平均を最終的な認識率とするk-fold cross validation法[23]を用いた．

表１　斜面の崩壊危険度診断の例

Table 1 Example of diagnostic case of Slope-Failure Possibility

斜面	条件属性			決定属性
	高さ	勾配	長さ	危険度
S_1	中	中	長	B
S_2	中	緩	短	C
S_3	高	急	短	A
S_4	高	中	短	A
S_5	低	緩	長	C
S_6	高	緩	長	B

2. ラフ集合とGAを併用した知識獲得方法

ラフ集合の概念は，1982年にポーランドの計算機学者Zdzislaw Pawlakによって提唱された[18]もので，その本質の1つは類別にあるということができる．人間が観測によって得られるいくつかの情報をもとに判断するとき，それらの情報の対象物（決定属性）を様々な属性（条件属性）によって類別する作業を行っている．対象物がこの類別結果に対して同じものであれば，それらの対象物は識別することができず，同じものとして取り扱い，推論したり決定したりしている．以下でラフ集合と遺伝的アルゴリズムを併用した極小決定アルゴリズムの導出方法について説明する．

2.1 決定表

表1は，斜面の高さ，勾配，長さという3つの条件属性によって，斜面の崩壊危険度という決定属性を診断した例である．表1は斜面の集合を

$$U=\{S_1, S_2, S_3, S_4, S_5, S_6\} \tag{1}$$

とし，属性の集合を

$$A=\{\text{高さ}, \text{勾配}, \text{長さ}, \text{危険度}\} \tag{2}$$

と考えた情報システム$S=(U, A)$である．ここで，表1は3つの属性

$$C=\{\text{高さ}, \text{勾配}, \text{長さ}\} \tag{3}$$

によって，もう1つの属性

$$D=\{\text{危険度}\} \tag{4}$$

を決定する情報システムであると考えることができ，

$$T = (U, A, C, D) \tag{5}$$

と表され，決定表と呼ばれる．

ここで，C を条件属性，D を決定属性と呼び，C と D の識別不能関係 $U/IND(C)$，$U/IND(D)$ による同値類をそれぞれ条件クラス，決定クラスと呼ぶ[18]．表1における条件クラスと決定クラスは，それぞれ

$$U/IND(C) = \{\{S_1\},\{S_2\},\{S_3\},\{S_4\},\{S_5\},\{S_6\}\} \tag{6}$$

$$U/IND(D) = \{\{S_1, S_6\},\{S_2, S_5\},\{S_3, S_4\}\} \tag{7}$$

となる．決定表はある条件が満たされたとき，どのような決定を下すべきかをまとめたものである．これは，ある条件クラスに属する対象がどの決定クラスに分類されるかという関係（決定規則）の集まりであると言うことができる．意思決定で取り扱われる多くの問題は，このような決定表によって表現することが可能であり，様々な分野で重要な役割を果たしている[24].

2.2 決定表の縮約

決定表によって表される意思決定において，決定属性の値を決定するために，すべての条件属性が必要であるとは限らない．情報システム $S = (U, A)$ において，

$$\{a \in B \mid IND(B-a) \neq IND(B), IND(B) = IND(A)\} \tag{8}$$

なる属性の集合 B が縮約[18],[25]である．縮約は情報システム S において，対象の識別能力を損なわない範囲で簡約化された属性の集合である．特に，情報システム S が，決定表 T であるとき

$$\{a \in R \mid POS_{IND(R-a)}(IND(D)) \neq POS_{IND(R)}(IND(D)), POS_{IND(R)}(IND(D)) = POS_{IND(C)}(POS(D))\} \tag{9}$$

となる属性の集合 R は D に対する C の相対縮約と呼ばれる．ここで，$POS_X(Y)$ は条件属性 X を用いることで，決定属性 Y を正しく決定クラスに分類できるものの集合を表し，R は決定表 T において，すべての条件属性 C を用いたときと同等の決定を行うことができる簡約化された条件属性の集合となる[18]．例えば，表1における決定表における縮約は｛高さ，勾配｝，｛高さ，長さ｝となる．

さらに，決定表 T において，U のひとつの要素に注目したとき，その要素をある決定クラスに分類するために，U のすべての要素に対するすべての条件属性の値が必要であるとは限らない．ここで，ある要素をその要素が属する決定クラスに属さない，すなわち，他のすべての要素と区別するために必要な条件属性の集合

$$\{a \in R \mid [x]_{(R-a)} \not\subset [x]_D, [x]_R \subseteq [x]_D\} \tag{10}$$

を満たす条件属性の集合 R は値の縮約と呼ばれ，ある要素について得られた値の縮約は，その決定規則が偽にならない範囲で，条件属性を簡約化した決定規則であるといえる．例えば，表1の斜面 S_1 に対しては，｛高さ，勾配｝，｛高さ，長さ｝，｛勾配，長さ｝という値の縮約が存在し，そこから

高さ=中 ∧ 勾配=中 → 危険度=B (11a)

高さ=中 ∧ 長さ=長 → 危険度=B (11b)

勾配=中 ∧ 長さ=長 → 危険度=B (11c)

という簡約化された決定規則を導き出すことができる．

2.3 決定アルゴリズムの最適化

決定表 T において属性 $a \in A$ が取りうる値の集合を V_a とし，属性と値の組を (a, v) で表す．また，情報システム S において，

$$P = \{a_1, a_2, \cdots, a_n\} \subseteq A, \quad v_i \subseteq V_{a_i} \tag{12}$$

とすると

$$(a_1, v_1) \wedge (a_2, v_2) \wedge \cdots \wedge (a_n, v_n) \tag{13}$$

で表現される論理式を P -基本式と呼ぶ．また，決定表 T において，C -基本式を φ，D -基本式を ψ とすると，φ → ψ の関係は CD -決定規則と呼ばれる．例えば，前

表 2　決定アルゴリズムの例
Table 2 Example of decision algorithms

決定規則	条件属性		決定属性
	高さ	勾配	危険度
R_1	中	中	B
R_2	中	緩	C
R_3	*	急	A
R_4	高	中	A
R_5	低	*	C
R_6	高	緩	B

節の決定規則(11a), (11b), (11c)は，表1における CD-決定規則の例となる．

任意の CD-決定規則からなる有限集合は決定アルゴリズムと呼ばれ[18],[26]，情報システム S に対して，すべての $x \in U$ に対応する CD-決定規則があるとき，その決定アルゴリズムは完全であるという．また，情報システム S において，すべての決定規則が真であるとき，その決定アルゴリズムは S において無矛盾であると言う．表1に示す決定表は矛盾のない決定表であるため，決定表の要素をすべて被覆するよう留意して，得られた決定規則を組み合わせれば，与えられた決定表に対して完全かつ無矛盾な決定アルゴリズムを得ることができる．表2は表1から得られた決定アルゴリズムの一例である．表中の”*”はドントケア記号を示す．表2は，2つの条件属性と1つの決定属性から構成され，6つの決定規則からなる決定アルゴリズムで，表1に対して完全かつ無矛盾な決定アルゴリズムである．例えば，決定規則 R_1 は

$$IF \text{ 高さ=中 } and \text{ 勾配=中 } THEN \text{ 危険度=B} \qquad (14)$$

というルール型の知識で記述することができる．

一般に，決定アルゴリズムは簡潔であるほど良いと考えられる．簡潔な決定アルゴリズムは見通しが良く，そこに表現されている知識も理解しやすい．決定アルゴリズムの中で最も簡潔なものを最小決定アルゴリズムと呼び，ラフ集合を用いた知識獲得ではその導出が目的の1つとなる．しかし，決定表の1つの要素に対して複数の縮約が存在する場合が多く，複数の決定規則が導出されるため，最小決定アルゴリズムの導出は，その組み合わせを考えると膨大な数になる可能性がある．単純には，各決定規則の値の縮約によって得られる簡約化された決定規則の組み合わせをすべて検証し，最も簡潔なものを最適解とすればよい．しかし，場合によっては数十程度の決定規則しか持たない決定表であっても，数千，数万の決定規則が得られる場合があり，容易に組み合わせの爆発が発生する．

2.4　遺伝的アルゴリズムの適用

以上よりラフ集合を用いた知識獲得は，決定表からの最適化された決定アルゴリズムの導出であると言うことができる．しかし，最適化の過程には，決定アルゴリズムを全体として評価することによって，より簡潔な知識を得ることが可能であると考えられる．ただし，条件属性や決定規則の数を考慮すると，その最適化には膨大な組み合わせの試行が必要となり，現実的とは言えない．そこで本論文では，決定表から得られた決定アルゴリズムを全体として評価しながら効率良く最適な解を得るため，遺伝的アルゴリズムを用いた．

遺伝的アルゴリズムでは，解の探索や最適化を生物の進化と遺伝の様相に模して考える．ここで重要なことは，最適化の対象となるシステムを，遺伝子構造としてどのように表現し，評価するかという点である．ラフ集合による知識獲得手法において最適化の対象となるのは決定アルゴリズムであり，エキスパートシステム等で利用することを考えると，その簡潔さが評価の中心となる．そこで本論文では，完全かつ無矛盾という性質を損なうことなく，可能な限り簡潔な決定アルゴリズム（極小決定アルゴリズム）を導出することを目的として，遺伝的アルゴリズムのコーディングを行った．

2.4.1　遺伝子表現

与えられた決定表から得られたすべての決定規則と遺伝子表現，および遺伝子表現と決定アルゴリズムの関係を図 1 に示す．図 1 の左側の表に示すように，与えられた決定表における各事例から簡約化された複数の決定規則が導出される．すべての事例から簡約化された決定規則を導出し，重複したものを除いた数を Nr とする．図 1 は表 1 から決定アルゴリズムを導出する例であり，表 1 からは 13 種類の簡約化された決定規則が得られたことを示す．ここで，決定アルゴリズムは任意の決定規則の組み合わせから導き出されるものと考えた．すなわち，Nr 個の決定規則から任意の n（$1 \le n \le Nr$）個の決定規則を組み合わせることで導出することができると考えた．本論文では，この考え方

に基づき，遺伝的アルゴリズムを適用するため，図 1 の右上に示すような長さ Nr の遺伝子列を用意して，対応する決定規則を用いるか否かを 1 と 0 でそれぞれ表現した．さらに，染色体のひとつの遺伝子座が，決定表から導出されたひとつの決定規則に対応しており，遺伝子情報から，図 1 の右下の表に示すような決定アルゴリズムが得られることになる．

2.4.2 評価関数

決定表の簡約化において，抽出した決定アルゴリズムに矛盾がない無矛盾性，および元の決定表を完全に被覆する完全性を満足した上で，最も簡約化された決定表，すなわち抽出した決定アルゴリズムにおける決定規則の数および前提部の条件属性値の数が最小となる決定表を最適解と定義した．したがって，各個体の評価値は，無矛盾性，完全性，決定規則の数，および条件属性値の数を考慮して

$$F = \frac{BF}{R_{val} \cdot R_{attr} \cdot R_{rule}} \cdot 0.2^{Nnc} \qquad (15)$$

によって求めることとした．ここで，BF は基準となる適応度で，本研究では 200 を用いた．R_{val}，R_{attr}，R_{rule} は，それぞれ与えられた決定表に対する，決定アルゴリズム中の条件属性値の数，条件属性の数，および決定規則の数の割合を示す．決定アルゴリズム中の条件属性値の数，条件属性の数，決定規則の数が少なくなり，これらの値が小さいほど簡潔な決定アルゴリズムとなる．すなわち，式(15)では，これらの値が小さくなるほど適応度 F の値は大きくなり，個体としては高く評価されることになる．また，N_{nc} はその個体によって表現される決定アルゴリズムを元の決定表の各事例にあてはめたとき，どの決定規則にも当てはまらなかった事例の数である．すなわち，N_{nc} は決定アルゴリズムの不完全性を表す指標であり，1 以上であれば適応度にペナルティが指数的に与えられることになる．

2.4.3 遺伝操作

選択方法はルーレット戦略を基本としてエリート戦略を併用し[27]，突然変異は任意の 1 ビットを反転させる方法を採用した．交叉方法は 2 点交叉を用いた．ただし，通常の 2 点交叉では長さ N_{nc} の個体に対して 2 点 m，n （$m < n \le Nr$）を無作為に決定し，その間の遺伝子を入れ替える操作を行う．これに対して，提案方法では，m，n の無作為な 2 点に対して整数 k を無作為に決定して，$m' = m + k$，$n' = n + k$ なる 2 点を求め，1 つの個体に対する m と n の間の遺伝子と，もう一方の m' と n' の間の遺伝子を入れ替えることとした．このように交叉位置を整数 k だけ移動させることによって，突然変異と類似した遺伝操作を行うことになり，局所解に収束することを避けることが可能となった．また，個々の決定規則の情報をこわすことなく，次世代に遺

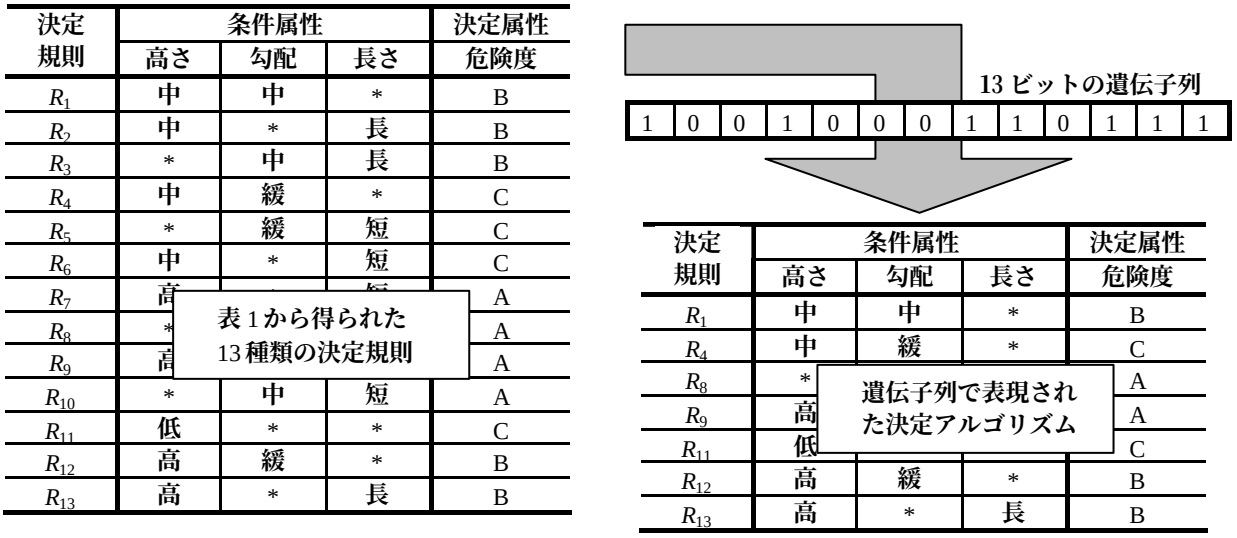

図 1　遺伝子コーディングの概念

Fig 1 Schema of Gene Cording

表 3　事故事例の書式

Table 3 Form of accident cases

事故の概要		足場解体作業中に墜落・転落					
事故に関する情報	発生日時	1997年4月14日　15時40分　月曜日			天　候	不明	
	発生場所	足場解体現場					
	作業工程	足場および支保工			作業名	足場解体	
	架設工法	固定支保工架設（タンクも含め）			架設様式	枠組式	
	事故の型	墜落・転落					
被災者に関する情報	被災者人数	1人					
	性　別	男	年　齢	59才	職　種	鳶工	経験年数	22年以上
	疾病程度	休業48日			疾病内容	手部骨折	
発生状況に関する情報	事故発生状況	時間	15時40分ごろ				
		状況	橋梁主桁の解体足場上で２段目の足場上にて３段目の鋼製布板を解体中,布板の支持部ストッパー４ヵ所のうち 1 ヶ所が外れにくいため,右手で布板を支え,左手にラジェットスパナを持って取り外しを行っていたところ,バランスを崩し約 3.8m 下へ墜落した.				
	原　因	・保護具を使用しない ・指示通りにやらなかった			対　策	・保護具の使用を徹底 ・指示の徹底	

表 4　事故情報の分類

Table 4 Classification of accident information

クラス / 属　性	1	2	3	4	5	6	7
(a) 時　　刻	0:00 ～8:00	8:00 ～10:00	10:00 ～12:00	12:00 ～13:00	13:00 ～15:00	15:00 ～17:00	17:00 ～24:00
(b) 場　　所	資材・製品等置場	足場	製品加工現場	橋脚・橋台	支保工	工場	その他
(c) 作業工程	仮設工	足場および支保工	主桁製作工	架設工	横組工	その他	
(d) 架設工法	ﾌﾟﾚｷｬｽﾄ桁架設	固定支保工架設	張出し架設	押出し架設	移動支保工架設	ﾌﾟﾚｷｬｽﾄｾｸﾞﾒﾝﾄ架設	その他
(e) 年　　齢	～20才	20～30才	30～40才	40～50才	50～60才	60才～	
(f) 職　　種	鳶工	PC工	大工	鉄筋工	運転手	その他	
(g) 経験年数	～1年	1～5年	5～10年	10～15年	15～20年	20年～	
(h) 障害程度	～30日	30～60日	60～90日	90～120日	120～150日	150日～	死亡
(i) 障害部位	頭 ※眼を除く	眼	胸・腹	肩・背中・腰	足	手	その他
(j) 事故の型	墜落・転落	挟まれ・巻込まれ	激突され	激突	飛来・落下	切れ・こすれ	崩壊・倒壊

伝させていくことも可能となった.

3. 事故事例における情報の分類

3.1　事故事例

事故事例の報告書は，企業やプロジェクトなどの組織単位で管理され，記述項目や書式が異なる場合が多い．提案方法によって抽出した決定アルゴリズムの有効性を検証するために，可能な限り多くの事故事例を収集することを考え，施工会社や現場の住所などの事故を特定できる情報は排除し，データの収集・整理を実施した．当面，表3に示す18項目に関する情報を収集して，データベースを構築した．表3に示す事故事例に関する情報は，(1) 事故に関する情報，(2) 被災者に関する情報，(3) 発生状況に関する情報の3種類に大きく分類することができる．また，実際に発生した事故に関する詳細な情報の確認も必要と考え，事故の大まかな内容が把握できる事故の概要を，作業名と事故の型を組み合わせることによって，表3の先頭行に付加した.

3.2　決定アルゴリズムの導出に用いた決定表

表3の書式に基づき記述された206件の事故事例を収

表 5　事故事例の決定表

Table 5 Decision table of accident cases

Case	事故事例に関する属性									
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
1	6	1	1	1	4	1	4	6	7	1
2	6	1	1	1	1	1	2	5	5	1
3	5	4	2	1	5	1	3	3	4	1
4	5	2	2	1	4	1	5	6	3	1
5	5	7	2	1	6	1	6	1	4	1
6	6	6	3	1	5	2	5	3	3	1
7	7	6	3	1	5	2	6	3	5	3
8	2	3	3	1	4	1	6	3	4	1
9	2	3	3	1	4	2	4	1	5	2
10	2	2	3	1	3	3	4	3	2	5
11	2	2	3	6	3	3	4	3	2	5
12	2	2	3	2	3	3	4	3	2	5
13	2	2	3	3	3	3	4	3	2	5
14	2	2	3	4	3	3	4	3	2	5
15	2	2	3	5	3	3	4	3	2	5
16	2	2	3	7	3	3	4	3	2	5
17	3	1	3	1	4	2	1	3	5	5
18	3	1	3	6	4	2	1	3	5	5
19	3	1	3	2	4	2	1	3	5	5
20	3	1	3	3	4	2	1	3	5	5
⋮ ⋮										
206	6	1	1	1	4	1	4	6	7	1

集することができた．収集した事故事例に対して，本論文で提案した決定アルゴリズムの導出方法を適用するため，事故に関するそれぞれの属性に対する情報を整理し，分類方法を検討した．収集できた事例数などを考慮して，決定規則の抽出で有効と思われる10種類の属性を表3から取り出した．表4は，これら10種類の属性と，それぞれの属性に対する分類（属性値）をまとめたものである．これらの分類は，収集した事故事例の情報を再整理すると同時に，専門家の意見や文献[28],[29],[30],[31],[32]に基づき作成したものである．(a)時刻に関する属性は，始業前，終業後，および昼休憩時間がそれぞれ1つの属性値に分類されるように考え，0:00～8:00を1，8:00～10:00を2として，17:00～24:00の7まで，合計7種類のクラスに分類した．(b)場所に関する属性は，資材・製品等置場や足場などの7種類のクラスに分類した．同様に，(c)作業工程から(j)事故の型に関する属性についても，それぞれの属性が6種類あるいは7種類のクラスからなり，すべての属性に対する属性値の数がほぼ均等となるよう配慮した．表5は，表4の分類に従い206件の事故事例の一部をまとめたものであり，この表に対して本論文で提案する決定アルゴリズムの導出方法の適用を実際に試みた．この表において，例えば(j)事故の型という属性を決定属性と考え，その他の属性を条件属性と考えることによって，複数の属性に基づき1つの属性を決定する決定表と考えることができる．また，決定表に含まれる各行は，1件の事故事例を表し，1つの決定規則と考えることができる．例えば，Case-1は(a)時刻に対する属性値が6，(b)場所に対する属性値が1であり，(j)事故の型に対する属性値が7であることを表し，これらの各行は条件属性を前提部，決定属性を結論部とした決定規則と考えることができる．

4. 決定アルゴリズムの導出

本論文で提案した決定アルゴリズムの導出方法を，表 5 に示す事故事例の決定表に適用し，実際に得られた極小決定アルゴリズムを評価した．決定アルゴリズムの導出は，表 5 における(j)事故の型を決定属性と考えた場合，および(i)障害部位を決定属性と考えた場合の 2 ケースについて実施した．提案方法を適用するにあたり，遺伝的アルゴリズムのパラメータは，1 世代

表 6　5-fold cross validation による実験結果（事故の型）
Table 6 Results by 5-fold cross validation (Accident Type)

Case	C4.5				提案方法			
	認識率	ルール数	属性数	属性値数	認識率	ルール数	属性数	属性値数
1	90.0%	32	8	79	95.0%	25	7	47
					95.0%	25	8	48
					95.0%	27	8	49
2	90.0%	35	7	95	95.0%	27	8	52
					95.0%	28	7	55
					95.0%	27	8	52
3	90.0%	37	6	103	97.5%	29	8	56
					97.5%	28	7	54
					95.0%	27	8	52
4	85.0%	37	5	102	95.0%	27	8	53
					95.0%	28	9	54
					97.5%	26	8	51
5	97.5%	34	7	84	100.0%	28	7	54
					100.0%	28	8	55
					97.5%	28	8	54
平均	90.5%	35.0	6.6	92.6	96.3%	27.2	7.8	52.4

あたりの個体数を 150，エリート個体数を 8 とした．また，交叉率を 60%，突然変異率を 10%とし，探索終了までの世代数を 4000 として極小決定アルゴリズムの導出を試みた．

導出した決定アルゴリズムの簡潔さや有用性を示すため，簡潔な知識を導出するために良く利用されているC4.5アルゴリズム[22]の結果との比較を行った．さらに，ルール数や属性数に基づく簡潔さのみならず，k-fold cross validation法[23]を用いて，導出した決定アルゴリズムの認識率による評価も実施した．

4.1　事故の型を決定属性とした決定アルゴリズム

表 6 は，実際の事故事例の中から(j)事故の型を決定属性と考え，5-fold cross validation 法における 5 回の試行結果をまとめたものである．また，提案方法は各ケースに対して 3 回の試行を繰り返した．この 3 回の試行においては，認識率，ルール数，属性数，および属性値数ともに，ほぼ同じ値を持つ決定アルゴリズムの導出が行われており，遺伝的アルゴリズムによる安定した準最適解の探索が行われているものと考えることができる．

提案方法による決定規則数は平均で27.2個であるのに対して，C4.5アルゴリズムによって導出された決定規則数は平均で35.0個となり，提案方法が簡約化という意味では優れていることが分かる．属性数に関しては，提案方法が7.8個，C4.5アルゴリズムが6.6個となり，わずかではあるが提案方法が多くの属性を残すという結果となった．これは，簡約化の過程として，C4.5アルゴリズムが情報エントロピーをもとに属性を絞り込んでいくためと考えられる．しかし，属性値の数では

表 7　決定アルゴリズムの一例（事故の型）
Table 7 Example of decision algorithms (Accident Type)

Rule	(a)	(b)	(c)	(e)	(g)	(h)	(i)	(j)
1	*	*	1	*	*	*	*	1
2	5	4	*	*	*	*	*	1
3	*	*	*	6	*	*	4	1
4	2	*	*	*	*	*	4	1
5	*	*	5	*	*	2	*	1
6	*	*	*	*	*	4	6	1
7	*	*	2	*	*	*	7	1
8	*	2	2	*	*	*	*	1
9	*	*	*	*	*	*	1	1
10	*	*	*	*	*	3	3	1
11	2	*	*	*	*	*	5	2
12	*	3	*	*	*	*	6	2
13	6	*	*	2	*	*	*	2
14	*	*	*	4	3	*	*	2
15	*	*	2	2	*	*	*	2
16	*	*	*	*	5	*	6	2
17	*	*	*	*	*	*	2	5
18	*	1	*	*	1	*	*	5
19	*	6	*	3	*	*	*	5
20	*	*	*	*	5	4	*	5
21	*	7	*	3	*	*	*	3
22	*	*	*	*	6	6	*	3
23	6	*	*	*	*	1	*	4
24	*	*	*	*	*	3	6	6
25	*	*	*	5	*	4	*	7

表 8　5-fold cross validation による実験結果（障害部位）
Table 8 Results by 5-fold cross validation (Injured part)

Case	C4.5				提案方法			
	認識率	ルール数	属性数	属性値数	認識率	ルール数	属性数	属性値数
1	85.0%	40	6	108	90.0%	36	9	67
					92.5%	35	9	65
					90.0%	34	9	63
2	95.0%	36	8	91	97.5%	34	8	67
					97.5%	35	9	68
					97.5%	35	9	67
3	95.0%	37	8	96	92.5%	36	9	68
					95.0%	34	9	64
					92.5%	34	8	64
4	85.0%	38	7	103	95.0%	34	8	63
					95.0%	35	7	65
					95.0%	34	8	63
5	95.0%	37	8	95	100.0%	37	9	71
					100.0%	35	9	67
					100.0%	37	9	71
平均	91.0%	37.6	7.4	98.6	95.3%	35.0	8.6	66.2

提案方法が52.4個，C4.5アルゴリズムが92.6個となり，全体として約56%のサイズに簡約化されていることが分かる．また，認識率に関しては，どちらも90%を越す高い認識率となったが，提案方法では96.3%という極めて高い認識率を得ることができた．

表7は，表6のCase1において，提案方法の初回の試行で得られた極小決定アルゴリズムの例である．表7における”*”はドントケア記号であり，これらの属性値はいずれの値でも良いということを意味する．元の決定表（表5）は，9個の条件属性から事故の型を決定している206件の事例からなる決定表であり，合計1854個の属性値から構成されていた．これに対して，提案方法によって得られた簡約化された決定アルゴリズムは平均で52.4個の属性値で構成されていることになり，全体として約2.8%のサイズに簡約化された結果が得られた．

4.2　障害部位を決定属性とした決定アルゴリズム

表 8 は，実際の事故事例の中から(i)障害部位を決定属性と考え，5-fold cross validation 法における 5 回の試行結果をまとめたものである．事故の型を決定属性とした場合と同様に，提案方法は各ケースに対して 3 回の試行を繰り返した．この 3 回の試行においては，このケースにおいても，各値がほぼ同じとなる決定アルゴリズムの導出が行われており，遺伝的アルゴリズムによる安定した準最適解の探索が行われているものと考えることができる．

提案方法による決定規則数は平均で 35.0 個であるのに対して，C4.5 アルゴリズムによって導出された決定規則数は平均で 37.6 個となり，事故の型を決定属性とした場合に比べて，簡約化という意味ではあまり差異が見られなかった．属性数に関しては，提案方法が 8.6 個，C4.5 アルゴリズムが 7.4 個となり，事故の型を決定属性とした場合と同様に，わずかではあるが提案方法が多くの属性を残すという結果となった．しかし，属性値の数では提案方法が 66.2 個，C4.5 アルゴリズムが 98.6 個となり，全体とし約 67%のサイズに簡約化されていることが分かる．また，認識率に関しても，事故の型を決定属性とした場合と同様に，どちらも 90%を越える高い認識率となったが，提案方法では 95.3%という極めて高い認識率を得ることができた．両方法の認識率が 90%を越え，比較的高くなった理由としては，似たような事故が多発しているという現状を物語っているものと考えられる．すなわち，作業工程や架設工法などの他の条件属性が特定されると，障害部位などの決定属性を比較的容易に推定することができるものと考えられる．

表9は，表8のCase1において，提案方法の初回の試行で得られた極小決定アルゴリズムの例である．表9における”*”はドントケア記号であり，これらの属性値はいずれの値でも良いということを意味する．提案方法によって得られた簡約化された決定アルゴリズムは平均で66.2個の属性値で構成されていることになり，全体

表 9　決定アルゴリズムの一例（障害部位）
Table 9 Example of the decision algorithms (Injured part)

Rule	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(j)	(i)
1	*	1	*	*	1	*	*	*	*	5
2	2	*	*	*	*	*	*	1	*	5
3	*	*	*	*	4	*	1	*	*	5
4	5	*	*	*	*	2	*	*	*	5
5	*	*	*	*	*	*	*	3	2	5
6	*	*	*	*	*	*	2	*	3	5
7	3	2	*	*	*	*	*	*	*	5
8	*	4	*	*	*	*	4	*	*	5
9	*	*	*	*	*	1	*	*	5	5
10	*	*	*	*	*	*	*	*	7	5
11	*	4	*	1	*	1	*	*	*	4
12	5	7	*	*	*	*	*	*	*	4
13	2	*	*	*	*	*	*	*	1	4
14	*	1	*	*	*	*	*	1	*	4
15	*	4	*	*	*	*	6	*	*	4
16	*	2	*	*	*	6	*	*	*	4
17	*	*	*	*	*	*	4	*	5	2
18	*	*	*	*	*	4	2	*	*	6
19	*	*	4	*	*	*	*	*	*	6
20	*	*	6	*	*	*	*	*	*	6
21	6	*	*	*	*	*	*	4	*	6
22	6	*	*	*	6	*	*	*	*	6
23	*	*	*	*	*	*	*	*	6	6
24	*	*	*	*	*	*	5	*	2	6
25	*	*	*	*	*	*	1	2	*	6
26	*	*	*	*	*	*	*	6	3	3
27	*	2	*	*	*	*	*	2	*	3
28	5	*	*	*	*	*	*	6	*	3
29	7	*	*	*	*	*	*	*	*	3
30	*	6	*	*	*	*	*	*	1	3
31	*	*	*	*	*	5	*	*	*	7
32	*	5	*	*	*	*	*	3	*	7
33	*	*	*	*	*	*	4	6	*	7
34	*	*	*	*	*	4	6	*	*	7
35	*	*	*	2	*	*	*	6	*	1
36	*	5	*	*	*	*	*	7	*	1

として約3.6%のサイズに簡約化された結果が得られた．事故の型を決定属性と考えた場合に比べて，簡約化という面では少し悪い結果となった．これは，表4に示す障害部位のクラス分けにおいて「その他」という曖昧な分類が含まれているため，簡約化のみならず認識率においても，事故の型を決定属性と考えた場合に比べ，わずかではあるが悪くなったものと考えられる．

5. おわりに

本論文では，ラフ集合を用いて事例から決定アルゴリズムを導出する過程に遺伝的アルゴリズムを適用し，データ量に依存しない比較的少ない計算量で，簡潔かつ有用な決定アルゴリズムを導出する方法を提案した．提案方法を簡潔さの面から検証するために，C4.5 アルゴリズムによる結果との比較を実施した．また，得られた決定アルゴリズムの有用性を検証するため，k-fold cross validation 法を用いて，認識率についての検証も実施した．実際に，提案方法を橋梁の架設現場で発生した事故事例に適用して，以下のような知見が得られた．

1) (j)事故の型，および(i)障害部位という 2 種類の属性値を決定属性と考え，提案方法により決定アルゴリズムを導出した．導出した決定アルゴリズムのサイズは，それぞれ約 2.8%と約 3.6%に簡約化され，C4.5 アルゴリズムの結果と比較しても，簡潔さという面では良好な結果が得られた．

2) 決定規則の数に関して，事故の型を決定属性と考えた場合に平均で 27.2 個，障害部位を決定属性とした場合に平均で 35.0 個となった．C4.5 アルゴリズムの結果である 35.0 個と 37.6 個に比べて，少ない決定規則で元の決定表を全て被覆している完全かつ無矛盾な決定アルゴリズムの導出ができた．

3) 属性数に関しては，事故の型を決定属性とした場合に平均で 7.8 個，障害部位を決定属性とした場合に平均で 8.6 個となり，C4.5 アルゴリズムの結果に比べてそれぞれ 1 個強多い結果となった．しかし，属性値の数では，事故の型を決定属性とした場合に平均で 52.4 個，障害部位を決定属性とした場合に平均で 66.2 個となり，C4.5 アルゴリズムの結果に比べて，それぞれ約 60%のサイズに簡約化された決定アルゴリズムが導出できた．

4) 導出した決定アルゴリズムをルール形式の決定規則で記述すると，その前提部は 1〜3 種類の属性と属性値の組で構成されている．人間が同時に判断できる数を考えると，見通しが良く，表現されている決定規則も理解しやすいものとなった．

5) k-fold cross validation の 1 ケースに対して，提案方法による決定アルゴリズムの導出をそれぞれ 3 回繰り返した．これら 3 回の試行において，認識率，ルール数，属性数，および属性値数のいずれの値もほぼ同じ値となり，遺伝的アルゴリズムによる安定した準最適解の探索が行われていることがわかった．

6) 事故の型を決定属性とした場合に比べて，障害部位

を決定属性とした場合では，認識率，ルール数，属性数，属性値数のいずれの値もわずかではあるが劣る結果が得られた．これは障害部位のクラス分けにおいて「その他」という曖昧な分類を設定したためと考えられる．

本論文では導出した決定アルゴリズムの簡潔さの指標として，ルール数，属性数，および属性値数を用いて，有用性の指標として，認識率を用いた．これらの指標に関しては，C4.5 アルゴリズムに比べ良好な結果が得られた．今後，さらに最適化された決定アルゴリズムの導出のためには，目的の異なる指標を全体的に評価できる，多目的最適化におけるパレート最適解などの検討が必要である．また，ラフ集合を用いた知識獲得手法では，離散化が重要な課題であり，離散化の方法は認識率や簡潔さに影響を与えることが分かった．今後，より最適な決定アルゴリズムの導出のためには，最適な離散化手法の検討が必要である．

文　献

[1] 中村秀治，松浦真一，松井正一，寺野隆雄，”知識工学的手法に基づく水力構造物の寿命予測，” 土木学会論文集，No.368/I-5，pp.301-310，1986.

[2] 中村秀治，松浦真一，寺野隆雄，篠原靖志，”水力構造物の寿命予測エキスパート・システムとその適用，” 土木学会論文集，No.374/I-6，pp.513-521，1986.

[3] 鴻池一季，”パソコンを用いた土留め工法選定支援エキスパートシステムの構築事例，”土木学会論文集，No.385/VI-7，pp.134-142，1987.

[4] 三上市藏，家頭圭晶，河野吉次郎，広兼道幸，”切土のり面保護工選定に関する知識ベース・システム，”土木学会論文集，No.403/VI-10，pp.121-129，1989.

[5] 上野春樹，”知識工学入門，”オーム社，pp.1-35，1989.

[6] 福田剛志，森本康彦，徳永豪，データマイニング，共立出版，2001.

[7] 山本英子，梅村恭司訳，”データマイニング，”共立出版，pp.1-10，1998.

[8] 有川節夫，”発見科学とデータマイニング，”共立出版，pp.1-3，2001.

[9] 森崎美紀，古田均，広兼道幸，”構造物同定問題への遺伝的アルゴリズムの適用，”土木学会年次学術講演会講演概要集，CS-70，pp.138-139，1998.

[10] 朱牟田善治，山本広祐，”ライフラインのネットワーク最適化問題における GA 固有の役割，”土木学会年次学術講演会講演概要集，CS-73，pp.144-145，1998.

[11] 山本広祐，佐々木康彦，”セルオートマトンと GA の組み合わせ利用による多目的構造最適化，” 土木学会年次学術講演会講演概要集，CS-81，pp.162-163，1999.

[12] 山本広祐，朱牟田善治，”工学的システムにおけるニューラルネットワーク利用の位置付け，” 土木学会年次学術講演会講演概要集，CS-74，pp.146-147，1998.

[13] 河村圭，宮本文穂，三宅秀明，中村秀明，”階層型ニューラルネットワークを用いた橋梁診断エキスパートシステムの実用化，” 土木学会年次学術講演会講演概要集，CS-104，pp.208-209，1999.

[14] 須藤敦史，星谷勝，”データマイニングにおける GA・情報エントロピーの適応について，” 土木学会年次学術講演会講演概要集，VI-438，pp.875-876，2002.

[15] 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，” 土木学会年次学術講演会講演概要集，VI-440，pp.879-880，2002.

[16] 宮本文穂，加賀山泰一，田中信也，中村秀明，河村圭，”データマイニングによる橋梁伸縮継手の損傷に関する知識の獲得，” 土木学会年次学術講演会講演概要集，VI-442，pp.883-884，2002.

[17] 佐藤丈晴，荒木義則，中山弘隆，水山高久，古川浩平，”ラフ集合を用いた土石流発生・非発生における地形の規則性に関する研究，”土木学会年次学術講演会講演概要集，VI-443，pp.885-886，2002.

[18] Pawlak, Z.,”Rough Sets - Theoretical Aspects of Reasoning about Data，”Kluwer Academic,” 1991.

[19] 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.

[20] 古田均，広兼道幸，田中成典，三雲是宏，”橋梁の損傷度診断事例からのラフ集合を用いたルール型知識の獲得方法，”構造工学論文集，Vol.44A，pp.521-528，1998.

[21] Furuta, H.，Hirokane, M.，Mikumo, Y.，”Extraction Method Based Rough Set Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels，” Rough Sets in Knowledge Discovery 2, Pphysica Verlag，Springer Verlag，pp.178-192，1998.

[22] Quinlan, J.R.，”C4.5－Programs for Machine Learning，Morgan Kaufmann Publishers，” 1993.

[23] Liang, K.H.，Krus, D.J.，Webb, J.M.，”k-fold Cross Validation in Canonical Analysis，” Multivariate Behavioral Research，pp.539-545，1995.

[24] 横森貴，小林聡，”ラフ集合と意思決定，”数理科学，No.375，サイエンス社，pp.76-83，1994.

[25] 日本ファジィ学会編，”ファジィとソフトコンピューティングハンドブック，”共立出版，pp.545-564，2000.

[26] 中村昭，”ラフ集合と論理・推論，”数理科学，No.374，サイエンス社，pp.86-91，1994.

[27] 電気学会 GA 等組合せ最適化手法応用調査専門委員会編，”遺伝的アルゴリズムとニューラルネット，”コロナ社，pp.1-34，1998.

[28] 阪神高速道路公団安全管理委員会編，都市高速道路の建設・管理における安全管理，理工図書，1997.

[29] 橘田敏之，小村敏，PC橋架設工法総覧，技報堂出版，1984.

[30] プレストレスト・コンクリート建設業協会編，PC 道路橋計画マニュアル，1997.

[31] プレストレストコンクリート建設業協会編，PC 工事安全管理指針，1999.

[32] 安全工学協会編，人身災害，海文堂出版，1982.

Extraction Method of Minimum Decision Algorithms Using Rough Sets and Genetic Algorithms

Michiyuki HIROKANE, Hideyuki KONISHI, Ayaho MIYAMOTO,
and Fumihiro NISHIMURA

Abstract

With the improvement and the popularization of computer in recent years, the quantity of data that is accumulated and available has been increasing steadily. So, the highly advanced and intelligent works are required by using the accumulated data. In this paper, the extraction method of the minimum decision algorithms by using rough sets and genetic algorithms are proposed. The proposed method can extract the efficient and the simplified decision rules. The decision algorithms were extracted from the accident cases in construction sites. The identification rate and the simplification of the extracted knowledge was evaluated by using the k-fold cross validation method.

Keywords

Data mining, Rough Sets, Genetic Algorithms, Decision Rules, Accident Cases