Vol. 33 No. 5 情報処理学会論文誌 May 1992

事例からのルール抽出：RF 2 アルゴリズム†

斉藤和巳†† 中野良平††

知識獲得はエキスパートシステム構築の最重要課題である．エキスパートからの知識（ルール）抽出は困難で，可能な限りの自動化が強く望まれている．本論文では，応用の広い分類問題を対象に，簡潔で十分に汎化した分類ルールを少ない事例からでも抽出可能とする RF 2 法を提案する．RF 2 法はルール候補の生成と精錬の2フェーズからなり，正の事例の記述を逐次一般化することによりルール候補を生成し，IDA* と呼ばれる探索手法を用いてそれらを最適なルールの集合に精錬する．RF 2 法を人工問題へ適用した結果，PAC-学習で必要とされる4分の1の数の事例から，十分に正確なルールを抽出できた．また，最新のルール抽出法である GREEDY 3 でも抽出できなかったルールを少ない事例からでも抽出できた．RF 2 法を医療診断問題へ適用した結果，未知の事例に対する正答率が医者の知識を用いて作成したエキスパートシステムのものに匹敵した．抽出した7個のルールは，人間にとっても容易に理解できる程度に簡潔であった．さらに，抽出に要した時間はワークステーションを用いて2分程度であり，実用規模の問題にも十分適用可能であることが分かった．

1. はじめに

知識獲得はエキスパートシステム構築の最重要課題である．エキスパートからの知識（ルール）抽出は困難で，可能な限りの自動化が強く望まれている．

事例からのルール抽出は，事例を用いて新たな知識の生成を行う記号表現による概念学習である．機械学習の分野では，近年最も活発に研究が行われている[17]．これらの手法の共通点の1つは，可能な限り簡潔な知識の生成を目標とすることである[9]．簡潔な結果は，人間にとっても理解が容易であり，ニューラルネットなどの学習では考慮されない知識獲得に望ましい性質である[9]．

記号表現による概念学習の代表的なアルゴリズムには，ID3[14]，AQ[8] 等があり，現実問題等にも適用され，広く研究されている．また，これらの手法をベースに機能拡張や改良を加えたアルゴリズムの提案もある[3),13]．しかしながら，可能な限り簡潔な知識を生成すること，すなわち，生成するルール（条件）の数を最小にすることは NP 困難な問題であり，一般に，結果の良さと計算量のトレードオフが重要になる．

エキスパートシステム構築の知識獲得という観点からは，ある程度の計算量を必要としても，より良いルールの抽出が望まれる．本稿では，比較的規模の大きな問題（例えば，80 属性，4,000 事例）でも，ワークステーションを用いて妥当な計算時間（数十分）で抽出が完了することを条件とし，より良いルールの抽出を可能とするアルゴリズムについて論じる．

従来の代表的なアルゴリズムの共通点は，属性を逐次選択しながらルールを生成することにある．しかし，この属性選択方式では，一般に，正確にルール抽出できない問題が存在する．例えば，不要な属性を持つパリティー問題である．理論的には，任意の1つの属性での値（オン/オフ）に着目しても，それぞれ，パリティー条件を満たす事例とそうでない事例は同数である．したがって，属性選択方式では，一般に，終了条件が成立するまで，ランダムな属性の選択を繰り返さねばならない．

概念学習の結果（ルール）の信頼性評価には，一般に，さらに進んで未知の事例に対する正答率を調べる実験が必要である．しかし，一定の確率分布に従って毎回独立に事例が現れ，学習する概念の属するクラスがあらかじめ分かっていれば，PAC (Probably Approximately Correct)-学習の理論[2),18]を用いて，確率的に汎化能力を保証することができる．すなわち，学習する知識の複雑さから，学習結果を保証するのに必要な事例数を求めることができる．現実の問題では，前もって知識の複雑さが分からず，適用できないことが多いが，人工問題を用いれば，アルゴリズム評価のための重要な指標を得られる．

本論文では，応用の広い分類問題を対象に，少ない事例からでも簡潔で汎化された分類ルールを抽出する RF 2 法について述べる．まず，ルール抽出の枠組みと従来手法の問題点について述べる．次いで，RF 2 アルゴリズムの特徴とその詳細について述べる．最後に，RF 2 法の人工問題，および，現実問題への適用

† Rule Extraction from Facts: RF 2 Algorithm by KAZUMI SAITO and RYOHEI NAKANO (NTT Communication Science Laboratories).

†† NTT コミュニケーション科学研究所

結果について述べる．

2. フレームワーク

分類ルールの抽出を行う枠組みについて説明する．概念（concept）は，既に学習した他の概念またはその概念の基本構成要素である属性（attributes）によって表現される．例えば，病名という概念は｛体温，血圧，性別，…｝という属性によって表現できる．また，事例は分類したい概念に属す正の事例と属さない負の事例からなる．例えば，正の事例はその病名の患者であり，負の事例は非患者である．そして，各事例は属性空間内の点として位置付けられる．

一般に，属性を用いて概念を表現する形式は論理式のクラスとして定式化される．ここではそのクラスとして，人間にとって直感的で理解しやすいと考えられる選言標準形（DNF：Disjunctive Normal Form）を扱う．DNF はターム（term）の論理和として表現され，タームはリテラル（literal）の論理積として表現される．各リテラルは対応する1つの属性にのみ関係し，その属性の値によって，リテラルの値は真または偽となる．本稿では，タームを1つのルール，DNF をルールの集合とみなす考え方をとる．

なお，本稿では，ノイズを考慮しない範囲でのルール抽出について論じる．なぜなら，新たな手法を提案する場合，理想的な環境下で実験を行うことにより，アルゴリズムの特徴が顕著になり，従来法との比較も容易になる．すなわち，この段階において手法の優位性が明らかとなった後に，ノイズへの対処法を検討するのが妥当であると考える．

3. 従来法とその問題点

生成するルール（条件）の数を最少にするという尺度の下で，厳密解を求めるアルゴリズムの計算量は莫大となり，到底実行することはできないと考えられる．なぜなら，単純に計算すれば，属性数が N のとき，各属性ごとに3通り（肯定的，否定的，無関係）の条件の現れ方があるので，可能なルール候補の個数は 3^N となる．さらに，求めたルール候補の集合には一般に冗長性が存在するので，その中から適当な組み合わせを選択する必要がある．したがって，効率的なルール抽出にはヒューリスティクスを用いたアプローチが必然になる．以下に，代表的な3種の概念学習アルゴリズムを概説し，それらの問題点を指摘する．図1には，ここで取り上げるアルゴリズムの系譜を示す．なお，本論文で提案する RF 2 法は，RF 法[16]の改良アルゴリズムである．

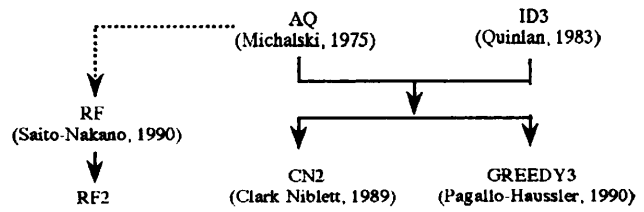


図 1 アルゴリズムの系譜
Fig. 1 Descent of algorithms.

第1の手法は事例から決定木（decision tree）を構築する方法であり，ID 3[14]が代表的アルゴリズムである．ID 3 は，すべての分割された事例集合が正または負の事例だけになるまで，分割を再帰的に実行する（divide and conquer）．ここで，事例集合の分割には，情報量の期待値に基づいて選択された属性の値が用いられる．

決定木学習の問題点は，決定木そのものの表現能力が不十分なことである．例えば，一般に，和結合した概念 $x_1x_2 \vee x_3x_4 \vee x_5x_6$ は簡潔な結果に学習できない．なぜなら，x_1 で最初の分割が行われた場合，2つの分割された事例集合のそれぞれが $x_3x_4 \vee x_5x_6$ を分類しなければならない．さらに，それぞれの部分集合で x_3 が選択された場合には，x_5x_6 を分類する部分木は全体で4つ必要になる．したがって，x_5x_6 を満たす事例も適当に4つに分割され，特に，元の事例数が少ない場合には，決定木の細部まで完全に学習することは極めて困難になる[13]．

第2の手法は，AQ システム[10]，概念クラスタリング[11]の基本アルゴリズムとして用いられる AQ[8] である．AQ の考え方に基づいたルール抽出アルゴリズム AQR[3] について説明する．AQR では，すべての正の事例が少なくとも1つのルール（スター）にカバーされるまで，ルールの生成を繰り返す．ルールの生成では，まず，どのルールにもカバーされていない正の事例を seed としてランダムに選択する．次に，ヒューリスティク評価関数に基づき，seed をカバーして，すべての負の事例が排除できるルールを Beam サーチにより生成する．ヒューリスティクスでは，より多くの正の事例をカバーして，より多くの負の事例を排除できる属性の組み合わせを優先して探索する．

AQR のアルゴリズムを一部修正して，すべての正の事例を seed としてルール生成を行うとすれば，結

果はかなり冗長なルール集合になってしまう．冗長なルールを消去することは，後述するように，ルールを集合，それに含まれる正の事例を集合の要素と考えることにより，SC 問題 (set covering)[6] として定式化できる．ところが，AQR のようにランダムに seed を選択することは，SC 問題において，すべての要素をカバーするまで，ランダムに集合（ルール）の選択を繰り返すことにほかならない．したがって，SC 問題の解法を追及することにより，より良いルール集合の抽出が可能になると考える．

第3の手法は決定リスト (decision list)[15] と呼ばれる if-then-else 形式のルールリストを構築する方法であり，CN 2[3], GREEDY 3[13] 等のアルゴリズムが提案されている．2つの方法では，すべての正の事例が消去されるまで，ルールの生成とそこでカバーされた正の事例の消去を繰り返す (separate and conquer)．ルールの生成では，ヒューリスティク評価関数に基づき，より多くの正の事例をカバーできる属性の組み合わせを逐次選択する．なお，CN 2 では，AQ と同様に Beam サーチが用いられる．ヒューリスティクスは，CN 2 では情報量の期待値が用いられるのに対し，GREEDY 3 ではベイズ規則により評価される．この2つの方法では，抽出ルールの表現形式を決定木から決定リストに変えることにより，第1の手法と比較して，表現能力を向上させ，SC 問題として定式化できるルール集合の精錬を欲張り法 (greedy algorithm)[6] に基づいて実行することにより，第2の手法と比較して，その精錬能力を向上させた．ここで，欲張り法とは，すべての点（事例）が消去されるまで，最も多くの点をカバーする集合（ルール）の選択とその集合にカバーされた点の消去を繰り返す方法である．

しかし，欲張り法は近似解法であり，さらに SC 問題の解法を追及する余地がある．すなわち，CN 2 と GREEDY 3 により順番に生成されるルールが，一般に，最終的なルール集合として適切なものを生成しているとは限らない．例えば，初めに抽出したルールが，後から抽出したルールの集合の和で完全に包含されるケースも考えられる．したがって，ルール生成段階で，ルール候補としてルールを冗長に生成し，次のフェーズとして，それらを簡潔なルール集合に精錬することにより，より良いルール集合の抽出が可能になると考える．

さらに，すでに指摘したように，以上説明した手法に基づいた4種のアルゴリズムの共通点は，属性を逐次選択しながらルール生成をすること，すなわち，属性選択方式を採用していることにある．したがって，不要な属性を持つパリティー問題などでは，極めてルール抽出が困難となる．一方，RF 2 法では，属性選択方式を採用せず，事例の記述を逐次一般化してルール抽出することにより，この問題点の克服を試みている．

4. RF 2 アルゴリズム

4.1 アルゴリズムの概略

RF 2 法の1つの特長は，正の事例の記述を一般化することによりルール生成をすることである．このため，属性選択方式では適当なヒューリスティク評価尺度が得られない問題でも，類似した事例の記述を逐次一般化することにより，良いルールを得ることが期待できる．しかし，一般に，この方法だけでは，冗長なルール集合が生じる場合も考えられる．すなわち，あるルールが他のルールの和集合で包含される場合である．この問題に対処するため，RF 2 法のもう1つの特長として，ルール候補の集合を最適なルール集合に精錬するフェーズを備えている．さらに，このフェーズは，IDA* と呼ばれる探索手法を用いることにより，効率良く解を得ることができる．

4.2 ルール候補の生成

RF 2 法によるルール候補生成の特徴は，各正の事例の記述を逐次一般化することにより，special-to-general 探索でルール候補を生成することである．すなわち，正の事例を seed として選択し，ヒューリスティクスにより選択する正の事例を順次用いて，seed の記述の最小限の一般化を繰り返し，ルールを生成する．一方，3章で説明した従来の方法では，ヒューリスティク評価関数等を利用して属性を選択することにより，general-to-special 探索でルールを生成する．

seed の記述の一般化には，参照統合オペレータ (Ref-Union operator)[11] を用いる．今後，このオペレータを RU と略記する．RU では，seed の記述と正の事例，および，すべての負の事例から，以下の手続きにより，seed の記述を出力する．

- RU オペレータ

1. 事例の記述と正の事例の属性値の論理和を各属性ごとにとり，中間記述を生成する．
2. もし，中間記述が負の事例を1つでもカバーすれば，元の記述を結果とする．さもなければ，

中間記述を結果とする．

例えば，Dを seed の記述，p_i を正の事例，n_j を負の事例とし，$D=(1,0,0\vee1)$，$p_1=(1,0,0)$，$p_2=(1,1,1)$，$p_3=(0,0,0)$，$n_1=(0,0,1)$，$n_2=(0,1,0)$，とすれば，

RU $(D, p_1)=(1,0,0\vee1)$,
RU $(D, p_2)=(1,0\vee1,0\vee1)$,
RU $(D, p_3)=(1,0,0\vee1)$,

が結果となる．

正の事例を選択するヒューリスティクスとして，以下の2つを考案した．第1に，既に得られているルール候補集合にカバーされた事例の優先度を低くする．第2に，seed とした事例との距離が小さい事例の優先度を高くする．すなわち，前者は，類似したルールを繰り返し生成することを極力避けるためであり，後者は，類似する事例を同じルールにカバーさせるためのヒューリスティクスである．

ルール候補の生成アルゴリズムの詳細を以下に示す．ここで，seed の記述は D_0 から D_f へ一般化される．ALPHA は最終的なルール候補の集合となり，BETA はアルゴリズムを制御するためのルール候補の集合である．また，ユーザの定義する MC は，一般化された seed の記述を何個ルール候補とするかを制御するパラメータである．ただし，以下での実験では，$MC=1$ とした最もシンプルな場合の結果である．

- ルール生成アルゴリズム

1. 各正の事例を seed として，以下のヒューリスティクスに基づき，各 seed の記述 D_0 を逐次一般化する．
2. BETA に含まれるルール候補集合にカバーされていない正の事例を対象に，最も多くの事例をカバーする seed の記述 D_f を BETA に加え，多くの事例をカバーする上位 MC 個の seed の記述 D_f を ALPHA に加える．
3. すべての正の事例が BETA に含まれるルール候補でカバーされたならば処理を終了する．さもなければ，各 seed の記述 D_f を D_0 に戻し，1. に戻って処理を繰り返す．

- ヒューリスティクス

1. BETA に含まれるルール候補にカバーされる回数が少ない事例の優先度を高くする．
2. 前項が同じ事例に対しては，seed の記述 D_0 との距離が小さい事例の優先度を高くする．

4.3 ルールの精錬

ルールの精錬には，冗長に生成したルールを除去するルールレベル精錬と，各ルールに現れる冗長なリテラルを除去するリテラルレベル精錬がある．まず，SC 問題とは，集合の族 $F=\{S_1,\cdots,S_n\}$ と各集合 $S_i=\{p_{i1},\cdots,p_{ik_i}\}$ が与えられたとき，

$$\{F'\subset F:\ \bigcup_{S\in F'}S=\bigcup_{S\in F}S\}$$

から $|F'|$ を最小にするものを求める問題である．2つのタイプの精錬は，次のように対応させれば SC 問題として定式化できる．ルールレベル精錬はルール候補を集合 S_i，それに含まれる正の事例 p_{ij} を要素と考える．また，リテラルレベル精錬では，各ルールをそれぞれ独立に扱い，リテラルを集合 S_i，それにより排除される負の事例 p_{ij} を要素と考える．

しかし，SC 問題は NP 困難な問題なので[5]，一般には，厳密解を求めることは大変難しく，近似解を求めるアプローチが採用されている．解法としては，最も多くの点をカバーする集合の選択とその集合にカバーされた点の消去を繰り返す欲張り法がよく用いられている．この方法の特徴は高速性と解の高品質にある．すなわち，解の最悪ケースが良い精度で保証され[6]，事例数を N，最適解の集合の個数を K，欲張り法で求めた集合の個数を S とすれば次式が成り立つ：

$$S\leqq K+(\log_e N+1).$$

ところが，欲張り法の弱点は，例えば次のようなケースに現れる：$A=\{1,2,3\}$，$B=\{2,3,4,5\}$，$C=\{4,5,6\}$，すなわち，まず，Bが選ばれ，続いてAとCがそれぞれ選ばれる．このケースでは，明らかにAとCの2つの集合で十分である．RF 2 法でのルール候補の生成でも，多くの正の事例をカバーする seed の記述をルール候補とするため，まさにこのケースが起こると考えられる．したがって，この方法では高精度のルールの精錬ができない．

既述のように RF 2 法では，第一フェーズにおいてルール候補を絞り込んで生成し，RU オペレータの働きにより，各ルール候補の不要なリテラルを消去しているため，厳密解を求めるアプローチも現実的である．すなわち，A^* 法[12]を用いた探索が可能である．まず，集合が選択されていない状態をルートとする．そして，次に展開するノード（選択する集合）は以下の目的関数 $f(n)$ を最小にするものである．この評価尺度 $f(n)$ は展開されたノードに対しても再帰的に用いられる．最終的に，最初にすべての点をカバーした

ノードをルートまで手繰ることにより結果を得ることができる．

$f(n)=g(n)+h(n)$,

$g(n)$=親ノードまでの探索木の深さ，

$h(n)$=|親ノード以前でカバーされていない点|
÷|ノード(集合)がカバーする点|．

しかし，A^* 法は最良優先探索であり，現在の計算機アーキテクチャでは，メモリ容量の限界からすぐに計算が不可能となる．そこで，RF 2 法では，探索の閾値を更新しながら繰り返し深さ優先探索を行う IDA* 法[7] を用いてルールを精錬する．また，処理の効率化のため，任意の探索段階で一般に用いられる以下の規則を適用した．

- 他の集合の部分集合となる集合は選択しない．
- 点を固有にカバーする集合を優先して選択する．
- バックトラックの結果，カバーできない点が現れた場合，さらにバックトラックする．

なお，RF 2 法で用いたヒューリスティク関数は最も単純なものである．SC 問題の効率的な解法を追及した研究[1),4)] では，工夫を凝らしたヒューリスティク関数が提案されている．これらの関数を導入することにより，RF 2 法の効率をさらに改善することができると考える．

5. 汎化能力評価

5.1 目標概念

RF 2 法の汎化能力を評価するため，以下に示す 3 タイプの目標概念を考える．これらの概念の要約を表 1 に示す．なお，これらの目標概念は GREEDY 3 に対して行われた実験と同じものである[13]．

ランダム DNF 概念：ランダムに生成した比較的小規模な DNF により記述された概念である．ただし，ターム数はあらかじめ指定され，各タームに現れるリテラル数は正規分布（平均と標準偏差を指定）に従って決定された．実験では 4 つのランダム DNF を扱う．その中で，2 つの DNF はモノトーン（負のリテラルが現れない）である．

MX (multiplexor) 概念：$k+2^k$ 属性（ビット）の始めの k 属性がアドレス，それに続く 2^k 属性がデータとなる概念である．実験では，$k=2$, 3 の場合を扱い，それぞれ 10, 21 個の不要な属性を付加した．

パリティ概念：k 属性だけに着目したとき，それが k ビットの偶パリティとなる概念である．実験では，4，5 パリティを扱い，それぞれ 12, 27 個の不要な属性を付加した．

表 1 目標概念の要約
Table 1 Summary of target concepts.

概念名	属性数	term 数	平均 term 数	学習事例数
dnf 1	80	9	5.8	3,292
dnf 2	40	8	4.5	2,188
dnf 3	32	6	5.5	1,650
dnf 4	64	10	4.1	2,640
mx 6	16	4	3	720
mx 11	32	8	4	1,600
par 4	16	8	4	1,280
par 5	32	16	5	4,000

5.2 事例数

事例の出現には，未知だが一定の分布があり，各事例がその分布に従って，毎回独立に出現すると仮定できるとき，目標概念の複雑さ (complexity) が分かっていれば，PAC 学習の理論を用いて，テスト事例に対する正答率を保証するのに必要な抽出事例数を計算できる．すなわち，概念記述に必要な属性数を n，概念記述に現れるリテラル数を k，テスト事例を正答できない確率を ε とすれば，抽出に必要な事例数 m は以下の式で近似できる[13]．

$$m=\frac{k\times\log n}{\varepsilon} \qquad (1)$$

本稿の実験ではすべて，$\varepsilon=10\%$，テスト事例数＝2,000 とし，すべての事例の属性値には，ランダムに 1 または 0 を与えた（一様分布）．

5.3 実験結果

図 2 に RF 2 法を用いて抽出したルールのテスト事例に対する正答率を示す．各目標概念における 4 つの棒グラフは，一番奥が式(1)で求めた値を抽出事例数

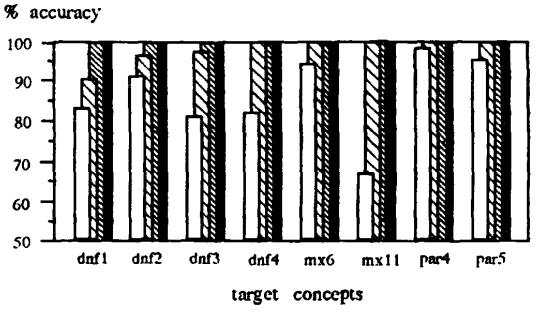


図 2 RF 2 法の学習結果
Fig. 2 RF 2 learning results.

とした結果であり，手前になるに従って抽出事例数が半分となる．すなわち，一番手前の棒グラフは，式(1)の8分の1の事例数で抽出した結果である．

図に示してある正答率は，ランダムに 10 回抽出事例を生成して実験を行った結果の平均値である．抽出の条件 (ε=10%) より，テスト事例に対して 90% 以上正答できていれば，抽出が成功したと言える．したがって，すべての目標概念において，式(1)の4分の1の事例数で十分に抽出が成功したと言える．このことは，分布が一様という条件付ではあるが，RF 2 法は少ない事例からでも汎化されたルールを抽出できることが実験的に明らかになった．

3つの目標概念 (dnf 1, dnf 2, dnf 3) では，式(1)の4分の1の事例数のとき，完全に正しいルールを抽出できなかった．その理由として，正の事例の全体の事例に対する割合が異なるからであると考える．すなわち，3つの目標概念では，その割合が 15%～25% 程度であるのに対し，それ以外の目標概念での割合は 50% 程度となっていた．したがって，RF 2 法では，正の事例数の割合が 50% に近いほうが，抽出が容易になると予想される．従来のアルゴリズムにおいてどのような性質があるかは，興味深い問題である．

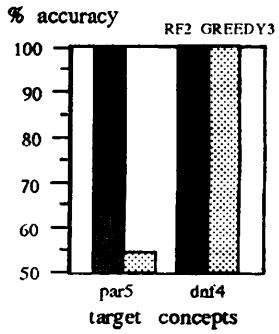
5.4 汎化能力比較

図3に RF 3 法と GREEDY 3 との汎化能力比較を示す．図3(a)には，2つの概念 (par 5, dnf 4) において，式(1)で求めた値を抽出事例数とした結果を示す．パリティ概念は，GREEDY 3 等の属性選択方式のアルゴリズムでは，抽出するのが一般に困難な概念だと言われる[13]．ところが，RF 2 法を用いれば，さらに少ない事例数からでも完全なルールの集合が抽出できた．

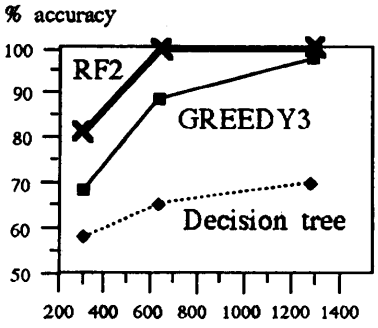
ランダム DNF 概念では，十分な数の事例が与えられれば，GREEDY 3 でも完全なルールの集合が抽出できる．しかし，図3(b)に示すように，事例数が少ない場合には，GREEDY 3 と比較して，RF 2 法が常に高い汎化能力を示している．なお，RF 2 法を決定木による方法と比較すれば，RF 2 法では決定木による方法の4分の1の事例数から抽出を行った場合でも，より高い汎化能力を示していることが分かる．

6. 医療診断問題への適用

現実問題での RF 2 法の能力を評価するため，患者への問診から患者の病名を判定する医療診断問題[19]へ



(a) 2概念の学習結果
(a) Learning results for two concepts.



(b) dnf 4 の学習曲線
(b) Learning curves for dnf 4.

図3 RF 2 対 GREEDY 3
Fig. 3 RF 2 v. s. GREEDY 3.

RF 2 法を適用した．ただし，この問題では患者の頭痛に関する症状と病名だけを扱っている．問診の例は，「あなたの頭痛はいつから始まりましたか？」という質問に対し，患者が「本日」「3日前」等の選択肢に応える形式である．これらの選択肢の合計数は 216 (各属性値は2値) となる．本実験では，約 400 人の患者データを利用し，患者が筋収縮性頭痛であるか判定するルールの抽出を行った．

抽出の結果，すべての患者データから，RF 2 法を用いて7つのルールが抽出できた．これらのルールの条件 (リテラル) の個数は，平均 22 個であり，人間にとっても理解容易なものであると考える．抽出したルールは，例えば，「頭痛の場所が一定で，痛みが4日以上前に始まっている，…，ならば，筋収縮性頭痛である」というものである．

すべてのデータから，300 事例をランダムに選択し，ルール抽出を行い，残りの約 100 事例に対する正答率を調べた．この実験を 10 回行った結果の平均正答率は 73% であり，この値は医者の知識を用いて作

成したエキスパートシステムの正答率と匹敵するものである．したがって，RF 2 法を用いれば，簡潔で汎化したルールの抽出が可能であることを実証できたと考える．さらに，抽出に要した時間はワークステーションを用いてわずか2分程度であり，実用規模の問題に十分適用可能であることも明らかとなった．

7. おわりに

本稿では，知識獲得ボトルネック解決のための試みとして分類ルールを自動抽出する方法について報告した．すなわち，少ない事例からでも簡潔で十分に汎化した分類ルールを抽出する RF 2 法を考案した．今後は，さらに多くのケースに適用してその有効性を検証すること，および，抽出ルールの信頼性指標を示すための検討が重要である．また，ノイズへの対処と領域知識を用いたルールの高度化の検討も必要である．

謝辞 本研究に際し，励ましをいただいた NTT 情報通信処理研究所 知識処理研究部 河岡司部長に感謝いたします．

参考文献

1) Balas, E. and Ho, A.: Set Covering Algorithms Using Cutting Planes, Heuristics, and Subgradient Optimization: A Computational Study, *Math. Program.*, Vol. 12, pp. 37-60 (1980).
2) Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M.: Learnability and the Vapnik-Chervonenkis Dimension, *J. ACM*, Vol. 36, No. 4, pp. 929-965 (1989).
3) Clark, P. and Niblett, T.: The CN 2 Induction Algorithm, *Machine Learning*, Vol. 3, No. 4, pp. 261-283 (1989).
4) Fisher, M. and Kedia, P.: Optimal Solution of Set Covering/Partitioning Problems Using Dual Heuristics, *Manage. Sci.*, Vol. 36, No. 6, pp. 674-688 (1990).
5) Garey, M. and Johnson, D.: *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York (1979).
6) Johnson, D.: Approximation Algorithms for Combinatorial Problems, *J. Comput. Syst. Sci.*, Vol. 9, pp. 256-278 (1974).
7) Korf, R.: Depth-first Iterative Deepening: An Optimal Admissible Tree Search, *Artif. Intell.*, Vol. 27, pp. 97-109 (1985).
8) Michalski, R.: Synthesis of Optimal and Quasi-optimal Variable-valued Logic Formula, *International Symposium on Multi-valued Logic*, pp. 76-87 (1975).
9) Michalski, R. and Kodratoff, Y.: Research in Machine Learning: Recent Progress, Classification of Methods, and Future Directions, Kodratoff, Y. and Michalski, R. (eds.), *Machine Learning: An Artificial Intelligence Approach, Volume 3*, pp. 3-30, Morgan Kaufmann, San Mateo, CA (1990).
10) Michalski, R., Mozetic, I., Hong, J. and Lavrac, N.: The Multi-purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains, *AAAI-86*, pp. 1041-1045 (1986).
11) Michalski, R. and Stepp, R.: Learning from Observation: Conceptual Clustering, Michalski, R., Carbonell, J. and Mitchell, T. (eds.), *Machine Learning: An Artificial Intelligence Approach, Volume 1*, pp. 331-364, Morgan Kaufmann, San Mateo, CA (1983).
12) Nilson, N.: *Principles of Artificial Intelligence*, Toiga Publishing Co., Los Altos, CA (1980).
13) Pagallo, G. and Haussler, D.: Boolean Feature Discovery in Empirical Learning, *Machine Learning*, Vol. 5, No. 1, pp. 71-99 (1990).
14) Quinlan, J.: Induction of Decision Trees, *Machine Learning*, Vol. 1, No. 1, pp. 463-482 (1986).
15) Rivest, R.: Learning Decision Lists, *Machine Learning*, Vol. 2, No. 3, pp. 229-246 (1987).
16) Saito, K. and Nakano, R.: Rule Extraction from Facts and Neural Networks, *INNC-90-PARIS*, pp. 379-382, Paris, France (1990).
17) Stefankis, P., Wnek, J. and Zhang, J.: Bibliography of Recent Machine Learning Research (1985-1989), Kodratoff, Y. and Michalski, R. (eds.), *Machine Learning: An Artificial Intelligence Approach, Volume 3*, pp. 685-789, Morgan Kaufmann, San Mateo, CA (1990).
18) Valiant, L.: A Theory of the Learnable, *Comm. ACM*, Vol. 27, No. 11, pp. 1134-1142 (1984).
19) 益沢ほか：医療知識ベース利用による医療診断支援システム (DOCTORS) の臨床評価，第四回医療情報学連合大会，pp. 672-677 (1984).

(平成3年8月15日受付)
(平成4年2月14日採録)

斉藤 和巳（正会員）

1963 年生．1985 年慶応義塾大学理工学部数理科学科卒業．同年，NTT 入社．以来，神経回路網，機械学習などの研究に従事．NTT コミュニケーション科学研究所研究主任．1991年9月より，オタワ大学客員研究員．人工知能学会，神経回路学会，日本認知科学会各会員．

中野 良平（正会員）

1947 年生．1971 年東京大学工学部計数工学科卒業．工学博士．同年，日本電信電話公社（現 NTT）入社．以来，統計解析，分散処理，データベース，人工知能の研究に従事．現在，NTT コミュニケーション科学研究所主幹研究員．人工知能学会，神経回路学会，日本応用数理学会各会員．